Originally:

11,075 words

50 pages

Target: 9414 words and 43 pages

Now:

8318 and 35 pages

Title: Making faces with computers: Witness cognition and technology

Running title: Witness cognition & technology

Authors: Graham Pike, Nicola Brace, Jim Turner and Sally Kynan

Affiliation: The Open University (for all authors)

Author responsible for correspondence:

Dr Graham Pike

Discipline of Psychology

The Open University

Walton Hall

Milton Keynes

MK7 6AA

United Kingdom

Acknowledgements

This research was conducted with the support of EPSRC grant GR/S06745/01(P). The authors would like to acknowledge the kind assistance of Dr Chris Solomon and the Forensic Imaging Group at The University of Kent.

Abstract

Knowledge concerning the cognition involved in perceiving and remembering faces has informed the design of at least two generations of facial compositing technology. These systems allow a witness to work with a computer (and a police operator) in order to construct an image of a perpetrator. Research conducted with systems currently in use has suggested that basing the construction process on the witness recalling and verbally describing the face can be problematic. To overcome these problems and make better use of witness cognition, the latest systems use a combination of PCA facial synthesis and an array-based interface. The present paper describes a preliminary study conducted to determine whether the use of an array-based interface really does make appropriate use of witness cognition and what issues need to be considered in the design of emerging compositing technology.

Biographical note

The authors are part of The International Centre for Comparative Criminological Research at The Open University. Recently they have been involved in two U.K. Home Office projects to investigate using computer technology to make the most of identification evidence and to study the visual identification of suspects. Their background is in applied cognitive psychology, particularly with regards to eyewitness evidence.

Introduction

Despite the recent advances made in physical and photographic identification, the eyewitness continues to play a central part in police investigations. That witnesses tend to be somewhat less than reliable has become a phenomenon well documented in research (for example see Cutler & Penrod, 1995) and, as a result, juries are warned against placing too much store in witness testimony throughout courts in both the UK and US. Perhaps the most important information that an eyewitness can supply is that relating to the identification of the perpetrator, but unfortunately this form of evidence is just as prone to error, if not more so, than more general forms of information about the crime.

Of key importance are the cognitive processes involved in encoding, storing and retrieving the face of the perpetrator. There is little, if anything, that can be done to improve the encoding and storing stages of this process, beyond only using witnesses who got a ‘good look’ at the time of the event, so it is the retrieval stage that researchers, practitioners and technology development have focused on in an attempt to improve the accuracy of eyewitness identification evidence. A considerable proportion of this research and development has concentrated on the construction and conduct of identification parades. Although of great evidential value, identification parades involve fairly minimal interaction with the witness, often limited to standardised instructions and the witness’ decision as to which, if any, of the people in the parade is the perpetrator.

In contrast to the single decision asked for at an identification parade, the construction of a facial composite (an image, typically constructed by recombining features from several faces, that attempts to capture the likeness of the person in question) of the perpetrator requires a great deal of interaction between the witness, the police operator and the system being used to aid construction. It is therefore vital that the interaction between witness cognition and technology is based on information that the witness can readily and accurately provide and avoids cognitively difficult or error prone tasks.

It is possible to produce an image of the perpetrator by employing a sketch artist and this technique has been found to produce accurate results (Laughery & Fowler, 1980), although it does require considerable skill and training. To overcome this requirement, systems such as Photo-FIT and Identikit were developed in the 1960s and 1970s and required witnesses to search through albums of individual facial features so that these could be assembled into a face-image. Research generally found that these early systems produced poor likenesses (e.g. Christie & Ellis, 1981; Ellis, Davies & Shepherd, 1978). Two explanations for the inaccuracy of the images produced were suggested: that the systems simply did not contain a database of features sufficient to cope with variability in appearance; and that the system was asking the witness to perform a task that was cognitively difficult. Although the former explanation undoubtedly accounts for some of the inaccuracy, the latter found great resonance with theories of face perception that suggested that the face was not simply encoded as a set of features but rather that recognition was also dependent on more holistic information, including the configuration of the features (see Rakover, 2002, for a review). In addition, some research had found that features were easier to recognise when presented as part of a face, i.e. within a face context (e.g. Tanaka and Farah, 1993). The poor results found with systems such as Photo-FIT may have, at least in part, been due to the fact that the witness had to search through individual features which were not presented within a face context: a task that did not match the essentially holistic nature of the cognitive processes involved with face perception.

Later systems, such as E-FIT, CD-FIT and more recently PROfit, took careful note of both applied and theoretical face perception research and were designed to make better use of witness cognition. To this end features were only ever presented within a face context and image manipulation tools were included to help make alterations to the configuration of features as well as to the individual features themselves. Attempts to evaluate these computerised systems have often produced inconclusive (and sometimes very complex) patterns of results (for example see Davies, van der Willik & Morrison, 2000). It should be noted that very often laboratory based research excludes many of the factors likely to lead to an accurate likeness being produced in the field, such as trained operators, lengthy interviews and construction times, and the use of artistic enhancement through image manipulation software (Gibling and Bennett, 1994). However, although this criticism undoubtedly prevents the forensic utility of any particular system to be estimated accurately, the fact that participants in the experiments find it difficult to produce an accurate image is still very interesting for what it reveals about the interaction of cognition and technology.

Looking at the generally poor rates of identification found with composites it is possible to conclude that witness memory is simply too inaccurate to ever produce usable images. However, as well as a simple comparison across systems, some researchers have looked in more depth at how well composite systems make use of witness cognition (e.g. Brace, Pike, Allen & Kemp, in press.) and have found that although more recent computerised systems such as E-FIT do interact more appropriately with the cognitive processes of the witness, they still involve tasks that are fundamentally difficult to achieve with any accuracy given the limitations of human cognition.

The composite construction process employed in systems such as E-FIT requires the witness to first recall the target face and verbally describe it to the operator. The operator then enters this description into the system by selecting appropriate feature descriptors, which are used by the system to rank order the exemplars for each feature and to produce an initial image comprised of the feature exemplars best matching the description provided by the witness. The witness is then shown this image, attempts to determine what is wrong with it and how it needs to be changed so as to better represent the target face. This is achieved by the operator using the system to replace individual feature exemplars with others from the database, altering the relative size and position of the features and by using image manipulation software to add finer touches to the face.

Examination of the above process shows it to be based on several tasks, each of which people generally find very difficult to perform with any accuracy. Firstly, the witness must recall the face of the perpetrator. Although research has found the act of identifying an unfamiliar face (such as at an identification parade) to be problematic (e.g. see Leippe & Wells, 1995; Levi & Jungman, 1995; Wells & Sealau, 1995), recognition at least appears to be the cognitive task that human face perception processes have fundamentally evolved to perform. Recalling a face involves retrieving the memory of that face and then somehow consciously bringing it to mind and is a far more complicated and difficult process to perform (Shepherd & Ellis, 1996). Once the face has been recalled, the witness must then verbally describe it to the operator. Research which has examined the descriptions provided by witnesses has generally found them to be of very poor quality, often missing out many features and containing little information about configuration (see, for example, Buckhout, 1974; Pozzzulo & Warren, 2003). Importantly, neither recall skills nor vocabulary seem to lend themselves to describing faces in detail which leads to sparse descriptions (Fahsing, Ask & Granhag, 2004).

To improve facial compositing technology so that it interacts more appropriately with witness cognition it is important, therefore, to address three factors: the reliance on facial recall; the reliance on verbal descriptions of the face; and the use of a database comprised of individual features. Recently several new facial composite systems, notably EVOfit and EigenFIT in the UK, have been developed with exactly these three factors in mind. Rather than use feature databases, both systems make use of a statistical technique known as Principal Component Analysis (PCA), which works by analysing the image properties of the whole face and thus capturing information that is intrinsically holistic, just as human cognition appears to (see, for example, Hancock, Burton & Bruce, 1996). PCA is applied to a training set of face images (usually several hundred to several thousand in size) to produce a set of eigenfaces (essentially the images corresponding to each eigenvalue resulting from the PCA). Although these eigenfaces in themselves do not represent any useful abstraction, they can be combined using different weighting mechanisms to form any face within the ‘face-space’ described by the original learning set. Thus by combining the eigenfaces it is possible to synthesise any face, as long as the learning set was suitably representative of the target population. The exact nature of this mathematical procedure is not of concern here: it is enough to note that the database is derived from entire faces, rather than individual features, and that construction therefore proceeds in a more holistic manner.

As not even a mathematically proficient witness can describe a face from memory according to its constituent eigenvalues, an interface radically different from that used by previous systems needed to be developed (see also O'Toole & Thompson, 1993). To avoid asking the witness to verbally describe the target face at any point in the procedure, the initial interfaces designed (for both EVOfit and EigenFIT) require the witness merely to select one facial image representing the closest match to the target from an array of several images (each array usually containing either 9 or 16 faces). The system then generates a new array of faces containing the previously selected face and with the other faces in the array resembling the chosen face to varying degrees and in varying ways. With each choice, the system gradually reduces the amount of variation in the array so that the faces should become more and more like the target. In fact, both EVOfit and EigenFIT make use of genetic algorithms to narrow down the variation (see Gibson, Solomon & Pallares Bejarano, 2003; and Hancock, 2000; for more detailed descriptions).

As all the witness needs to do is select the closest match to the target in each array, and to tell the operator when the system has produced a good likeness of the perpetrator, there is never a need to involve verbal descriptions. The third factor described above, that of avoiding recall, is also helped by the new interface. However, although it has become common to refer to array-based (or parallel) interfaces as being based on the cognition of recognition rather than recall, the situation is not actually quite so clear-cut. In fact the task would only be one purely of recognition if the witness looked at the first array produced and told the operator that one of the faces was an excellent likeness of the perpetrator. The chances of a good likeness being present in the first array are, unfortunately, statistically remote to say the least. Instead, the witness must look at each face in the array in turn, compare them and decide which one of them is the best likeness. When presented with the variation necessarily inherent in the initial array, this task might be akin to a pure recognition decision as the best match is likely to be easily evident. However, once the system begins to narrow the amount of variation between faces in the array, the witness' task becomes much harder and they undoubtedly have to look at each face systematically and make comparisons both between faces in the array and between each face and their memory of the target face. Thus, the task involves elements of both recognition and recall. So, although the use of a PCA database and array-based interface can be argued to overcome the problems associated with verbal descriptions and working featurally, they will still necessitate the use of recall, albeit it in a restricted form.

The design of PCA facial compositing systems therefore *appears* to be a step in the right direction in terms of a good fit between technology and cognition. However, early tests of images made with Evofit found them to have poor utility (Frowd, Hancock and Carson, 2004) and that E-FIT and sketch artist images were generally superior (Frowd, Carson, Ness, Richardson, Morrison, McLanaghan and Hancock, 2005). In essence, array-based interfaces do away with the need for lengthy and complex interactions between the witness and the system via the operator. One point to note here is that, although it is often very difficult and can undoubtedly lead to poor results, witnesses required to use feature-based compositing systems are generally able both to communicate with the operator and to describe what they want the system to do. A key question is therefore whether the problems associated with earlier systems (such as E-FIT) were due to the fact that they involved *any* recall and verbalisation or because the witness was forced into using recall and verbalisation throughout the construction process. Could it be that at least some of the information that the witness wants to provide is both accurate and useful? If this is so it could be argued that by not being able to incorporate this information, array-based interfaces are missing out on potentially vital cues and are in fact not interacting as well as they could with the cognitive processes of the witness.

The present paper describes a study that was designed to look at the above questions in more detail. In particular, the interaction between the technology involved in array-based facial compositing systems and witness cognition was examined. This was achieved by asking witnesses to interact with an array-based system and recording their reactions and comments. Rather than make use of an actual PCA compositing system, arrays of nine faces were created in advance, so that with each array the faces became more like the target face. This was done so that each participant worked through exactly the same images, making their comments directly comparable, and so that the images displayed did actually end up looking like the target face. This technique meant that the participants were not actually affecting the construction of the composite, although they did think that they were, and therefore allowed the interaction to be observed under more controlled circumstances. Critically, it allowed the amount and type of variation between the faces in the array to be controlled and pre-determined. In addition, the individual faces used in the array were constructed using the E-FIT compositing system, rather than a PCA system. Although this does mean that the arrays and images used differ from those that would be generated by a PCA system, the method adopted allows the separate manipulation of features and the configuration of features and also provides semantically meaningful differences between the arrays, e.g. the faces in one array may have shared five features or had just one feature in common.

As well as the amount of variation between the faces in each array, two other factors were examined: the number of arrays that the participants were asked to work through was manipulated to create a shorter sequence of 30 arrays and a longer sequence of 60 arrays; and the task the participant was asked to perform was manipulated so that some participants were asked to select the best match for the target from each array, whilst others were asked to select the most masculine face from each array. By asking participants to simply select the most masculine face from each array, any active recall of the target face and comparison with the faces in the arrays was removed.

Method

Design

This experiment employed a mixed design examining the within-participant factors of sequence length, with two levels (short, 30-array sequence; long, 60-array sequence) and array type, with six levels (five features changed; four features; three features; two features; one feature; configural changes only) and the between-participant factor of task type, with two levels (best match for the target selection or most masculine selection). Two target faces were used and the design was fully-crossed and counterbalanced for both array sequence order and target face order. As the study was exploratory, many dependent variables were included (and are detailed in the results and discussion), such as both quantitative and qualitative feedback on the array interface and process in general, consistency in selections both between and within participants and performance on a line-up task designed to test memory for the target face.

Materials: Target photographs

Two target "suspect" photographs were used as stimuli, each showing a full-face head-and-shoulders view of a person unknown to the participants. The targets were Caucasian middle-aged males, who were shown clean-shaven, with no spectacles or other facial paraphernalia and no distinguishing marks. Colour images of both targets were presented in high resolution on a 17"computer screen.

Materials: Images, arrays and sequences

For each target an "optimum" likeness was produced using E-FIT v3.1a by an operator with several years of experience with the E-FIT system and checked for accuracy by a panel of 5 trained operators. These composites formed the "base image" for each target, from which the subsequent E-FIT images were derived as detailed below.

In order to produce an image sequence that would develop a progressively closer likeness to the target, the "base image" was gradually modified using the E-FIT system to become increasingly *less* like the target. These modifications ranged from small changes in facial configuration (the smallest differences, which should remain fairly good type-likenesses of the target) to entirely new facial images sharing no features with the base-image. Placed in reverse order these images would therefore produce a sequence running from entirely dissimilar, effectively randomly-generated images bearing little or no likeness to the target (other than by coincidence) through progressively more similar-to-target images until finally reaching the optimum base-image itself.

Five sets of each array type (five feature, four feature, three feature, two feature, one feature and configural changes) were created, with nine images being created for each set. This gave six “types” of image, according to the number of features on which they differed from the optimum. Each type had five sets of images, differing in terms of which features were changed for the 1, 2, 3, 4 and 5-features changed types, and amount of feature displacement for the configural change type. Each set consisted of nine individual images, giving a total of 270 images for each target. The nine images within each set were cropped and resized to an approximately uniform size and positioned in a Microsoft PowerPoint slide to form a 3x3 image.

The arrays of images for each target were then arranged in PowerPoint slide sequences such that the *least*-like image arrays (sharing none or few of the optimum feature exemplars) would appear towards the beginning of the sequence and the *most*-like image arrays (sharing most or all of the optimum feature exemplars) would appear towards the end of the sequence. When run, the sequence therefore proceeded from composites bearing little or no resemblance to the target, through composites bearing progressively more of a resemblance to the target, and ending with composites bearing a good resemblance to the target. However, as it was envisaged that a composite system operating on a genetic algorithm would be expected to occasionally take a backwards step, there was some mixing of the image types in the array sequence to simulate such backwards steps. Specifically, the all features change and 4-changes arrays were intermixed, the 3-changes and 2-changes arrays were intermixed, and the 1-change and configural-changes arrays were intermixed.

For the long sequence condition for each target (60 arrays) the same overall pattern of array presentation was maintained as was used in the short sequence, however each set of images appeared twice in the sequence. For the second appearance of an image set, the images were pseudo-randomly rearranged so that they appeared in a different position in their second array than they had in their first. Thus, for example, an image which had appeared in the centre of the array on its first appearance in the sequence might appear in the top-left of the array on its second appearance. The occurrence of the repeated image sets was pseudo-randomised such that an image set never occurred twice in a row. For both the short and long sequences of arrays, the final array shown consisted of images with the smallest configural changes (single-feature, single-pixel displacements) including the optimum base-image itself.

Materials: Photospread line-ups

For each target a target-absent and a target-present photospread line-up was created. The target-absent photospreads consisted of 9 foil images, whilst the target-present photospreads consisted of an image of the target different to that seen previously (lighting, background, clothing and hair cues were changed) and 8 foils (different to those used for the target-absent photospread). The foils for both photospreads were judged to resemble the target visually as well as verbal descriptions provided of the targets by participants who were unfamiliar with them. The foil images were partly sourced from the Pics image database maintained by Stirling University (http://pics.stir.ac.uk). All of the images for the photospreads were standardised so as to be pictorially similar.

Participants

60 Open University staff volunteered as participants in this study. There were 47 females and 13 males, with ages ranging from 23-55. None were familiar with either of the targets or any of the photospread line-up foils.

Procedure

Participants were briefed as to the nature of the research and those in the 'best match' condition informed that their role would be to act as a participant witness and study the face of a target "suspect" for 30 seconds before working through a sequence of arrays of faces on the computer. They were also instructed that in each array they should select the face that was most like the target and, based on their selections, the computer should progress towards an improved likeness of the suspect. Those in the 'most masculine' condition were told simply that they would need to select the most masculine from each array. The participant was also told to study each array of nine face-images, and decide either which bore the most resemblance to the suspect or which was the most masculine (depending on condition). Participants were also encouraged to verbalise their thoughts as much as possible when working through the arrays, which, along with their decisions, were recorded by an experimenter. When the final array was reached, the participant was also asked to rate on a 10-point scale how good a likeness of the suspect the chosen image was (whilst it was still in view). The participant was then shown first a target absent and than a target present 9 image photospread line-up and told that the target suspect may or may not be present before each one. As well as indicating whether the target was present, and if so which image was of the target, the participant also gave a rating of their confidence in that decision using a 10-point scale. After completing the line-up task, the participant then worked through the sequence of arrays for the other target, following exactly the same procedure. The final stage of the study involved asking the participant questions about their experience of the arrays and eliciting any further comments or observations that the participant wished to make. Following this free-commentary session the participant was thanked for their assistance and de-briefed.

Results and Discussion

Participant feedback from best match selection conditions

The first thing to note is that all of the participants managed to work through both the 30 and 60 array sequences and to select a 'best match' in each array, with no participants giving-up due to task difficulty (or any other reason). This demonstrates that selecting the best match from an array of faces is at least something that the participants seemed able to do.

The verbalisations made by each participant were analysed to determine the frequency with which certain comments were made. This analysis revealed that in 56.25% of the sequences (each participant worked through two sequences, one short and one long), the participant commented in response to at least one array that they found the task of selecting the best match to be difficult and in 75% of sequences the participant commented for at least one array that they thought that none of the faces in the array looked sufficiently like the target to be selected as a ‘best match’. Analysis of the comments also revealed that participants quite often wanted to interact with the array in a different manner from simply choosing the best match. For example, in 28% of sequences the participants responded that they would prefer to indicate the face(s) that looked the least like the target, rather than the most; in 27.5% of sequences comments were made about wanting to select multiple faces from an array rather than just one; and in 70% of sequences the participants said they would like to alter or select individual features, rather than the whole face. Further analysis revealed that participants often felt that they had become confused, with mention being made in 31.25% of sequences that the participant was having difficulty remembering what the target face looked like.

One of the standardised questions asked after the array task, concerned the length of the sequences and analysis revealed that 60% of participants thought that 30 arrays was ‘the right number’ to work through and further that 58% of participants believed that 60 arrays was ‘far too many’ to work through. This is an important figure, as one potential advantage of PCA based systems is that they can produce a credible likeness much more quickly than feature-based systems. The data collected here suggest though, that the repetitive nature of the task means that witnesses may grow fatigued or disillusioned unless a good likeness can be reached within about 30 generations, even though this probably involves less than a third of the time it takes to construct a composite using systems such as E-FIT and PROfit. Informal qualitative analyses of the comments made by participants suggests that being presented with screen after screen of similar looking faces can make it difficult to remember what the target face looks like. It is also likely that requiring witnesses to make relative judgements about which of the faces presented is the best match involves more than the cognition required to decide if a presented face is, or is not, the target. The use of relative judgements to simultaneously presented faces has been shown to be problematic in the psychological literature on eyewitness identification, as it tends to lead witnesses to make misidentifications by choosing the best match for the perpetrator, even if the best match is not actually a particularly good match. Although selecting the best match is the aim of the array-based interface, the participants' responses suggest that this can become problematic if a large number of arrays are required to produce the final composite.

The participants were also asked about the selection method employed, i.e. selecting the best match, and other potential selection methods. 58% of participants responded that choosing the single best image was 'frequently' difficult, with 5% saying it was 'always' difficult and none 0% saying it was 'never' difficult. Opinion proved to be more divided when they were asked whether they would have preferred to use a selection method based on choosing the best two or three images in each array, with 35% responding 'probably' and 35% responding 'probably not'. A third type of selection method, that of providing a score out of ten for each of the faces in the array, has the potential to provide the system with a great deal of information, as a response will be given to each of the nine faces which would allow each to be weighted differentially when generating the next array and would allow the genetic algorithms employed in PCA systems to converge on the target more swiftly. However, 42% of participants responded that they would 'probably not' have preferred to use this method and 40% said that they would 'definitely not' have preferred to use it.

The final question asked about selection methods concerned giving specific feedback about the array, so that the participant could respond however they wished and the system would be able to accommodate their comments. With 75% of participants responding that they would have ‘definitely’ preferred this type of interaction with the system and a further 20% saying they would ‘probably’ have preferred it, this selection method proved to be the most popular. Systems such as E-FIT were designed so that a trained operator would be able to take virtually any comment made by the witness and translate it into a useful alteration to the composite being constructed. However, it is exactly this form of verbalisation that some researchers have suggested is a potential source of error in composite construction and is one that array-based interfaces can be designed to avoid if they employ the best-face selection method. The point is, though, that whilst the interface used by E-FIT *requires* the witness to verbalise every aspect of their decisions, if greater flexibility of selection were built into an array-based interface the witness would not *have* to verbalise every aspect of selection, but could do so if they desired.

Individual differences in faces selected from arrays

As well as examining the interaction between technology and cognition, the experiment conducted also allows an examination of individual differences in the cognitive processes employed. One benefit of the methodology employed in the experiment is that each participant was presented with exactly the same arrays. It was therefore possible to see whether participants tended to all select the same face as either the best match for the target (or the most masculine face) or whether different participants selected different faces even though they were presented with exactly the same faces in exactly the same order. Logically, it would be expected that there would be greater consistency in the faces selected from arrays containing a large degree of variation, such as those where the faces differed on four or five features, than those containing less variation, such as the arrays containing faces sharing all but one feature or those where only the configuration of the features was different.

The frequency with which each face in each array was selected as either the best match to the target or the most masculine face was calculated and these responses were then grouped according to the number of features changed in the array (i.e. five, four, three, two or one feature or just configural changes). From these data it was possible to calculate the mean number of participants who selected the modal (most selected) face. A summary of these data are presented in Table 1.

INSERT TABLE 1 HERE

These data were subjected to a 2 (task type - best match or masculinity selection) x 2 (length - 30 or 60 arrays) x 6 (number of features changed – 5 features through to configural only changes) between-participants ANOVA which revealed: a statistically significant main effect of task type ($F(1,156)=15.067$; $p<0.001$), with an effect size (partial eta^2) of 0.091; a statistically non-significant main effect of sequence length ($F(1,156)=0.825$; $p=0.365$), with an effect size of 0.005; a statistically non-significant main effect of array type ($F(5,156)=1.886$; $p=0.1$), with an effect size of 0.057; statistically non-significant two-way interactions between task type and sequence length ($F(1,156)=0.2797$; $p=0.598$), effect size of 0.002, task type and array type ($F(5,156)=0.645$; $p=0.695$), effect size of 0.02, and between sequence length and array type (*F*(5,156)=0.774; *p*=0.57), effect size of 0.024; and a statistically non-significant three-way interaction (*F*(5,156)=0.185; *p*=0.968), effect size of 0.006.

The only statistically significant result was the main effect of task type, with participants generally showing more agreement about which face was the most masculine (mean of 28.22% of participants choosing the most selected face) than about which was the best match to the target (mean of 24.5% of participants choosing the most selected face). By chance alone, each face would be picked by 11.11% of participants, so the means from both the masculine and best match selection conditions are more than twice as much as would be expected from chance. However, on average approximately only a quarter of participants selected the most chosen face, suggesting there are considerable individual differences in the perception of faces; whether this be in determining masculinity or the task of comparing the faces to the memory of the target.

From Table 1 and the results of the inferential analysis it is apparent that there was a great deal of variation in the faces selected as the best match. In addition, there was not a great deal of difference in this variation according to the similarities of the faces within the arrays, with 28% of participants selecting the modal face on average in the five-features changed arrays and 24.5% on average in the one-feature changed array. In many ways this result is surprising, as the task of selecting the best match from an array of very different looking faces should be far easier (and should therefore lead to fewer individual differences) than selecting the best match from an array of very similar looking faces. One possible explanation is that the participants had simply forgotten what the target looked like, but a check of their memory following the sequence of arrays revealed that their memories appeared intact (see later analyses). Instead the results suggest that there are considerable differences in the way that faces, at least unfamiliar faces, are encoded, stored and retrieved from memory.

That there are such large individual differences in face processing cognition is important in considering the design of facial compositing technology. Perhaps most importantly there is a need to consider whether the aspects of the face that the participant is concentrating on to make their decision about which is the best match, are similar to the aspects of the face that the system will be using to generate the next array. For example, the large individual differences could have resulted from some participants selecting the best match because they though the face-shape was a good match, some because the eyes were a good match and some because the overall look of the face was a good match for the target. If the compositing system is based on a PCA analysis of the entire face, then the next array generated may well exclude the aspects of the face that the participant concentrated on, especially if these were particular individual features. It could well be that a more flexible system, one that includes technology that can respond more directly to the cognition of the witness, by allowing a specific feature to be used prominently in generating the next array for instance, would be better able to cope with the obvious large individual differences that exist.

Within-participant consistency in faces selected from arrays

As well as allowing an examination of individual differences in the interaction of cognition and technology, the design of the present study also allowed consistency *within* the selections made by each participant to be investigated. This was only possible in the longer (60 array) sequences where each array appeared twice, so that the same faces were seen but in different positions within the array. It was therefore possible to determine how often each participant selected the same face from these matched arrays and how often they chose a completely different face, even though the face they selected the first time the array was presented was still available.

These data are presented in Table 2, and as before grouped according to whether they related to arrays that had either five features, four features, three features, two features, one features or just configural changes.

INSERT TABLE 2 HERE

The data were subjected to a 2 (task type - best match or masculinity) by 6 (array type - from 5 to 0 changes) mixed design ANOVA, which revealed: a statistically non-significant main effect of task type (*F*(1,58)=0.802; *p*=0.374), with an effect size (partial eta^2) of 0.014; a statistically significant main effect of array type (*F*(5,290)=13.528; *p*<.001), with an effect size of 0.19; and a statistically non-significant interaction between the two factors (*F*(5,290)=0.375; *p*=0.865), with an effect size of 0.006.

From an examination of Table 2 and the results of the inferential analysis, it is apparent that participants were generally more consistent when there was greater variation present in the array (five or four feature changes) than when there was less (one feature or configural changes). In other words, when there was greater variation a participant was more likely to select the same face from the two presentations of the array than when there was less. However, the data still suggest that there is considerable variation in the selections made by the same participant, as even in the five features changed condition over half the participants selected a different face the second time they were presented with the array and this figure rose to approximately 75% of participants making inconsistent selections in the configural changes condition.

The above analysis also shows that the best match and masculinity selection conditions produced similar results, as the main effect of task type and its interaction with array type were statistically non-significant. These results could suggest that the order with which the faces are presented in the array is important, as this order was changed between presentations of the matched arrays and most participants tended to select a different face. If participants were working sequentially through the faces, starting say with the top-left and finishing with the bottom-right, it could be that the serial order in which the face appeared was affecting its selection.

An alternative explanation is that the arrays seen between presentation of the two matched arrays were affecting the selection made. This could either be because the memory for the target face was being changed by exposure to the faces in the array or because the decisions made to the intervening arrays were leading the participant to concentrate on an aspect of the faces that led to a different face being selected from the second presentation.

Whatever the actual mechanism, this result suggests that the design of compositing technology could well benefit from allowing a witness to make a selection based on several faces rather than one; a suggestion that was made quite frequently by participants in this experiment (see analysis of feedback above). This would allow them to incorporate aspects from several faces if they felt that more than one face had something in common with the target. In addition, the results suggest that narrowing the amount of variation present within the arrays too greatly could be problematic, particularly if participants focus on one aspect of the face at a time. For example, if they first focus on getting the face-shape correct and then move onto the internal features, the arrays will need to contain sufficient variation so that a suitable selection of features is displayed. Without this variation the participant will be stuck with elements of the face that they had not actively selected.

Memory for the target face

Analysis of the feedback received from participants revealed that in nearly a third of all the sequences (31.25%), the participants reported having difficulty remembering what the target face looked like. Obviously, if their memory for the target face has been eroded in some way, either by the passage of time or by viewing the faces presented in each array, the task of composite construction will become very difficult indeed, if not impossible. However, it is possible that the participants' comments were more a reflection of the difficulty they were having in actively recalling the target face than any deterioration in the actual stored memory of the face *per se*. Although problems in recalling the face are far from trivial, they are not nearly as limiting as would be the case if the process of working with the array interface actually affected the stored memory.

To determine whether working through the arrays had indeed altered or eroded the memory for the target face, each participant's memory for the face was tested after they had completed each sequence. First of all they were asked to study a photo line-up consisting of 9 faces, which (unknown to them) did not contain an image of the target, and to say whether the target was present, and if so, to indicate the appropriate image. They were then shown a second photo line-up containing different faces, one of which was the target, and asked for the same decisions. A summary of the data relating to how accurately these tasks were performed is presented in Table 3.

INSERT TABLE 3 HERE

As can be seen from Table 3, the vast majority of participants were both able to determine correctly that the target was not present in the target-absent line-ups and to pick-out the target from the target-present line-ups. To complete this task, the participant's memory of the target's face must have been reasonably intact, and not particularly altered by exposure to the faces in the arrays.

Although performance across task type was similar in the case of the target present conditions, more mistakes were made with the target absent line-ups in the masculine selection condition than in the best match selection condition. Analysis of the data using the chi-square procedure (making use of the Fisher's exact test to compensate for expected counts less than 5 and using Cramer's V as an estimate of effect size) revealed that the difference between best match and masculine selection conditions was statistically significant for the target absent line-ups (chi^2=5.11; df = 1; p<0.05;

effect size = 0.206), but statistically non-significant for the target present line-ups (chi^2=0.43; df = 1; *p*=0.835; effect size = 0.019).

Thus there appears to be some advantage to interacting with the arrays in a manner that requires constantly accessing the memory for the original target. The fact that the difference was significant for the target absent and not the target present parades is probably due to the target absent parades presenting a more difficult task; as it is often easier to decide that the target face is present than to decide that none of the faces is that of the target. Moreover, it is possible that seeing the faces in all the arrays but *not* continually accessing the memory for the target face had a more detrimental effect on memory for the target than resulted from the constant recall of the target in the best match conditions. The main point here is that whatever negative effect exposure to the arrays seems to have on memory for the target appears to be overcome if the target is recalled and compared to the faces in the arrays.

Conclusions

The main conclusion to be drawn from the current study is that it could be inadvisable to limit the interaction of a witness/participant with array based facial compositing technology to making just a single selection from each array. Instead the large differences between individual participants (and also within a single participant throughout the process) suggests that greater flexibility is required in order to arrive at a good likeness. In particular, participants want to be able to respond to multiple faces in each array and also to specific aspects (such as individual features) of certain faces.

These findings appear at odds with the wider literature that suggests that both witness recall and verbalisation are important sources of error when constructing a facial composite. However, the studies these suggestions were based on made use of compositing technology that forced the witness to *always* verbalise their thoughts and to rely very heavily on recall. It could be that involving a greater amount of cognition based on recognition and allowing the witness to provide more specific comments only when they want to and only about those aspects that they want to cover, could lead to a more efficient and accurate compositing system.

References

Brace, N.A., Pike, G.E., Allen, P. and Kemp, R. I. (in prep). “Identifying composites of famous faces: Investigating memory, language and system issues”. Submitted to *Psychology, Crime and Law*.

Brigham, J.C. and Cairns, D.L. 1988. “The effect of mugshot inspections on eyewitness identification accuracy”. *Journal of Applied Social Psychology* 18(16, Pt 2): 1394-1410.

Buckhout, R. 1974. “Eyewitness testimony”. *Scientific-American* 231(6): 23-31.

Christie, D.F. and Ellis, H.D. 1981. “Photofit constructions versus verbal descriptions of faces”. *Journal of Applied Psychology* 66(3): 358-363.

Cutler, B.L. and Penrod, S.D. 1995. *Mistaken identification: The eyewitness, psychology, and the law*. New York: Cambridge University Press.

Cutler, B.L., Penrod, S.D. and Martens, T.K. 1987. “The reliability of eyewitness identification: The role of system and estimator variables”. *Law and Human Behavior* 11(3): 233-258.

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

Ellis, H.D., Davies, G.M and Shepherd, J.W. 1978. "A critical examination of the Photofit system for recalling faces". *Ergonomics* 21(4): 297-307.

Fahsing, I.A., Ask, K. and Granhag, P.A. 2004. "The Man Behind the Mask: Accuracy and Predictors of Eyewitness Offender Descriptions". *Journal of Applied Psychology* 89(4): 722-729.

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

Frowd, C.D., Hancock, P.J. & Carson, D. 2004. "EvoFIT: A Holistic, Evolutionary Facial Imaging Technique for Creating Composites". *ACM Transactions of Applied Perceptions* 1(1): 19-39.

Gibling, F., and Bennett, P. 1994. "Artistic Enhancement in the Production of Photo-Fit Likenesses: An Examination of its Effectiveness in Leading to Suspect Identification". *Psychology, Crime and Law* 1: 93-100.

Gibson, S.J., Solomon, C.J. and Pallares-Bejarano, A. 2003. "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* 1: 221-230.

Gorenstein, G.W. and Ellsworth, P.C. 1980. “Effect of choosing an incorrect photograph on a later identification by an eyewitness”. *Journal of Applied Psychology* 65(5): 616-622.

Hancock, P.J. 2000. “Evolving faces from principal components”. *Behavior Research Methods, Instruments and Computers* 32(2): 327-333.

Hancock, P.J.B., Burton, A.M. and Bruce, V. 1996. “Face processing: Human perception and principal components analysis”. *Memory-and-Cognition* 24(1): 26-40.

Laughery, K.R. and Fowler, R.H. 1980. “Sketch artist and Identi-kit procedures for recalling faces”. *Journal of Applied Psychology* 65(3): 307-316.

Leippe, M.R. and Wells, G.L. 1995. “The police lineup: Basic weaknesses, radical solutions”. *Criminal Justice and Behavior* 22(4): 373-385.

Levi, A.M. and Jungman, N. 1995. “The police lineup: Basic weaknesses, radical solutions: Reply”. *Criminal Justice and Behavior* 22(4): 386-396.

McClure, K.A. 1998. “The use of participant free-hand drawings and written verbal descriptions as practice for a facial recognition task: Implications for improving eyewitness identification accuracy”. *Dissertation Abstracts International: Section B: The Sciences and Engineering* 59(5-B): 2448.

O'Toole, A.J. and Thompson, J.L. 1993. "An X Windows tool for synthesizing face images from eigenvectors". *Behavior Research Methods, Instruments and Computers* 25(1): 41-47.

Pozzulo, J.D and Warren, K.L. 2003. "Descriptions and identifications of strangers by youth and adult eyewitnesses". *Journal of Applied Psychology* 88(2): 315-323.

Rakover, S. S. 2002. "Featural vs. configurational information in faces: A conceptual and empirical analysis". *British Journal of Psychology* 93(1): 1-30.

Shepherd, J. W., and Ellis, H. D. 1996. "Face Recall - Methods and Problems". In S. L. Sporer (Ed.), *Psychological Issues in Eyewitness Identification*. Mahwah: Lawrence Erlbaum Associates, 87-115.

Tanaka, J. W., and Farah, M. J. 1993. "Parts and Wholes in Face Recognition". *Quarterly Journal of Experimental Psychology: Human Experimental Psychology* 46A(2): 225-245.

Wells, G.J. and Seelau, E.P. 1995. "Eyewitness identification: Psychological research and legal policy on lineups". *Psychology, Public Policy and Law* 1(4): 765-791.

Table 1: Percentage of participants selecting ‘most chosen’ face, by condition, sequence length and number of features on which faces differed

			No. of features changed					
			Five	Four	Three	Two	One	Configural
Best match	Short sequence	Mean	28	26	26.5	21.5	24.5	22
		SD	8.91	6.75	3.79	3.79	3.71	3.26
	Long sequence	Mean	26.25	26.5	24.75	23.75	24.75	20.25
		SD	6.37	5.43	5.2	6.26	7.02	3.22
Most masculine	Short sequence	Mean	34	30	27	26	29	29
		SD	8.9	5	5.7	4.18	7.41	12.45
	Long sequence	Mean	28	28	26.5	29	28	27
		SD	7.45	6.75	5.79	4.59	7.53	5.37

Table 2: Mean percentage of matched array pairs from which the same face was selected, by array type

		No. of features changed					
		Five	Four	Three	Two	One	Configural
Best match	Mean	47	39	34	30.63	22	21.37
	SD	27.76	25.2	21.82	27.2	18.56	18.95
Most masculine	Mean	53	47	32	31.5	26	23
	SD	21.79	22.73	20.93	27	19.58	26.92

Table 3: Participants' selections from photospread line-ups (TA = Target Absent; TP = Target Present)

			Hit	Miss	False Alarm	Correct rejection
Best match	Short sequence	TA	-	-	0	100
		TP	90	10	0	-
	Long sequence	TA	-	-	2.5	97.5
		TP	87.5	12.5	0	-
Most masculine	Short sequence	TA	-	-	5	95
		TP	90	10	0	-
	Long sequence	TA	-	-	15	85
		TP	90	0	10	-