International Journal of Bio- Science and Bio- Technology
Vol. 1, No. 1, December, 2009

Further Automating and Refining the Construction and Recognition of Facial Composite Images

[1]Charlie Frowd, [2]Clare Lee, [3]Anna Petkovic, [4]Kamran Nawaz, and [5]Yasmeen Bashir

University of Central Lancashire
[1]cfrowd@uclan.ac.uk, [2]Clare.Lee@hmps.gsi.gov.uk, [3]petkovic.anna@googlemail.com, [4]MNawaz@uclan.ac.uk, [5]YBashir@uclan.ac.uk

Abstract

Bringing a criminal to justice is a labour intensive process. In the current paper, we explored ways of reducing the police time involved with the construction and identification of facial composites (these are images of wanted persons made by witnesses and victims of crime). A software system called EvoFIT was used that 'evolved' a composite by the repeated selection and breeding of complete faces. In the first part, a standalone version of EvoFIT was designed and evaluated in the laboratory. This performed similarly to the full system that normally requires several hours of a police officer's time. It was also found that composite quality did not change overall, although was more variable in one of the measures used, if users were asked to make fast rather than slow face selection judgements. In latter work, a small database of composites was built that could be used to search for matching identities. It was found that pixel intensity (texture) information was valuable for composites produced from a traditional 'feature' based system, but feature shape information was valuable for composites produced from EvoFIT. The results show promise for the automated construction and identification of composite images.

1. Introduction

Bringing a criminal to justice involves considerable human resources. This is particularly true when collecting evidence: descriptions of events and persons, identity parades, DNA, fingerprints, CCTV footage, facial composites, etc. In the case of facial composites, witnesses must be interviewed, to obtain a verbal description of the face, and interact with computer software or a sketch artist to produce an image the face. Later, this image is shown to other people in the hope that someone will name it to the police and provide additional lines of enquiry. Therefore, both the method of constructing a composite face (from witnesses) and procedures used to identify it (showing the composite to members of the public) are time-consuming and labour intensive.

There are two broad approaches for constructing facial composites. The first requires witnesses to describe the appearance of the criminal and to select individual features from a kit of parts – hair, eyes, noses, mouths, etc. The UK uses two computerized systems to do this, E-FIT and PRO-fit [[1]], although there are many such systems available elsewhere [[2]]. The second approach requires witnesses to repeatedly select from arrays of complete faces, and the system itself provides alternatives based on these selections, to allow a composite to be 'evolved'. The process used is a working example of Charles Darwin's theory of *evolution by artificial selection.* To the authors' knowledge, there are three such systems in existence: EvoFIT [[3]] and EFIT-V [[4]] in the UK, and ID in South Africa [[5]]. All systems require several hours of a police operative's time to construct the face and complete the paperwork necessary to record the evidence. Composites are then circulated within the force, and more

generally in the media, to obtain the relevant identity. For this reason, composites are generally restricted to serious crime, such as indecent assault and murder, rather than to more common but less serious crime such as antisocial behavior or petty theft. (Although composites are sometimes constructed in criminal investigations involving particularly prolific offenders.)

One part of this process that has been overlooked, and where computer algorithms may be of value, is the use of composites as a mechanism to search for other composites of the same identity. Police forces tend to accumulate composites over time, and may not be aware that there are multiple images drafted of the same person. Therefore, a tool that allows law enforcement to reliably detect repeat offenders based on composite images would be valuable. Also valuable would be a database of existing composites that could be regularly interrogated as new faces are constructed by witnesses.

The focus of the current work was twofold. Firstly, we explored the possibility of a composite system that could produce an identifiable composite without being controlled by a police software operator. A version of the EvoFIT system has been developed with this in mind. The work explored the effectiveness of such a system in comparison to the normal version of EvoFIT that does require manual assistance. We also explored whether good quality composites would be produced from EvoFIT in general when used in a more timely fashion, simply by asking users to make rapid selection from the presented face arrays. If good performance were found using this procedure, then this could be applied to police work, similarly reducing time. Secondly, we investigated the feasibility of a searchable composite database. An initial study is presented to test the effectiveness of several potential metrics for searching a database of composites produced from two leading face production systems.

Figure 1. Example celebrity composites from a typical 'feature' system. Each was constructed from a user's memory. The identities are listed in Section 8.

1.1. Facial composite systems

There are two broad approaches to composite production, those based on the selection of individual facial features [[1]] and those based on the selection of complete faces [[3]-[5]]. These are described below.

1.1.1. Feature-based composite systems: The most popular composite systems are computerized and contain a large database of individual facial features. These facial parts are cut electronically from photographs of faces and classified. In use, a witness describes the criminal's face to a software 'operator', who then uses the classification system to locate examples that match the witness's memory of the face. Computer graphics technology allows

features to be resized and repositioned as required. Example composites from this type of technology are presented in Figure 1.

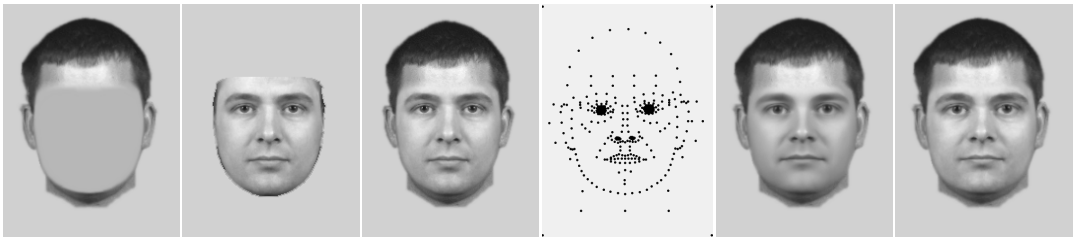
There are two fundamental problems with this approach [[6]-[8]]. Firstly, the basic feature-by-feature mechanism used to construct a composite is contrary to the way in which faces are naturally perceived, as complete entities [[8]], and not as lists of individual features. Secondly, many witnesses are unable to describe a face from their memory in sufficient detail for the classification system to be effective, and are thus denied the opportunity of constructing a face in the first place.

1.1.2. Recognition-based composite systems: A new type of system has emerged that is based more on our superior ability to recognise faces rather than to describe them. The basic approach is to present screens of complete faces to user. The person selects a few that bear an overall resemblance to the criminal's and these are bred together by combining facial characteristics. Repeated a few times allow a composite to be 'evolved'. The process is therefore based on whole face selection rather than selection by facial parts.

There are several systems of this kind [[3]-[5]]. Each one uses an underlying model to generate faces. The models are constructed using Principal Components Analysis, or PCA, which is a statistical technique that extracts the major axes of variation in a data set. In the current application, the dataset typically comprises of carefully photographed frontal-pose images of faces of a given age, race and gender. PCA provides a set of reference faces (Eigenvectors) and coefficients (Eigenvalues) that allow the original items to be r; here, the coefficients are assigned random values in order to generate novel faces.

The EvoFIT system has been extensively developed and evaluated [[21]], and will be the main focus of the current work. The method used to generate faces with this system is described in detail in Frowd et al. [[3]]; it is presented in summarised form here. The underlying EvoFIT model is in two parts, each built separately using PCA. The first is *shape* and describes the shape and position of individual features on the face; the other is *texture*, for the colour of the eyes, brows, mouth and overall appearance of the skin. The shape model is built from so-called 'landmark data', files of 298 co-ordinate points that define the outline of features of the face. In order to build the texture model, each reference face is morphed to a standard shape – normally referred to as a *shape-free* face defined as the average of the landmark data – so that the facial features are co-aligned. A second PCA is carried out on these greyscale pixel values.

In practice, only the *internal facial features* are contained in the texture model, as illustrated in 2(b), the central region of the face encompassing the brows, eyes, nose and mouth. To generate a random face, a random texture is blended into a set of *external facial features* – hair, ears and neck – as selected by a user. This provides an image that is then morphed (distorted) to provide a final, random face using a random shape. Example images illustrating the process are presented in Figure 2. EvoFIT can construct faces of white, black and other ethnic groups; there are separate PCA models for these, which themselves are subdivided by age: see [[21]] for further details.



(a) (b) (c) (d) (e) (f)

Figure 2. Production of a random face: (a) external facial features; (b) random texture; (c) blend of (a) and (b); (d) co-ordinates of random shape; (e) representing (d) in facial form, to see the shape more clearly; (f) image distortion of (d) to (c) to give a random face.

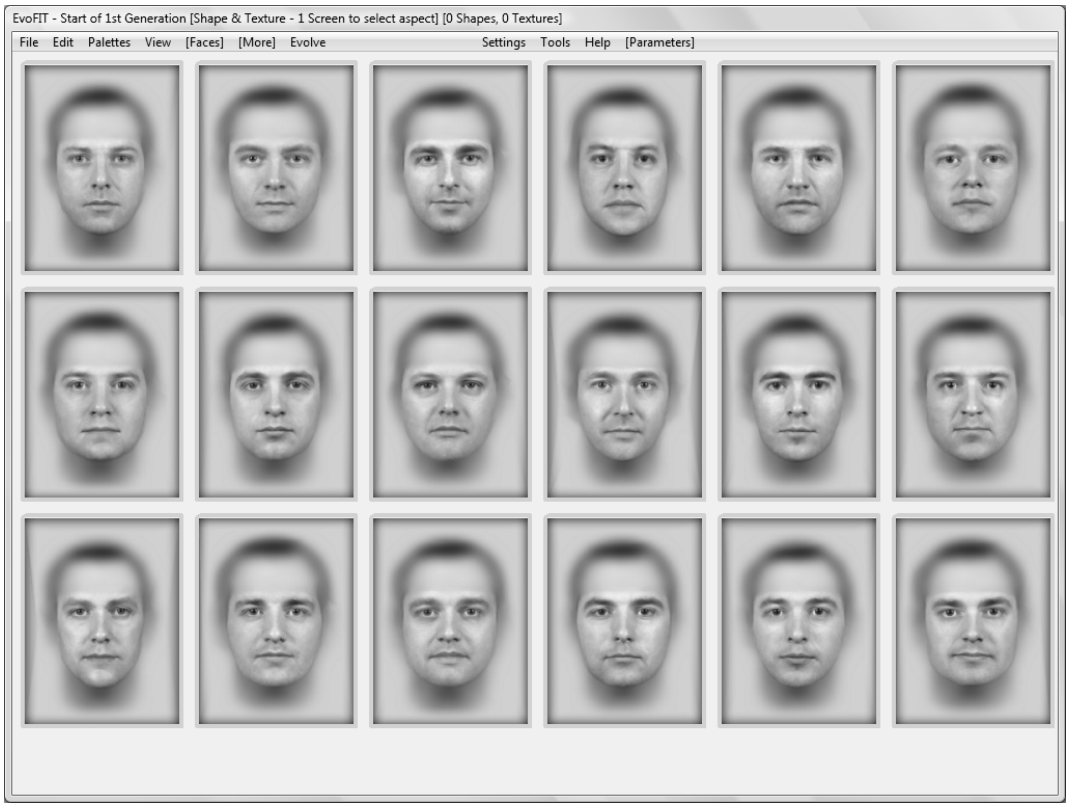


Figure 3. An example EvoFIT screenshot. The external region is blurred, as shown here, to help a user to focus on the important central region of the face. After evolving, the blur is disabled to reveal all parts clearly.

EvoFIT presents users with screens of 18 faces, the maximum number that can be sensibly displayed on a computer monitor. The procedure for constructing an EvoFIT is fairly complicated in order to produce an identifiable face. In brief, witnesses first choose a set of *external facial features* – hair, ears and neck – to be displayed on each face. Next, they select a single face that best represents the target's *facial aspect* ratio, the relative width and height of the face. Following this, they select the best overall likenesses from four screens of facial shape, followed by four screens of facial texture; witnesses select two faces per screen up to a maximum of six. Combinations of selected shape and texture are then presented and witnesses select the best single face. These selections are then bred together using a Genetic Algorithm that combines shape coefficients with uniform crossover and a mutation rate of 0.05; the same is carried out for coefficients of the selected textures. The process is normally repeated once more to produce the 'composite'. Note that users are encouraged to select faces based on their overall appearance, rather than by specific facial features. The aim here is to

produce a face with an overall appearance as close as possible to a target (rather than one with accurate facial parts).

Two additional procedures have been found to be helpful for evolving a face. Firstly, when the external features have been selected, a Gaussian (blur) filter is applied to this region of the image. This filtering allows witnesses to focus on the internal parts of the face, making it more identifiable, which is important since other people will rely on the same region when attempting to recognise the composite later – e.g. [[9],[21]]. A screen of typical faces presented in this way is shown in Figure 3. Secondly, software tools were designed [[10]] to allow an evolved face to be enhanced along a number of psychologically-useful scales: age, attractiveness, masculinity, weight, etc. A software 'Shape Tool' is also available towards the end of the process that allows the shape and position of features to be manipulated on demand. Example composites are presented in Figure 4. These have been evolved using the basic system plus enhancements.

In recent tests, EvoFIT produced composites that were correctly named with a mean of 25% following procedures that mirror those of real witnesses as far as possible [[17]]. This includes the normal situation where a witness is required to wait 2 days between seeing a target and constructing a composite of the face. In comparison, correct naming levels for 'feature' composites produced under the same conditions are typically 5% or less [[2],[6],[7],[17]].

Figure 3. Example celebrity composites from EvoFIT constructed from a user's memory. The identities are listed in Section 8.

2. Automating composite construction

In this part, the standard version of EvoFIT, where an operator controlled the software, was compared with a standalone version, where no such assistance was required. We also evaluated the standard version of EvoFIT where users were asked to more rapidly select faces, in an attempt to encourage holistic face processing, improve performance and reduce the overall construction time. For both, police procedures for constructing the face [[1]] were followed as closely as possible in the laboratory (for the operator-assisted system).

2.1. Standalone EvoFIT

A standalone version of EvoFIT has been designed, involving written instructions appearing at the bottom of the screen to guide a user through each stage of the process. These instructions typically require a user to make selections from the presented face array and to click the 'Next' button to continue. We attempted to mirror normal construction procedures with EvoFIT as far as possible, including the latest enhancements: external features blurring and Holistic Tools. Due to limitations in time, however, two aspects were not implemented in

the standalone version. Firstly, selection of the appropriate face model was done by the operator. Secondly, there was no Shape Tool. This tool appears to be quite useful and so we expected composite quality to be less than optimal. Note that, to make the conditions in the experiment as equivalent as possible, the Shape Tool was used for neither version.

2.2. System evaluation

Two stages were required to investigate whether the full and standalone versions of EvoFIT were equivalent. In the first, composites were evolved by volunteers with or without an operator; in the second, they were given to other people to name.

A set of 12 photographs of UK international-level footballers were used as targets. This enabled non-football fans to be recruited as 'witnesses', so that the targets would be unfamiliar to them, as in real life. The images were of Emmanuel Adebayor, Nicolas Anelka, Ashley Cole, Joe Cole, Jermaine Defoe, Didier Drogba, William Gallas, Frank Lampard, Gary Neville, Paul Scholes, Alan Smith and John Terry. Images depicted a front view of the face, in colour, without glasses and with at most minimal facial hair.

Each of these 12 targets was constructed once by a participant-witness working with an operator and once by a different participant-witness who worked alone. The assignment of targets to participant-witnesses and construction type (operator / standalone) was randomized. Each person looked at a target for 30 seconds. Then, those working with the operator followed procedures used in police work, which are described in full in [[1]]. In brief, they received a Cognitive Interview (CI) to help them recall details of the face. To do this, they were asked to visualise the appearance of the face and then to describe it in an uninterrupted way. Next, they were asked to focus in more detail on each facial feature and attempt further recall. The operator made written notes while this description was being produced.

After the interview was complete, EvoFIT was started with the correct age Caucasian male database. Witnesses selected an appropriate set of external features and evolved a single composite as described in 1.1.2. Those who worked on their own first wrote down a description of the face, and then followed the on-screen instructions in the standalone version. A total of 24 composites were evolved, 12 with an operator and 12 without.

Eighteen football fans were recruited to evaluate the quality of the composites. They were told that they would be shown composites of well-known UK international-level footballers and to try to name them; also, that there were repeated identities in the set. Each of the 24 images were presented in sequence and participants attempted to provide a name where possible. The order of presentation was randomized for each person.

2.2.1. Results: Ten of the 12 EvoFITs in each condition were correctly named by at least one person. Composites were named 21.3% correct when constructed via an operator and slightly less, 17.1%, when witnesses worked alone. This difference approached significance using a two-tailed t-test, $t(17) = 1.7$, $p = 0.095$; the items analysis was not significant, $t(11) = 0.7$, $p = 0.449$. Thus, there is a slight benefit for the operator-assisted images. An analysis of incorrect names was carried out as this can provide a further indication of composite quality (guessing); scores also differed little whether faces were produced with an operator, 4.2%, or without, 1.9%.

2.3. Discussion

Faces constructed in criminal investigations are labour intensive for police personnel. In the current work, composites from a version of EvoFIT not requiring assistance from an operator were named about 4% less than with one, a small but reliable difference in the by-

subjects analysis. Naming levels were about 17% from the standalone version, and thus were fairly good anyway.

A reason for the decrement in performance for the standalone EvoFITs is likely to be increased task difficulty: these witnesses not only had to read and follow the instructions on the screen but also think about which faces to select. Witnesses in the other condition were verbally guided through the process by the operator. In fact, many of the witnesses using the standalone system commented that the procedure was hard. Thus, in spite of some software pilot testing, task difficulty remained high.

One way to simplify the procedure would be to provide instructions in a verbal rather than written form. We have tried this already in a simple version of EvoFIT that is installed in the Sensation Science Centre, Dundee [[11]]. Anecdotal evidence from this exhibit is that such prompts are very effective. We plan to add these in due course and carry out a further evaluation.

An additional improvement would be to implement the Shape Tool in the standalone system. As mentioned above, this enables the shape and position of features to be modified on demand. It would seem that this tool is fairly effective, given that naming levels from the operator-assisted composites were somewhat lower than those found elsewhere using a similar design which had included it – e.g. [[9]].

In general then, results were positive for the standalone version of EvoFIT. In the following experiment, we explored whether an effective composite could be evolved in a shorter time than normal, by asking users to make rapid face selection judgements.

2.4. Face selection speed

Current procedures with EvoFIT ask users to base selections on the overall appearance of the face (and not on individual features). Anecdotal evidence is that some users do this quickly, while others take much longer. But, is this factor important for good performance? On the one hand, it is known that face recognition processes occur automatically and rapidly, taking around 250ms, and therefore selection time may not be important. The reader is referred to [[12]] for a general overview of this area. On the other, faces are recognised as complete entities rather than by their individual features – e.g. [[13]]. As mentioned above, this is perhaps one reason why face construction using a 'feature' system is not an effective interface to human memory. Similarly, users of EvoFIT appear to have a tendency to compare individual features when they select from arrays of faces and therefore (non-optimally) may make judgements based on individual features. Simply asking them to make rapid face recognition judgements may encourage whole face selection and thereby promote an overall more recognisable composite.

In the following experiment, users were asked to select faces either quickly or slowly. The same methodology as 2.2 was followed, except there was a longer, more realistic delay of two days between a target face being seen and a composite being evolved. Targets were photographs of five male members of Psychology staff at the University of Central Lancashire, Preston, UK. Each was photographed as before.

Participant-witnesses were 20 students at the University of Central Lancashire. All did not study within the School of Psychology, to allow the target faces to be unfamiliar (as in police work). The same procedure as 2.2 was used (normal construction using an operator and a CI), except that each participant-witness looked at a target face for 60 seconds and was asked to select faces either slowly or quickly (with random assignment). Each of the five targets was constructed twice in each face selection condition, also with random assignment, after the 2 day delay, to make a total of 20 composites. Examples are presented in Figure 4.

Figure 4. Composites constructed by two different people following instruction to select faces either slowly (left pair) or quickly. The image on the right of each pair are the internal facial features that were used as part of evaluating composite quality. The target face used here is shown on the far right.

Thirty six additional participants evaluated the quality of the composites. These participant-evaluators were staff and students from Psychology at the University of Central Lancashire, to have good familiarity with the target identities. Half of these participants were shown the 20 composites and asked to provide a name for each where possible. They were then presented with the five target photographs and asked to name those, as a check that the faces were well-known. The other group did the same, but were presented with internal features of the composites and a list of written target names from which to select. This second task is a useful check to guard against composites being named on the basis of hair alone, or lack thereof, which can potentially happen (and is not of primary interest here). Tasks were self paced and, to limit systematic bias, composites were presented in a different random order for each person.

2.6. Results

The photographs were named very well, at 80.0%, thus indicating that the participant-evaluators did indeed know the target set well. The composites were named at 33.3% correct in the 'slow' condition and slightly higher, at 36.1%, in the 'fast' condition. In a two-tailed t-test carried out on the participant naming data, this difference did not differ significantly, $t(16) = 0.7$, $p = 0.508$. For the second naming task using the internal features, correct naming was somewhat higher for slow selection, 52.2%, than for fast, 44.4%, but again this difference was not reliably different, $t(16) = 1.5$, $p = 0.172$.

Inspection of the variability of the naming data revealed a consistent standard deviation by selection type for naming of complete composites, but the standard deviation was about twice as high for internal features naming in the fast (SD = 19.2%) relative to the slow condition (SD = 10.6%); the kurtosis was also fairly high and positive in the fast condition, $G_2 = 1.94$, indicating the presence of a peaky distribution. F-tests were run to investigate this further. These found that the variance of the naming scores was not significantly different for complete images, $p = 0.916$, but there was significantly more variability for fast than slow selection using internal feature composites, $p = 0.019$. Therefore, while neither type of instruction promoted an overall more identifiable composite, asking participant-witnesses to make rapid face selection judgements did produce a more variable outcome.

2.7. Discussion

The present study sought to investigate whether asking participants to make speeded face perception judgements would promote a more identifiable face from EvoFIT. However, the results did not support such a mean improvement in composite quality, with both naming tasks indicating equivalent performance for the two types of face selection. Note, though, that there is a (non-significant) trend in the internal features data that clearly favours slow selection; similarly, in the same data, instructions to make slow judgements prompted less variable results than those to make fast judgements.

It was predicted that instructions which encouraged rapid selection of faces would promote a more natural, holistic procedure. The notion here is that rapid face selection would tend to prevent users from looking in detail at the individual parts of the face, a procedure that is unlikely to be optimal for face recognition [[13],[14]]. The number of composites constructed here, and the number of participants used to evaluate them, are normally sufficient to detect medium-to-large effect sizes, which would be of practical value (i.e. not just small and theoretically interesting). However, there was no evidence that this was the case in terms of the mean level of correct naming. In contrast, prompting faster face selection resulted in composites that were more variable in quality. This in itself would suggest that some users benefited from this instruction (naming rate was higher than normal) while other users were put at a disadvantage (naming rate was lower). Overall, the study indicates that it is perhaps more prudent to ask users to take their time when making selections with EvoFIT, as this would promote a more consistent end result. However, it would be sensible now to see if this basic result replicates. As part of that follow-up work, a third condition should be included whereby participant-witnesses are not given any instructions about face selection rate. This addition would allow comparison against the current procedure, where no specific instructions are given.

In the next part of the work, another way to automate the composite process is explored: the ability to search composites against each other.

3. A searchable database for composites

One consequence of a standalone system is that it can be deployed much more often than normal, potentially resulting in a large number of composites. Of course, composites are only ever valuable if attempts are made to identify them: by circulating them within a police force, or by publishing them on TV or on wanted person's web pages. Thus, a standalone system is likely to create more police work unless managed properly. A solution to this potential problem is to build a database of composites and allow them to be identified by searching them against each other. This basic idea has been applied to mugshots (photographs) of suspects – e.g. [[18],[19]].

The effectiveness of such a searchable database is explored here. This is considered for a typical 'feature' system, PRO-fit, a 'standard' from which to compare, and for EvoFIT. In the following, different metrics are discussed and a formal evaluation is presented to indicate the best method and system for searching.

3.1. Metrics for searching

There are many methods available to search for facial information. For a composite database, this could include the facial shape information (the coordinate landmarks defining the outline of facial features) and the facial texture information (the greyscale pixel values in the image). The simplest method for establishing similarity is to compare pairs of composites using the root mean-square error (RMSE) measure by shape or by texture. One compares corresponding landmarks or pixels and computes the square-root of the average of the square

of the differences across the set. Comparing all pairs of items in this way within a database produces a similarity matrix; ultimately, error scores that are below some kind of threshold can be taken to indicate an 'identity match'.

This type of similarity estimation can be carried out in the *physical* space (co-ordinates and pixel values) and applied to composites from both feature and recognition systems. While the latter system uses an inherent landmark coding mechanism, as illustrated in Figure 2(d), co-ordinates need to be established for feature composites, a manually intensive procedure requiring about 20 mins per face. Comparisons can also take place in the face *coefficient* space for EvoFITs (but not for PRO-fits, as there are no underlying coefficients available.) Recall that for EvoFIT, each face has a small number of Eigenvalue coefficients that are used to represent faces in terms of facial shape and texture. While the shape and texture information in the physical space is large, only 72 floating point numbers are required in the coefficient space. Thus, the compact code produced by PCA is potentially ideal for carrying out a large number of similarity estimations.

There are a range of metrics that can be easily applied to the coefficient space. The simplest is the Euclidean Distance (ED), which has the same algorithm as RMSE above, but uses either the 72 shape coefficients (rather than the 298 landmarks) or the 72 texture coefficients (rather than the 5,000 or so pixels). A slightly improved version of the ED is the Mahalanobis Distance (MD) [[16]]. This measure is similar to ED except that each squared difference is multiplied by the variance accounted for by the relevant Eigenvector. This is done because some Eigenvectors account for more variance in the dataset than others and so using MD will result in pairs of items being considered more similar to each other if they have closer matching values along such dimensions.

A third potential metric is Angle, often used in the document searching domain as an alternative to ED [[20]]. Angle metrics consider the shape or texture coefficients as vectors and the mathematical cosine function is often used to compute the angle between them; as for ED and MD, lower values indicate a closer match.

In the following, the RMSE measure is used to evaluate the effectiveness of EvoFITs and PRO-fits in the physical space; and, the ED, MD and Angle for a set of EvoFITs in the coefficient space.

3.2. Method of evaluation

We were interested in finding the best metric to search EvoFITs (Euclidean Distance / Mahalanobis Distance / Angle) and which data type (Shape / Texture / Both Shape and Texture) to use in general. To achieve these objectives, a set of 12 composites were extracted from past research projects that had been constructed from a person's memory after seeing a photograph of the face. The requirements for selecting these images were that each had to have been constructed from the same face model, so that the shape and texture coefficients had an equivalent Eigenface mapping, and that a PRO-fit was also available, for evaluation in the next part. The set comprised of snooker players (Ken Doherty, Stephen Maguire, Alan McManus, Shawn Murphy, Neil Robertson, Mark Selby and Ronnie O'Sullivan), international footballers (David Beckham, Alan Smith and John Terry) and other celebrities (TV host, Anthony 'Ant' McPartlin; and UK pop singer, Will Young). Each of these had been made from the 30 year EvoFIT white male model, and also from PRO-fit.

A second set of composites was constructed of each of these identities by EvoFIT (30 year face model) and by PRO-fit. To do this, an experienced user looked at a photo of the relevant identity for 1 minute and then constructed an EvoFIT using the procedure outlined in 1.1.2; he then looked again at the photo for the same amount of time and used PRO-fit, 1.1.1. This resulted in 24 images constructed by witnesses in past studies, and 24 by an experienced

operator. While the EvoFITs already had landmark data available, which EvoFIT automatically produces, co-ordinates were manually located for the PRO-fits. Example composites are presented in Figure 5.

Figure 5. Composites of the UK footballer, David Beckham, used as part of evaluating the searchable database. The left pair is EvoFITs, the right PRO-fits; the image on the left of each pair was made from witnesses, the right one from the experienced user.

The first analysis used the EvoFITs. We asked the question as to which metric and data type were best for searching? A database was constructed containing the shape coefficients of the 12 EvoFITs constructed by the witnesses. Then, the first EvoFIT constructed by the experienced user was compared against each item in the database and an ED score computed. These scores were then ranked from 1 to 12 (best to worst) with respect to the relevant target. The ED was next computed for each composite constructed by the user. To increase the statistical power of the analysis, these calculations were repeated by swapping over data sets (i.e. composites constructed by the witnesses were used as probes for composites constructed by the user).

This procedure was repeated for the two other metrics – Mahalanobis Distance (MD) and Angle – and also for metrics for the texture coefficients. Next, the metric scores were averaged together for shape and texture for each identity and the data re-ranked to provide a combined score, which we refer to as data type 'Both'. Finally, the above was repeated for the RMSE metric for the coordinate (shape) and pixel (texture) information; this relates to the *physical* space (rather than the *coefficient* space) and for which we refer to as 'Image'. The MD metric and the Both data type were expected to be superior (i.e. to have the lowest overall ranking scores).

In the next part of the analysis, EvoFITs and PRO-fits were searched against each other. This analysis used the co-ordinate (shape) and pixel (texture) information. All possible combinations using the RMSE measure were considered: EvoFITs to EvoFITs, PRO-fits to PRO-fits, EvoFITs to PRO-fits and vice versa. The expectation was that matching based on the same type of composite system would be best.

3.3. Results

The mean rank score by data type (Shape / Texture / Both) and metric type (Image / Euclidean / Mahalanobis / Angle) for the EvoFITs are presented in Table 1. Note that scores are out of a possible 12 and that lower values represent better matches between corresponding

composites. It can be seen that the Image (physical) was the best metric and Angle (coefficient) was the worst; also, that differences by data type were fairly small.

To further increase statistical power, two inferential analyses were carried out. The first compared the main metrics – Image, ED and Angle – and the second compared the two similar metrics, ED and MD. For the former, a repeated-measures ANOVA approached significance for metric type, $F(2,46) = 2.6$, $p = .083$, but was significant for neither data type, $F(2,46) = 2.0$, $p = .151$, nor the interaction, $F(4,92) = 0.6$, $p = .569$. Simple contrasts of the ANOVA provided weak evidence that Angle was significantly worse than Image, $p = .051$; no other reliable contrasts were found. For the latter, there were no significant differences between ED and MD for the main effects or interaction, $Fs < 1.1$, $p > .332$.

Table 1. EvoFIT-to-EvoFIT searching by data type (columns) and metric (rows). Scores are mean rankings out of 12; lower values represent better matches.

Metric	Shape	Texture	Both	Mean
Image	5.00	5.50	4.54	5.01
Euclidean Distance	5.75	6.08	5.08	5.64
Mahalanobis	6.33	5.96	5.46	5.92
Angle	7.46	6.33	6.92	6.90
Mean	6.90	6.15	6.19	6.41

Table 2. Searching for EvoFITs or PRO-fit composites by data type (Shape / Texture / Both). Values are mean rank with a maximum of 12: lower scores represent better matches.

	Shape	Texture	Both	Mean
EvoFIT	5.00	5.50	4.54	5.01
PRO-fit	6.71	4.83	5.33	5.63
Mean	5.85	5.17	4.94	5.32

The second analysis involved EvoFITs and PRO-fits. The RMSE *physical* (Image) metric was used, and converted to rank order data (as above). This involved co-ordinate values for shape and pixel values for texture. The mean ranking scores are presented in Table 2; as above, lower scores indicate better matching and all are out of 12. Note here that the scores for EvoFIT are the same as those for 'Image' in Table 1 (these are the same data). Performance was slightly better for EvoFITs searching other EvoFITs, and when both shape and texture information was combined. This time, neither data type, $F(2,92) = 2.2$, $p = .118$, nor composite type, $F(1,46) = 0.4$, $p = .509$, was significant. However, the interaction was, $F(2,92) = 3.4$, $p = .036$, as (a) for EvoFITs, there was some evidence that the combined data type (Both) was better matched than by texture, $p = .057$; and (b) for PRO-fits, shape matching was significantly worse than matching by texture, $p = .039$, and by Both, $p = .007$.

Next, we compared same- and cross-system searching using the more sensitive root-mean-square error (RMSE) measure. We used composites of the same type – EvoFITs to search EvoFITs, PRO-fits to search PRO-fits – with cross-composite searching – EvoFITs to search PRO-fits, and vice versa. As can be seen in Figure 6, shape matching by the same type of composite was preferable, and matching EvoFITs to EvoFITs was best overall. The ANOVA was significant by composite type, $F(1,92) = 484.6$, $p < .001$, as EvoFITs were matched overall better than PRO-fits, and by search type, $F(1,92) = 103.2$, $p < .001$, as matching using the same composite technology was also best. However, these factors interacted with each other, $F(1,92) = 107.1$, $p < .001$, since these main effects were consistent except for cross-composite matching where there was no significant difference by system, $p = .893$. Note that

the data for texture followed the exact same pattern of effects and, for brevity, are omitted here.

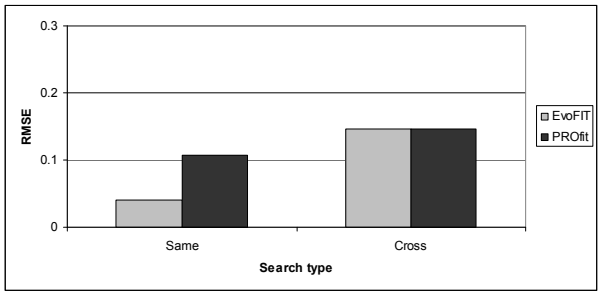


Figure 6. Shape matching for composites of the same (EvoFITs-EvoFITs, PRO-fits-PRO-fits) and different type (EvoFIT-PRO-fit and vice versa).

The above analyses used mean scores. Another method of assessment is based on the number of correct matches for low ranking items. As the size of the database used was fairly small, only 12 items, a fairly strict analysis was carried out such that a 'success' was taken to have occurred for matches that were ranked first, second or third. The result of such a 'podium-position' analysis is presented in Table 3. It can be seen that there is little difference overall by either system or data type. However, these factors appear to interact with each other such that, at best, about 40% of the time, shape matching was effective for EvoFIT, and texture matching for PRO-fit.

Table 3. Number of correct matches ranked in the first, second or third position. Scores are out of a maximum of 24.

	Shape	Texture	Both	Mean
EvoFIT	11	7	8	8.7
PRO-fit	6	10	7	7.7
Mean	8.5	8.5	7.5	8.2

3.4. Discussion

A searchable database that could accurately identify people from their composites could be useful for law enforcement. In the current work, a small set of composites were constructed from two modern systems, PRO-fit and EvoFIT, and were used to evaluate the effectiveness of a number of metrics and data types. Results indicate that the angle measure performed somewhat worse than the Image (RMSE) type; and, that there were no reliable differences by the type of information used to search: shape, texture or both. In an analysis involving non-coefficient (Image) data alone, there was some evidence that the combined data type (Both) was better than texture for EvoFIT, and that shape matching was reliably the worst for PRO-fit. Using the more sensitive RMSE measure than rank, matching scores were better for composites of the same type and that this was even better for EvoFITs. In the final part, there were more correct matches ranking in the top three with EvoFITs for shape and with PRO-fits for texture.

While a fairly small data set was used, a pattern appears to be emerging. Firstly, the angle metric is likely to be ineffective in this application. Angle-based measures are sometimes used in the document retrieval area [[20]], whereby the vectors (coefficients) are based on frequencies of words found within a document. As such, vector angles tend to be driven by

high frequency word counts, which are often a feature in that domain; for PCA coefficients, values are normally bound (e.g. within +/- 1) and thus do not have extreme values.

Secondly, for searching within the same technology, texture information appears to be more valuable than shape for searching PRO-fits; but, for EvoFITs, the approaching significant advantage of the combined metric (Both) over texture, and the high number of top three matches for shape together suggest value in shape information for EvoFIT. These data no doubt reflect biases on the part of the composite constructors: the focus is primarily on texture information for PRO-fit, and (initially) shape for EvoFIT. They do make intuitive sense. For PRO-fit, the emphasis is very much on selecting individual features (predominantly texture) but less on placement (predominantly shape). In contrast, shape information is specifically probed during the construction of an EvoFIT, and witnesses seem to be able to take advantage of this (and are better at selecting that aspect of the face than texture). This conclusion arguably explains to some extent why cross-system matching was ineffectual: information correct in one type of technology was more error-prone in the other.

4. General Discussion

To our knowledge, the current work is the first that has looked formally and in detail at the feasibility of more automated mechanisms for constructing and identifying facial composites. The standalone version of EvoFIT driven by written prompts, performed almost as well as the full system controlled by a software operator. This was in spite of users reporting that the task was challenging in the standalone version; users who worked with the operator did not report thus. Clearly, improvements can be made fairly easily by using spoken rather than written instructions; and, by developing a shape manipulation tool for the standalone system. With both of these improvements made, a sensible next step would be to carry out a similar experiment to the one conducted here. Perhaps in such a test, the delay from seeing a target to evolving the face could be much longer: two days is the norm in criminal investigations. While overall performance is expected to be lower, as longer delays tend to promote worse quality composites [[1],[6],[7]], good performance from the standalone system is likely to be maintained. As part of this work, it would be useful to test the system on members of the public, rather than on university students: ultimately, a design is necessary for use by all, not just by students.

The project also looked at whether there was any difference in system performance when participant-witnesses were instructed to make fast or slow selection of faces from the presented arrays. It was expected that the former would be more similar to how we naturally perceive faces, as wholes, and so would promote a more identifiable composite, and in less time. The design involved a realistic procedure, including a 2 day delay between seeing a target and evolving the face. The data did not support this hypothesis, with composites being named equivalently between the groups, but instructions to select faces more quickly did result in more variable results.

In the research and development of EvoFIT, we have observed that EvoFIT does not always behave in ways that one would expect of such an interface. In [[15]], for example, we found that an early version of EvoFIT did not benefit from a holistic type of face learning (encoding) while a feature-by-feature encoding did. Taken together with the results of the current project, it would seem that emphasising the featural information in a face is of value during face learning, but it does also seem to promote more reliable (stable) results for face decoding (face construction by slower face selection). As mentioned previously, and before forming strong conclusions here, it would be prudent to explore selection at a user-defined pace, as opposed to just encouraging fast or slow choosing.

In the second part of the paper, a searchable database was developed and evaluated. Searching appeared best for matching to occur for composites of the same type; matching was just as effective in the Image (physical) space as in the coefficient (PCA) space for Euclidean or for Mahalanobis Distance metrics. In the database used, correct matches occurred about 40% of the time in the top three with shape for EvoFIT, and texture for PRO-fit. While the size of the database was quite small, the overall result is encouraging. A searchable system that could return correct matches in the top three at this level of success would appear to be worthwhile. We envisage that such a system would involve a human observer relying on a search mechanism to screen out obvious non-matches but returning potential 'hits' within the first half a dozen or so. Of course, the next stage is to scale up and see if the results replicate for a much larger database.

Part of future work could also explore matching for different parts of the composite image. At present, all information contained in the internal facial features of the texture is used, but most of these pixels are for the area of skin and not the individual features themselves; arguably a better method would mask out such areas to allow a more representative measure of the colouring of features. A similar situation applies to the shape co-ordinates: all 298 are used but it is likely that only the co-ordinates for the inner face would be useful, areas that are likely to carry information most useful for identification [[21]].

5. Conclusion

The current work explored the feasibility of automating processes involved in the construction and identification of facial composites. A standalone version of the EvoFIT system was found to produce faces almost as identifiable as the full system that used an operator. It was also found that more variable results were observed when users were instructed to select faces quickly. Further, that there was some utility in matching composites produced from the same type of technology against each other in a small database; specifically, shape matching was best for EvoFIT and texture matching for PRO-fit. Overall, the work demonstrated promise for automating the production and identification of facial composite images.

6. Answers

The celebrities in Figure 1 are 'feature' composites of (left to right) singers Mick Jagger and Robbie Williams, footballer Wayne Rooney and actor Tom Cruise; for the EvoFITs in Figure 3, they are of actor David Tennant, politician George W. Bush, TV celebrity Simon Cowell and singer Noel Gallagher. All are generally well-known identities for people living in the UK.

Acknowledgement

The current work was gratefully funded by Crime Solutions, University of Central Lancashire, Preston, UK.

References

[1] C.D. Frowd, D. Carson, H. Ness, J. Richardson, L. Morrison, S. McLanaghan, and P.J.B. Hancock, "A forensically valid comparison of facial composite systems", *Psychology, Crime and Law*, 11, 2005, pp. 33-52.

[2] C.D. Frowd, D. McQuiston-Surrett, S. Anandaciva, C.E. Ireland and P.J.B. Hancock, "An evaluation of US systems for facial composite production", *Ergonomics*, 50, 2007, pp. 1987–1998.

[3] C.D. Frowd, P.J.B. Hancock, and D. Carson, "EvoFIT: A holistic, evolutionary facial imaging technique for creating composites", *ACM Transactions on Applied Psychology (TAP)*, 1, 2004, pp. 1-21.

[4] S.J. Gibson., C.J. Solomon, and A. Pallares-Bejarano, "Synthesis of photographic quality facial composites using evolutionary algorithms." In R. Harvey and J.A. Bangham (Eds.) Proceedings of the British Machine Vision Conference, 2003, pp. 221-230.

[5] C.G. Tredoux, D.T. Nunez, O. Oxtoby, and B. Prag, "An evaluation of ID: an eigenface based construction system", *South African Computer Journal*, 37, 2006, pp. 1-9.

[6] C.D. Frowd, D. Carson, H. Ness, D. McQuiston, J. Richardson, H. Baldwin, and P.J.B. Hancock, "Contemporary Composite Techniques: the impact of a forensically-relevant target delay", *Legal and Criminological Psychology*, 10, 2005, pp. 63-81.

[7] C.D. Frowd, V. Bruce, H. Ness, L. Bowie, C. Thomson-Bogner, J. Paterson, A. McIntyre, and P.J.B. Hancock, "Parallel approaches to composite production", *Ergonomics*, 2007, 50, pp. 562-585.

[8] G.M. Davies, J. Shepherd, and H.D. Ellis, "Remembering faces: acknowledging our limitations", *Journal of Forensic Science*, 18, pp. 19-24.

[9] C.D. Frowd, J. Park., A. McIntyre, V. Bruce, M. Pitchford, S. Fields, M. Kenirons, and P.J.B. Hancock, "Effecting an improvement to the fitness function. How to evolve a more identifiable face." In A. Stoica, T. Arslan, D. Howard, T. Higuchi, and A. El-Rayis (Eds.) *2008 ECSIS Symposium on Bio-inspired, Learning, and Intelligent Systems for Security*, 2008, pp. 3-10, NJ: CPS.

[10] C.D. Frowd, V. Bruce, A. McIntyre, D. Ross, and P.J.B. Hancock, "Adding Holistic Dimensions to a Facial Composite System", *Proceedings of the Seventh International Conference on Automatic Face and Gesture Recognition,* 2006, pp. 183-188, Los Alamitos: CA.

[11] http://www.uclan.ac.uk/scitech/psychology/research/evofit/science_centre.php

[12] V. Bruce, and A. Young, "In the eye of the beholder: The science of face perception". 1998, New York: Oxford University Press.

[13] J.W. Tanaka, and M.J. Farah, "Parts and wholes in face recognition". *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 46A, 1993, pp. 225-245.

[14] G.L. Wells, and B. Hryciw, "Memory for faces: encoding and retrieval operations". *Memory and Cognition*, 12, 1984, pp. 338-344.

[15] C.D. Frowd, V. Bruce, H. Ness, L. Bowie, C. Thomson-Bogner, J. Paterson, A. McIntyre, and P.J.B. Hancock, "Parallel approaches to composite production". *Ergonomics*, 50, 2007, pp. 562-585.

[16] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey", *ACM Computing Surveys*, Vol. 35, 2003, pp. 399-458.

[17] C.D. Frowd, M. Pitchford, V. Bruce, S. Jackson, G. Hepton, M. Greenall, A. McIntyre, and P.J.B. Hancock, "The psychology of face construction: giving evolution a helping hand", *Applied Cognitive Psychology*. Under revision.

[18] E. Baker, and M. Selzer, "The mug-shot search problem", Visual Interface '98 Proceedings, Vancover, BC, Canada, 1998.

[19] A.M. Levi, N. Jungman, A. Ginton, A. Aperman, and G. Noble, "Using similarity judgments to conduct a mugshot album search", *Law and Human Behavior*, 19, 1995, pp. 649-661.

[20] T. Koreniusa, J. Laurikkalaa, and M. Juhola, "On principal component analysis, cosine and Euclidean measures in information retrieval", *Information Sciences*, 177, pp. 4893-4905.

[21] C.D. Frowd, V. Bruce, and P.J.B. Hancock, "Evolving facial composite systems", *Forensic Update, 98,* 2009, pp. 25-32.

[22] H. Ellis, J. Shepherd, and G.M. Davies, "Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition", *Perception*, 8, 1979, pp. 431-439.