Soft Comput
DOI 10.1007/s00500-008-0391-z

ORIGINAL PAPER

Evolving the memory of a criminal's face: methods to search a face space more effectively

Charlie Frowd · Vicki Bruce · Melanie Pitchford · Carol Gannon · Mark Robinson · Colin Tredoux · Jo Park · Alex Mcintyre · Peter J. B. Hancock

Abstract Witnesses and victims of serious crime are often required to construct a facial composite from their memory, a visual likeness of a suspect's face. The traditional method is for them to select individual facial features to build a face, but often these images are of poor quality. We have developed a new method whereby witnesses repeatedly select instances from an array of complete faces and a composite is evolved over time by searching a face model built using Principal Components Analysis. While past research suggests that the new approach is superior, performance is far from ideal. In the current research, face models are built which match a witness's description of a target. It is found that such 'tailored' models promote better quality composites, presumably due to a more effective search, and also that smaller models may be even better. The work has implications for researchers who are using statistical modelling techniques for recognising faces.

Keywords Face generation · Evolution · Face perception · PCA · Genetic algorithms

C. Frowd (✉) · M. Pitchford · C. Gannon
School of Psychology, University of Central Lancashire,
Preston PR1 2HE, UK
e-mail: cfrowd@uclan.ac.uk

V. Bruce
School of Psychology, Newcastle University, Newcastle, UK

M. Robinson · J. Park · A. Mcintyre · P. J. B. Hancock
Department of Psychology, University of Stirling, Stirling, UK

C. Tredoux
Department of Psychology, University of Cape Town,
Cape Town, South Africa

1 Introduction

For people we know well, face recognition normally occurs accurately and effortlessly. For less familiar faces, recognition involves greater error and uncertainty (e.g. Hancock et al. 2000). Likewise, computer recognition systems involve error and appear to perform similarly to human observers perceiving unfamiliar faces (e.g. Hancock et al. 1996; Phillips et al. 2003). One approach to improve performance from both unfamiliar face recognition and computer recognition is to converge information from several sources. For example, verbal descriptions have been used for many years to help locate visual exemplars in semi-automatic systems (e.g. Shepherd 1986). A recent trend has been to combine different biometric modalities—face, iris and fingerprints (e.g. Jain et al. 2004)—or different types of descriptive information (Annesley et al. 2006). Other approaches may combine different viewpoints or different images of the same person (Burton et al. 2005; Prince et al. 2006). Here, we evaluate an enhancement to a face production system that allows witnesses to construct faces of criminals. The heart of this technology is a model built using Principal Components Analysis (PCA), a typical component of many face recognition systems (e.g. Kirby and Sirovich 1990; O'Toole et al. 1993). The work combines visual and verbal information provided by a witness to both fine-tune the face model and evolve a better likeness of a target.

Witnesses and victims of crime are often asked to describe the appearance of an unfamiliar face—a suspect—and then to construct a facial composite. These images are sometimes shown in the newspapers and on TV crime programmes in an attempt to identify and locate a criminal. The traditional procedure is for witnesses and victims to select individual facial features from a kit of 'face parts'.

Published online: 13 January 2009

There are many such systems available: for example, in the UK, E-FIT and PRO-fit; in the US, FACES and SuspectID. However, this is not an optimal procedure since we do not perceive faces as a set of parts—rather, more as a complete or *holistic* image (e.g. Tanaka and Sengco 1997). Unsurprisingly then, composites constructed using the 'feature' method are not recognised very well (e.g. Davies et al. 2000; Frowd et al. 2004, 2005, 2007a, b). Following a realistic delay of several days, recognition tends to be very poor indeed (e.g. Frowd et al. 2005, 2007c).

An alternative approach is to present sets of complete faces for a witness to select from, a more natural procedure than viewing individual facial features; it is also somewhat similar to a witness selecting faces from a police line-up or mugshot album. While there are several such whole face systems (e.g. ID, Tredoux et al. 1999; EigenFIT, Gibson et al. 2003), the focus here will be on the EvoFIT composite system (Frowd et al. 2004, 2006, 2007c; Hancock 2000; Frowd 2002) developed jointly by the Universities of Stirling and Central Lancashire. With EvoFIT, witnesses select from a set of complete faces, and the selected faces are bred together using a Genetic Algorithm to produce another set for selection. While the faces contain random characteristics at the start, repeating the selection and breeding process a few times enables a specific face to be 'evolved'—hence the name Evolving Facial Identification Technique, or EvoFIT. In practise, witnesses' first select facial *shape*, corresponding to the size and position of features, then facial colouring or *texture*, the colour of the eyes, brows, mouth and overall skin tone. To model hair, a specific hairstyle is chosen from about 500 alternatives and presented on each face (which is subsequently blurred, see below). Additionally, tools are available to manipulate the shape and position of features on demand as well as the more overall or *holistic* properties of the face (e.g. age and masculinity).

Central to the current EvoFIT system are a set of face models that are able to generate a large number of synthetic faces. The models are constructed from PCA, a statistical technique that extracts the dimensions of variation—the eigenvectors—from a set of items, in this case faces. These 'standard' models are built from 72 front-view, white male faces for a given age range. In recent experiments, EvoFIT produced composites that were correctly named on average about 25% of the time after a realistic delay of 2 days. While this figure is still rather low, it is at least twice that of a current 'feature' system such as E-FIT or PRO-fit (Frowd et al. 2006, 2007c). More recent developments have improved performance further, including the use of ageing and other 'holistic' tools (Frowd et al. 2006) and by caricaturing (Frowd et al. 2007). As such, EvoFIT is now being piloted by Lancashire and Derbyshire constabularies, and it is reported to be valuable in criminal investigations between 20 and 30% of occasions.

A problem with EvoFIT has been to ensure reliable convergence on a target face. One of the challenges with PCA face models is their complexity: in being able to generate a wide range of faces, they contain a great deal of information, which is difficult to search, even with genetic algorithms. One successful approach is to run the system more than once with a witness, each time using a new set of random faces (Frowd et al. 2006), capitalising on the fact that random starting points in a complex search space can produce quite different results. An alternative is to be more selective about the faces used to build the model in the first place. For example, if a witness remembers a suspect's face to be thin with small eyes, a model built containing faces that match this description is likely to be valuable. Such a model would not generate wide faces, nor faces with large eyes (since it is based on the statistics of the reference faces). A 'tailored' model of this type could be searched more effectively and thus better likenesses would be evolved.

The thrust of the current work is to explore the effectiveness of evolving faces using standard face models that are matched on age, the approach used with EvoFIT to date, versus tailored face models, which are built specifically to match a witness's memory of a target. Three experiments are presented. The first explored the quality of composites constructed from standard and tailored models built using the same number of faces; the second compared a range of tailored models of different sizes; finally, the third, utilised the 'best' model size found in the second experiment and compared it with the standard model. It was expected that a tailored model would outperform a standard one, and that reducing the complexity of the model still further, by reducing its size, would be even more effective.

1.1 Detailed background to EvoFIT

The procedure used to build the current face models follows that of Sirovich and Kirby (1987) and Troje and Vetter (1996). The initial stage was to carefully photograph 200 white male faces in a front-view pose (the system currently works for Caucasian male faces). The reference faces covered a wide age range, from 16 to 75, as shown in Table 1, and allowed 4 overlapping face models to be constructed, each containing 72 faces centred at 10-year intervals: 20, 30, 40 and 50 years. Models were constructed

Table 1 Age distribution of reference faces used to build the PCA face models

Age	15–20	20–29	30–39	40–49	50–75
Count	13	63	51	45	28

in greyscale, since there is presently no evidence that colour improves results (Frowd et al. 2006).

The next stage was to identify the position of individual landmarks on the 200 reference faces. This is largely a manual procedure that involves careful identification of approximately 300 standard locations on each face (see Fig. 1, far right). For the standard models, these shape co-ordinate files are then subjected to a PCA to provide a set of 72 reference shapes, or eigenshapes. While the eigenshapes can be added together in different amounts to reconstruct the original shapes, adding them in random proportions allows a novel shape to be generated. Note that the face model constructed by PCA is itself holistic in nature, since the eigenshapes change the overall appearance of the face (Hancock et al. 1997). For example, one of the eigenshape components normally adjusts (or models) face length and width.

The features of the reference faces are then morphed to a standard size and position, and a second, texture PCA is carried out on the image pixel values. This produces a set of reference eigentextures that allow random facial textures to be generated. To produce a random face itself—a face that changes by both shape and texture—a random texture is generated that is morphed to a random shape. Hair is selected via the PRO-fit 'feature' system and added to the facial texture prior to the shape morph. The result is a good quality synthetic face, as shown in Fig. 1. A screen shot of 18 such randomly-generated faces (they change by shape and texture) is shown in the "Appendix".

Witnesses are presented with 72 facial shapes, on four consecutive screens each containing 18 faces (the maximum number that can sensibly be displayed on a computer monitor). They then select about 6; in the same way, 72 facial textures are presented and they similarly select about 6. Witnesses go on to select a single face that contains the shape and texture with the best likeness: known as the 'best' face. The selected faces are then subjected to a Genetic Algorithm (GA), (refer to Mitchell 1996, for an introduction to GAs). To do this, faces are selected in pairs, first for shape, then texture. Each time, an offspring is produced containing a random mix of coefficients from both faces (uniform crossover). All faces are selected as parents for breeding with equal probability, except the 'best', which is given twice the number of breeding opportunities. The 'best' is also copied directly to the next generation, an elitist strategy which avoids 'damage' through crossover and mutation operators—i.e. as a result of breeding. For all other faces, a small amount of mutation, a probability of 0.1, is applied that replaces each coefficient with a random value. The breeding process is repeated to create a new set of 72 shapes and 72 textures for selection. Witnesses normally look at three generations of faces before selecting one that is saved to disk as the composite image. They are also given the opportunity to enhance the face artistically: for example adding shading, stubble or eye bags, as required.

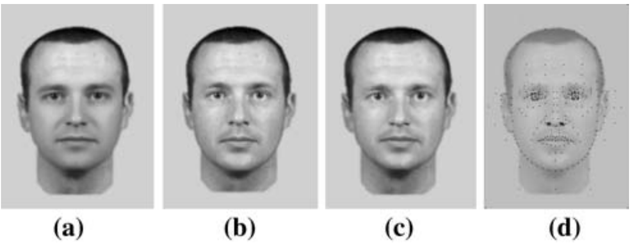


Fig. 1 Representations used to produce a random face with EvoFIT: **a** random shape, **b** random texture, **c** combined shape and texture and **d** location of shape landmarks

An additional utility called the Feature Shift is available to change the shape and position of features in the 'best' face. Such changes might include moving the eyes closer together or making the mouth larger. To do this, the points of the face are moved and a best fit is carried out in the shape model to maintain a holistic code for the face. Furthermore, two additional enhancements have been found to be of benefit to composite construction. First, once a hairstyle has been selected, the set of so-called *external facial features*—the region comprising the hair, ears and neck—are blurred until the end of the evolution. This procedure allows a witness to focus on, and thereby improve, the internal part of the composite, which is important for recognising the face later (e.g. Frowd et al. 2007; Ellis et al. 1979). Second, a set of 'holistic' tools has been developed that allow an evolved face to be re-worked holistically (Frowd et al. 2006). These tools allow alterations such as making a face appear older, more masculine, or even more threatening; eight such dimensions have been implemented to date.

1.2 Building a 'tailored' face model

To build a face model that matches a witness's description of a target, referred to here as a 'tailored' model, descriptive labels were first assigned to each of the 200 reference faces. The descriptive labels used were taken mainly from the 'Aberdeen' Index (Davies et al. 1986), which are also used to classify features in the UK PRO-fit and E-FIT composite systems. Examples include brow thickness, eye colour and mouth shape.

A total of 70 facial features were classified by giving each a whole number value (rating). These features were themselves grouped into seven categories: face shape, hair, eyes, nose, mouth, brows and holistic attributes. Some features were rated along an interval scale—e.g. brow colour, size and thickness. Ratings were typically given a value between 0 and 2; brow thickness for example: 0 = thin, 1 = average, 2 = thick. Some features were categorical and were classified along dimensions that were broadly ordered by feature similarity. For example, for chin shape: 0 = square, 1 = oval, 2 = round, 3 = angular, 4 = triangular. Some features were clearly dichotomous, and were given a value of 0 or 1. Examples include: broken nose, slanting eyes and crow's feet. Faces were also given a value along each of the eight dimensions used in the above holistic tools; these were the mean participant scores obtained from Frowd et al. (2006).

In use, a witness first recalls about three distinctive features of their target face. An error score (the absolute numerical difference) is then calculated between each specified feature and the relevant rating for each face in the database. Those faces with the lowest overall error score are used to build the model. To avoid building models that are too old or too young, an additional classification is made that excludes all faces that are not within about 15 years of the target face, also specified by the witness.

2 Experiment 1—standard versus tailored models

The first evaluation explored the effectiveness of a standard face model, one built using faces of a similar age to a target, compared to a tailored model, built from faces that broadly matched on age as well as other facial characteristics. The basic procedure involved recruiting 24 participants to serve as 'witnesses', with each constructing a single composite using EvoFIT. Half of these participants constructed a face using a standard EvoFIT model, and the other half provided three to four distinctive features from which a tailored model was built; all models contained 72 faces. The targets in the experiment were six members of the Psychology staff at Stirling University.

Witnesses who were unfamiliar with the target faces were recruited: in order to effect this important parallel to police practise, witnesses were drawn from another university department who did not know the targets in the study. Composites constructed by these witnesses were given to psychology staff and final year psychology students, all of whom were familiar with the targets, and who were asked to name the person represented by the composite. Thus, composite quality was assessed by composite naming, and as such can be considered an analogue to real life usage of composites. An additional task was administered that required further participants, also familiar with the targets, to identify each composite from just the inner part of the face, the so-called internal facial features, which are important for naming (e.g. Ellis et al. 1979)—see Fig. 2 (far right). This was carried out to check that composite naming was not being driven exclusively by the presence of hair, which can be an important cue to recognition when the number of potential targets is fairly small: there are about 15 male staff members in the present study there were six staff members from the Psychology Department.

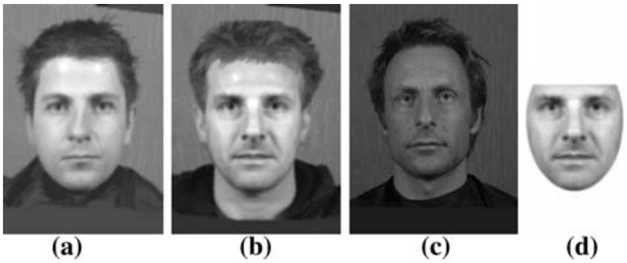


Fig. 2 Example stimuli from Experiment 1. **a** Composite face evolved using a standard model, **b** a face evolved of the same target using a tailored model built from the witness's description of 'spiky hair', 'chiselled jaw' and 'large nose bridge', **c** a photograph of the target face and **d** an example internal features composite used in the evaluation

2.1 Procedure

Laboratory witnesses were tested individually and asked to watch a short video of an unfamiliar member of staff. This was carried out with the knowledge that they would be required to construct a composite the following day. There were six target videos, each of a different member of staff giving directions to a local train station, and lasted for about a minute. Each video was shown to four people, two of whom went on to construct a composite of the target using a standard model, the other two, a composite using a tailored model.

Witnesses returned 24 h later and were met by an experimenter, a person experienced in the use of EvoFIT, who helped them to construct a composite. The experimenter was blind to the identity of the targets during the process of the composite construction. Each person first described the appearance of their target face with the assistance of a Cognitive Interview (e.g. Geiselman et al. 1986). This involved describing the face freely (free recall) and then attempting to recall further details about each feature (cued recall). Those witnesses in the tailored model condition were additionally asked to identify three to four distinctive features of the face and a bespoke model was then built; those in the other condition used the standard model that matched on age.

Both groups went on to evolve a composite with EvoFIT, as described above. Thus, they first selected an appropriate hairstyle that was displayed in blurred form, along with the ears, neck, etc. They then selected 6 facial shapes from a set of 72 shapes, then 6 facial textures from a set of 72 textures. Next, the best face was selected from the best combination of shape and texture. The faces were then bred together and this procedure repeated until a good likeness was evolved, whereupon the blurring filter was switched off. The tool for manipulating specific facial features, the Feature Shift, was offered for use on the best face from the second cycle onwards. The final face was reworked using the holistic tools, to manipulate the face's perceived age, masculinity, etc. Lastly, witnesses were given the opportunity to enhance the face using the GIMP, an artistic package available at no cost (http://www.gimp.org/). The experimenter improved the likeness of the hair, mainly by adding or deleting textured areas, or adding stubble to the chin, as directed by the witness. Composites took about an hour to construct.

The 24 composites were given to a group of 18 participants to name (as described earlier). Twelve additional composites of unfamiliar faces were added to this set to make the task more life-like. Participants were told that many of the composites were constructed of members of staff and were to try to name them. They were also told to expect more than one composite of the same member of staff. Each participant was tested individually and shown the composites sequentially. Afterwards, as a check to verify that the targets were known, they were shown a static photograph of the targets and similarly asked to name them. The order of presentation of composites and target faces was randomised for each person. A further group of 17 participants were similarly shown just the inner part of the face, the so-called internal facial features, and for each were asked to select the most likely person from a list of 12 written names, a list containing the original six names mixed in with a further six staff.

2.2 Results and discussion

Participants were very familiar with the photos of the target set, correctly naming them 97.2% of the time. Composites constructed using the standard face models were correctly named on average quite well, at 35.2%; those using a tailored model were much better, at 54.2%. Example composites are presented in Fig. 2. A two-tailed paired samples t-test applied to the participant data confirmed that the tailored models performed significantly better than the standard variety, $t(17) = 6.03$, $p < 0.001$; the relatively weaker items analysis approached significance, $t(5) = 4.43$, $p = 0.06$. An analysis of the incorrect names mentioned by participants was included, which provides an indication of guessing, and was very similar across both conditions ($M \approx 14\%$). For internal feature composites, identification was significantly better for the tailored model ($M = 37.3\%$) compared with the standard ($M = 24.5\%$), $t(16) = 2.85$, $p = 0.012$, by-subjects. While the items analysis was not significant, a more powerful by-item test was run with the type of task (complete composite/internal features) as a between-subjects factor and the type of model (standard/tailored) as within-subjects. The ANOVA indicated that tailored models were better, $F(1,5) = 7.10$, $p = 0.045$; all other Fs < 2.34, $p > 0.1$.

It would appear valuable then to build a model from faces that matched a target on a few distinctive features mentioned by a witness. The improvement in naming was sizeable, at almost 20%. Note that only five of the six target faces were better named overall when a tailored model was used. It is likely that the sixth target was problematic to construct as this person is in his 60s and there were relatively few faces from which to produce a model (refer to Table 1). As such, the tailored model may have behaved much the same as the standard one. Better performance would be expected for a tailored model had a greater number of older aged reference faces been available.

3 Experiment 2—different sized tailored models

The previous experiment demonstrated benefit for a tailored model, one built from a subset of faces whose features matched a witness's memory of a target face. In Experiment 2, we explored whether further benefit could be obtained using a tailored model built from fewer faces. One would expect smaller models to contain reference faces that better match a description, thus promoting an even better composite. Clearly, there is a lower limit for the number of faces that should be used to build a model, since it needs to generalise well for faces that match the description. Early work on the development of the system used about 30 faces (Frowd 2002), which appeared to work fairly well, and therefore this was taken as the smallest model size. The number of faces used in a standard model was taken as the largest, and rounded down to 70 faces for convenience, and compared with an intermediate one of 50 faces. Thus, 3 tailored models were built from 30, 50 and 70 faces and then evaluated.

A procedure similar to the previous experiment's was employed to compare the ability of these models to construct recognisable composites. Recall that Experiment 1 used a somewhat realistic procedure, with witnesses recruited to construct a single composite of an unfamiliar face (a between-subjects design). For the current experiment, a more powerful within-subjects design was followed whereby an experienced EvoFIT operator evolved all the composites from her memory alone. To allow the resulting faces to be evaluated by adults in general, thus facilitating ease of participant recruitment, the targets were of well-known UK celebrities, and included Gordon Brown (politician), David Tennant (actor), Declan Donnelly (TV presenter), Simon Cowell (TV celebrity), Daniel Craig (James Bond) and Wayne Rooney (footballer). Each of these identities was a familiar face to the operator and was evolved for each size of model (30/50/70), a total of 18 composites.

3.1 Procedure

The EvoFIT operator looked at one of the celebrity photographs for 1 min, to refresh her memory of the face, constructed a tailored face model as before based on three distinctive features of the face, and then evolved a composite. This procedure (including looking again at the photograph) was repeated for all three model sizes and then for the remaining five celebrity targets. The same hairstyle was selected for each target to maintain consistency across conditions. Each composite took about an hour to construct and, working full-time with sensible breaks, the operator produced the set of 18 within 3 days. The order of construction using the three sizes of model was randomised and based on a Latin Square design that allowed all possible combinations to be used.

The composites were printed at 8 cm wide by 10 cm high on A4 paper, one per page. A group of 17 participants volunteered to identify the composites, comprising adult visitors to a newsagents and health club in Wigan, UK. They were presented with each composite in sequence and asked to select the most likely candidate from a list of 6 written names. Participants were tested individually and were self paced. The order of presentation of the composites was randomised for each person.

3.2 Results and discussion

Example composites produced are presented in Fig. 3. Identification was lowest from composites evolved from the largest model ($M = 63.7\%$, $\mathrm{SD} = 32.3\%$); it was about 10% better from the smallest face model ($M = 71.6\%$, $\mathrm{SD} = 19.8\%$) and better again by a similar amount from the intermediate one ($M = 81.4\%$, $\mathrm{SD} = 13.6\%$). A repeated-measures ANOVA of these subject data was significant for model size, $F(2,32) = 5.9$, $p = 0.006$, and simple contrasts of the ANOVA confirmed that the 50 face model was superior to the other two, $p < 0.05$; a two-tailed paired t-test provided weak evidence for a benefit of the 30 over the 70 face model, $t(16) = 1.73$, $p = 0.104$. While the by-items ANOVA was not significant, $F < 2$, relative to the largest model, the intermediate one produced composites that were on average better identified on 5 of the 6 celebrities; a one-tailed t-test also suggested a non-significant trend between these two model sizes, $t(5) = 1.57$, $p = 0.089$.

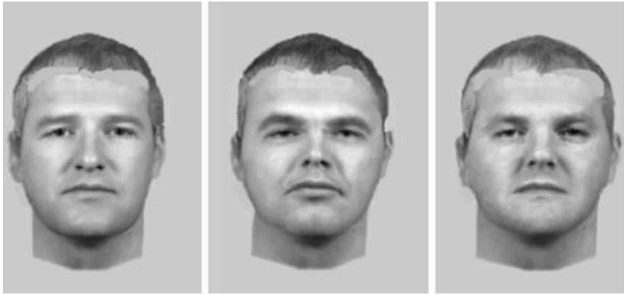


Fig. 3 Example composites of the British footballer, Wayne Rooney. They were evolved from tailored models of size 30, 50 and 70 faces (from *left* to *right*)

In summary, the intermediate sized face model was found to evolve a better quality composite than either the larger or the smaller versions. The expectation was for an inverse relationship between model size and composite quality, but this was only found to be partially true: model performance improved considerably when reducing the build from 70 to 50 faces; this trend did not continue to 30 faces, though there was weak evidence of benefit relative to the largest. It would appear therefore that a somewhat smaller sized tailored model is of value for evolving composites. Note that the variability in composite quality from the intermediate model was considerably less than the largest: the standard deviation was approximately 50% less and approached significance on an F test, $p = 0.081$. Thus, the evidence is that the intermediate sized model not only evolves an overall better composite than the largest, it is also more consistent.

In the final experiment, the intermediate and standard models were evaluated using a more realistic composite construction procedure, as in Experiment 1.

4 Experiment 3—standard versus intermediate sized tailored models

In this experiment, the potential of the intermediate sized model—the one containing 50 faces—was evaluated against the standard aged model. The design of Experiment 1 was followed and involved recruiting participants to act as witnesses and construct a composite. However, an even more realistic design was employed, this time where participants were required to wait 2 days between seeing a target face and constructing a composite (the delay was 24 h previously). The standard model used contained the normal 72 faces, and therefore the experiment tested a smaller tailored model (containing 50 faces) against the one in current police use (72 faces).

The target faces were six photographs of Caucasian international cricket players. These would not be known to non-cricket fans, who would act as witnesses; but would be very familiar to those who follow the game, who would evaluate the resultant composites. Targets were drawn from the England and Australian teams and were Paul Collingwood, Adam Gilchrist, Matthew Hoggard, Simon Jones, Ricky Ponting and Shane Warne. Each face was constructed twice in each condition to produce a total of 24 composites.

4.1 Procedure

The design and procedure of Experiment 1 was followed to construct the composites, except that non-cricket fans were recruited as witnesses; photographs were used instead of videos, and each participant inspected a photograph for 60 s; all participants waited 2 days (rather than 24 h); and the tailored models were built from 50 faces (rather than 72). The participants were staff and students from the University of Central Lancashire (UCLan).

To evaluate the composites, 16 participants were recruited from members of the Wigan Cricket Club. They were presented with the 24 composites in a random sequence and selected the most likely cricketer from a list of ten written names. The list contained the six targets plus four additional cricketers, namely Andrew Flintoff, Glenn McGrath, Brett Lee and Kevin Pietersen. A further 16 cricket fans, staff and students from UCLan, were recruited via a global email. They did the same task using internal feature composites from a list of six written target names.

4.2 Results and discussion

Complete composites turned out to be identified slightly better from the standard ($M = 31.8\%$) than the Tailored ($M = 26.0\%$) model. However, tailored models ($M = 32.3\%$) were much better identified than standard ($M = 23.4\%$) for the more important, internal features test. Example composites from the study are presented in Fig. 4.

A two-way repeated-measures ANOVA failed to find an overall significant effect for neither model type, $F(1,30) = 0.64$, $p = 0.641$, nor composite type, $F(1,30) = 0.11$, $p = 0.746$. However, these factors did interact, $F(1,30) = 4.82$, $p = 0.036$, as there was some evidence that tailored models produced better quality composites than those from a standard model, $p = 0.069$; and (although perhaps less interestingly) composites from standard models were better identified when complete than for the internal features only, $p = 0.099$.

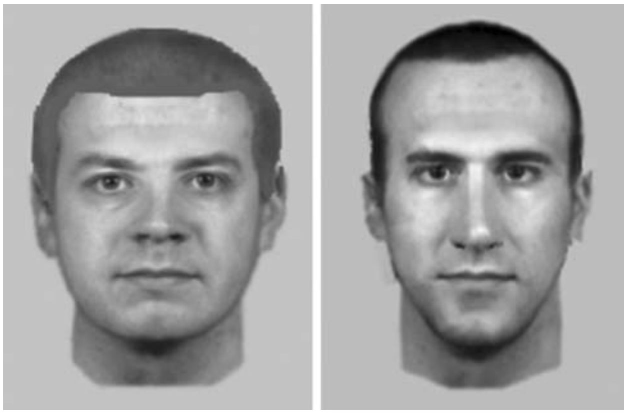


Fig. 4 Example composites of the cricketer Simon Jones who plays for England. A composite constructed from a standard model (72 faces) is on the left; from a tailored model (50 faces), on the right

In summary, witnesses saw a picture of an unknown cricketer and constructed a composite of him 2 days later, a situation that follows real life procedures. While complete composites were not of better quality from tailored models, there was evidence that they were when considering the inner face region alone. The lack of benefit for complete composites was likely to have arisen from limitations in the range of hairstyles. While there were about 500 available, few are modern, and this is likely to have caused mismatching in the complete composite task, reducing discriminability. It is exactly for this reason that we have included analyses of internal feature composites throughout.

5 General discussion

EvoFIT allows a witness to produce a likeness by the selection and breeding of complete faces. While the general approach is more appealing than selecting individual facial features, overall performance is by no means ideal. In this paper, we explored whether tailoring a face model to more closely match a witness's description of a target might help to produce a better quality composite. In the first experiment, a fixed face model containing 72 faces for a given age, a 'standard' face model, was compared with a tailored model. The tailored model was found to be better. In the second experiment, different sized tailored models were evaluated. An intermediate sized model built from 50 faces produced better quality composites than ones built from 30 or 70 faces. In the third experiment, we demonstrated that a tailored model containing 50 faces is generally better than a standard one in a mock witness paradigm.

The results from all three experiments indicate value in tailoring a face model to match a target. The evolving process includes a GA that searches face space iteratively, identifying a sub-space likely to contain the target face. Witnesses select faces that are bred together to produce another set of solutions. If a model is not constrained, it will tend to generate examples with potentially irrelevant features and so produce less accurate solutions: if a target has a thin appearance, for example, then generating wide faces is ineffectual. As Fig. 2 illustrates, building a model from thin faces will preclude wide faces from being generated. The other advantage of this approach concerns the nature of the average face. While the initial faces have random characteristics, they are solutions that radiate from the model's average. With a tailored model, the average should be closer to the desired region of face shape and therefore the initial solutions should similarly be closer.

We have previously evaluated a face model built on the basis of a very detailed description, but this did not appear to work very well (unpublished data). This is perhaps due to a limitation in the number of reference faces available—i.e. 200. Our original approach may have worked better given a larger set of references and thus more candidates from which to select. This is difficult to achieve currently given the appreciable time necessary to locate features in each face and to classify them. Using a few distinctive features appears to overcome this limitation.

One surprising result was that our smallest tailored model did not perform the best, despite the likelihood of it containing the most accurate set of reference faces (the best ranked faces). PCA is a statistical technique that, when applied to face stimuli, captures variations in facial shape and texture. However, to be effective, the model must generalise well to faces that match the description. One possibility then, is that the model of 30 faces was too small to effectively generate a sufficient number of faces for that description; a larger model may generalise better. An alternative explanation is that people's descriptions of faces are not consistent: one person's idea of 'a thin face' may be different to another's. A somewhat more general model, like that built with 50 faces, might give a more consistent match on average.

One possible way to investigate these anomalies might be to evolve faces from standard models of a specific age (like the model used in Experiment 1)—i.e. containing 30, 50 and 70 faces of a given age. In this case, a feature description is not necessary and one would merely explore the generalisation ability of the model. If it is the case that an intermediate size model still comes out the best, this would also suggest that a smaller model might be preferable to larger one for witnesses who are unable to produce a description of a suspect (as is often the case in police work).

There was evidence from Experiment 3 that an intermediate sized tailored model was of benefit relative to a standard one when used realistically. The same experiment also indicated the impact of the external features of the face, a result we have found previously with facial composites (Frowd et al. 2007a). The quality of the hair is likely to be an issue here, indicating the importance of appropriate styles. The PRO-fit composite system from which the hairstyles were taken has a separate database of more modern hairstyles and these should be available for use with EvoFIT in the near future.

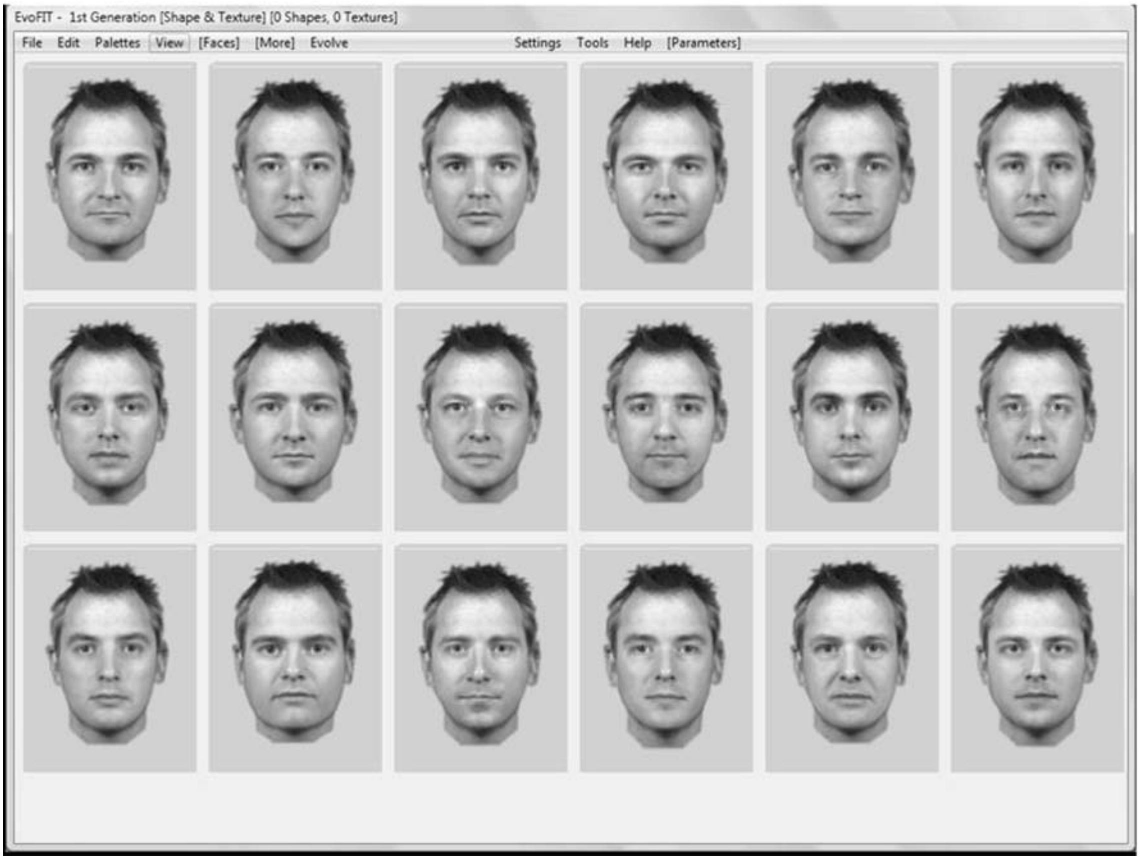
There are a number of potential extensions to this project. Arguably the most pressing is to explore the generalisation ability of the intermediate sized tailored model. All experiments employed six targets, a fairly small number, and so a sensible next step would be a replication with a larger set of targets. Experiments 1 and 3 together show that the tailored approach has value after one or two overnight delays (respectively). It would also be interesting to further investigate model size. While 50 faces was the optimum number here, what about 40 or 60? However, as human evaluations are time-consuming, a better approach might be to use computer simulations initially. A third possibility might be to further refine face selection. At present, the same faces are used to build both shape and texture models. Selecting a face on the basis of a wide mouth, which, while perhaps appropriate for its shape, may not be for texture if other properties are inappropriate (e.g. wrong eye or brow colour). More careful face selection is likely to help. A fourth avenue might be to use verbal information in searchable database systems—in this case, for searching composites against each other. We have been exploring such a notion recently using shape and texture matching of EvoFITs, but it is clear from the current work that verbal cues may be to some extent valuable in this context (possibly along the lines suggested below).

Our work would appear to resonate with other researchers who use face recognition applications premised on PCA. These typically involve a PCA shape model where the reference faces are themselves the targets (e.g. Kirby and Sirovich 1990; O'Toole et al. 1993). 'Recognition' occurs when the (Euclidean or Mahalanobis) distance is minimal between the correct reference and the projected probe in shape space. However, the distance metric for similarity tends to be noisy when the number of principal components is large and therefore a smaller space may yield better results; it is also difficult to know, due to their complex behaviour, which components are the most important. It should be possible to rank the reference faces according to an incoming probe and re-build the space on the fly to compute similarity. Such a procedure would need to be fairly rapid—within a second or two for good performance—but would appear to be conceivable for a manageable face set (e.g. up to 100 candidates) and shape-only matching, the norm.

In summary, the current work sought to improve the underlying model for a face evolving system. It found that a more identifiable composite was produced from a model built with faces that matched on key aspects of a target rather than from a more generic model. The work also found that reducing the model size by about 30% was also valuable, but a further similar sized reduction was less effective. Further, the slightly smaller tailored model would appear effective after a long retention interval. In general, the research suggests that tailoring a face model is of benefit to face evolution with EvoFIT in situations where a witness is able to report a few distinctive features of a suspect's face.

Acknowledgments The work was funded by a grant from the Engineering and Physical Sciences Research Council, EP/C522893/1, and from Crime Solutions at the University of Central Lancashire, Preston, UK. We would like to thank Martin's Newsagent, Winstanley, Wigan, and Wigan Cricket Club, Parsons Walk, Wigan, for allowing participant recruitment for Experiments 2 and 3, respectively.

Appendix: EvoFIT screen shot



References

Annesley J, Leung VL, Colombo A, Orwell J, Velastin SA (2006) Fusion of multiple features for identity estimation. IEE conference on crime and security. IET, London, pp 534–539

Burton AM, Jenkins R, Hancock PJB, White D (2005) Robust representations for face recognition: the power of averages. Cogn Psychol 51(3):256–284

Davies GM, Shepherd J, Shepherd J, Flin R, Ellis H (1986) Training skills in police photofit operators. Policing 2:35–46

Davies GM, van der Willik P, Morrison LJ (2000) Facial composite production: a comparison of mechanical and computer-driven systems. J Appl Psychol 85(1):119–124

Ellis H, Shepherd J, Davies GM (1979) Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition. Perception 8:431–439

Frowd CD (2002) EvoFIT: a holistic, evolutionary facial imaging system. PhD thesis, University of Stirling (unpublished)

Frowd CD, Hancock PJB, Carson D (2004) EvoFIT: a holistic, evolutionary facial imaging technique for creating composites. ACM Trans Appl Psychol (TAP) 1:1–21

Frowd CD, Carson D, Ness H, Richardson J, Morrison L, McLanaghan S, Hancock PJB (2005a) A forensically valid comparison of facial composite systems. Psychol Crime Law 11(1):33–52

Frowd CD, Carson D, Ness H, McQuiston D, Richardson J, Baldwin H, Hancock PJB (2005b) Contemporary composite techniques: the impact of a forensically-relevant target delay. Leg Criminol Psychol 10:63–81

Frowd CD, Bruce V, McIntyre A, Ross D, Hancock PJB (2006a) Adding holistic dimensions to a facial composite system. In: Proceedings of the seventh international conference on automatic face and gesture recognition, Los Alamitos, pp 183–188

Frowd CD, Bruce V, Plenderleith Y, Hancock PJB (2006b) Improving target identification using pairs of composite faces constructed by the same person. IEE conference on crime and security. IET, London, pp 386–395

Frowd CD, Bruce V, McIntyre A, Hancock PJB (2007a) The relative importance of external and internal features of facial composites. Br J Psychol 98(1):61–77

Frowd CD, Bruce V, Ross D, McIntyre A, Hancock PJB (2007b) An application of caricature: how to improve the recognition of facial composites. Vis Cogn 15(8):1–31

Frowd CD, Bruce V, Ness H, Bowie L, Thomson-Bogner C, Paterson J, McIntyre A, Hancock PJB (2007c) Parallel approaches to composite production. Ergonomics 50(4):562–585

Geiselman RE, Fisher RP, MacKinnon DP, Holland HL (1986) Eyewitness memory enhancement with the cognitive interview. Am J Psychol 99:385–401

Gibson SJ, Solomon CJ, Pallares-Bejarano A (2003) Synthesis of photographic quality facial composites using evolutionary algorithms. In: Harvey R, Bangham JA (eds) Proceedings of the British machine vision conference, pp 221–230

Hancock PJB (2000) Evolving faces from principal components. Behav Res Methods Instrum Comput 32(2):327–333

Hancock PJB, Burton AM, Bruce V (1996) Face processing: human perception and principal components analysis. Mem Cogn 24:26–40

Hancock PJB, Bruce V, Burton AM (1997) Testing principal component representations for faces. In: Bullinaria JA, Glasspool DW, Houghton G (eds) Proceedings of fourth neural computation and psychology workshop. Springer, London, pp 84–97

Hancock PJB, Bruce V, Burton AM (2000) Recognition of unfamiliar faces. Trends Cogn Sci 4–9:330–337

Jain AK, Dass SC, Nandakumar K (2004) Soft biometric traits for personal recognition systems. In: Proceedings of international conference on biometric authentication, Hong Kong

Kirby M, Sirovich L (1990) Application of the Karhunen–Loeve procedure for characterization of human faces. IEEE Trans Pattern Anal Mach Intell 12:103–108

Mitchell M (1996) An introduction to genetic algorithms. MIT, London

O'Toole AJ, Abdi H, Deffenbacher KA, Valentin D (1993) Low dimensional representation of faces in high dimensions of the space. J Opt Soc Am A 10:405–410

Phillips PJ, Grother P, Michaelis RJ, Blackburn DM, Tabassi E, Bone JM (2003) Face recognition vendor test 2002. Evaluation report NISTIR6965

Prince SJD, Elder JH, Hou Y, Oleviskiy Y (2006) Towards face recognition at a distance. IEE conference on crime and security. IET, London, pp 570–575

Shepherd JW (1986) An interactive computer system for retrieving faces. In: Ellis HD, Jeeves MA, Newcombe F, Young A (eds) Aspects of face processing. Martinus Nijhoff, Dordrecht, pp 398–409

Sirovich L, Kirby M (1987) Low-dimensional procedure for the characterization of human faces. J Opt Soc Am 4:519–524

Tanaka JW, Sengco JA (1997) Features and their configuration in face recognition. Mem Cogn 25(5):583–592

Tredoux C, Nunez DT, da Costa L, Rosenthal Y (1999) Face reconstruction using a configural, eigenface-based composite system. In: Presented at SARMAC III, Boulder, Colorado

Troje NF, Vetter T (1996) Representation of human faces. Technical report. Max-Planck-Institut, Tubingen