THE INFLUENCE OF SOME FACTORS AFFECTING FACIAL COMPOSITE PRODUCTION AND THEIR APPLICATION IN PRACTICAL POLICING

Stephanie Plews B.Sc. (Hons)

Thesis submitted for the degree of M.Phil. Department of Psychology, University of Stirling, 2006

Table of contents

Acknowledgements

I would like to thank my Principal Supervisor, Dr Peter Hancock for his overall supervision throughout this period of study and for his invaluable comments and feedback on draft chapters.

Thanks also go to ABM United Kingdom Limited and Dr Leslie Bowie for the use of PROfit software and their financial backing, without which it would have been extremely difficult to undertake this research project.

My greatest appreciation goes to Dr Charlie Frowd who has provided me with a great deal of support and advice throughout this project. In addition my thanks go to Dr Hayley Ness, Robbie Cooper and Katharine Walsh for their composites and identification data and, of course, to all those who participated in these experiments and aided with data collection.

Finally, I would like to thank my husband, Michael, for giving up our spare time at weekends during my final year and for doing the household chores whilst I have been concentrating my efforts on writing-up the thesis!

Abstract

To date, identification performance from composites remains poor, especially where forensically valid procedures are adopted in the construction process. Several experiments have assessed some of the reasons for poor performance. In the first experiment, the effects of operator performance were assessed through the construction of composites of the same targets across novice and experienced operators. Performance was assessed through uncued naming and likeness ratings and results indicated that the performance of the experienced operator may have been no better than novice operator performance (however, there were procedural differences). Target distinctiveness was also manipulated and targets rated as being highly distinctive were identified more often than targets with low distinctiveness. The second experiment concentrated on the effects of retention interval (2 days and 1 week) and artistic enhancement. Naming and likeness ratings were poor. Likeness ratings revealed an advantage for composites constructed with the shorter retention interval. The use of artistic elaboration appeared advantageous with the longer retention interval of one week. A third experiment implemented retention intervals of 3 - 4 hours and 2 days. Again, naming levels and likeness ratings were poor. There was a trend in the direction of the shorter retention interval providing better identification results.

1

Review and approach

What are facial composites?

A facial composite is an eyewitness's account of the general facial appearance of a crime perpetrator and is constructed with the aim of including the 'unique' elements of that particular person's face. Composites are used by US and UK law enforcement agencies to trace a suspect to a crime or to eliminate large numbers of individuals, in their investigations. They can be in either sketched form or formed using a database of facial features from one of the computerised or non-computerised composite production systems available. Future systems offer selection and breeding of whole faces to evolve a composite (for example, EvoFIT, Frowd et al., 2004).

The Association of Chief Police Officers (Scotland) (ACPO(S)) guidelines on facial imaging, term the facial composite a 'composite likeness' and define it in the following ways: "made up of various parts or blended" and "as bearing close similarities, characteristics and resemblances to the person portrayed (pg.2)." A common fallacy surrounding facial composites is that they are intended as an exact replication of the facial form of a crime perpetrator. This is not so, and would indeed be highly unrealistic to achieve from a witness memory given what psychologists know of the weaknesses in human memory performance, in general, and particularly in relation to eyewitness accounts (Loftus, 1979). The purpose of the composite is to trigger the recognition of a person that is highly familiar with the crime perpetrator by replicating basic facial appearance and including the most prominent features of the particular individual's face. The composite is not used as a photograph or mug shot would be, for triggering recognition in an individual who may have only seen the person of interest on one or two prior occasions. In fact, the main reason for the use of composite systems in their early stages of development was in the elimination of suspects, and not necessarily in the identification of a suspect (Bennett, 1986).

Psychologists have provided a great deal of research into the most favourable methods required to produce a facial composite from a witness, based upon extensive research on memory and facial recognition in general. The Facial Imaging Operator (who is either a facial imaging specialist trained to produce facial composite images using a computerised system, or a police artist who uses sketches), employs a Cognitive Interview Procedure, or similar technique, to elicit the optimal recall of the facial form of a crime perpetrator from an eyewitness. During the procedure, the event or scene is recreated in the mind's 'eye' of the witness, so that he relives every aspect of the incident, and is able to recall as much detail as possible.

The first approach used to create a facial likeness from a witness was the use of a Sketch Artist. These artists were skilled in the use of portrait drawing with the use of pencils and crayons. Later, two systems were developed that were intended for use by those who were not as skilled in artistic method: IdentiKit and PhotoFit. IdentiKit was a system developed in the US and included a resource of sketch-like facial features. Photo-Fit, on the other hand, was used in the UK, in its early stages from 1970 and included a library of features sourced from photographs. The Field Identification System was an alternative composite system; this was a book-like device which allowed for minimal intervention from an operator. The subject selected facial features by turning the pages of the book.

Advancements in computer technology over the years, have led to the development of computerised facial composite systems. The proposed advantages of these systems over the other systems were that they included increased databases of features and they professed to be accessible by those who did not necessarily possess artistic skills. Of these systems, Mac-a-Mug-Pro was developed which included databases of sketch-like features. Following on from the Mac-a-Mug-Pro system were those systems that are in more frequent use today; E-FIT, PROfit, FACES, IdentiKit 2000 and the more recent EvoFIT-type systems (see Figure 1.1 below for PROfit and E-FIT examples). These systems were deemed to be of benefit by including features that were more realistic or human-like and many systems included databases of features compiled from actual photographs (as did PhotoFit), or adopted an approach of selection and breeding of the facial form (eg. Hancock, 2000).

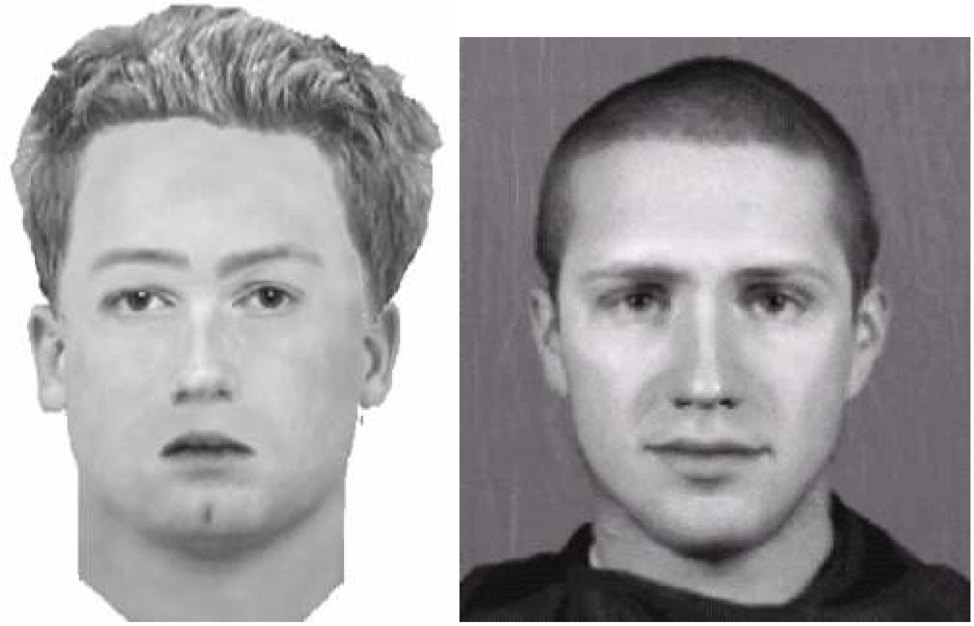

Figure 1.1: Composites produced using PROfit and Evo-FIT respectively

In recent years, with increased accessibility and availability of computerised systems for facial composite production, police forces throughout the UK and the US are faced with the choice of knowing which of these systems to utilise, if any, or whether the more traditional method of employing a sketch artist is preferred. The trend in psychological research has been towards assessing the effectiveness of various composite production systems against each other and the more traditional method of the sketch artist, concentrating on the benefits for the operator, the witness and, most importantly, which systems produce favourable results in the form of identification of a suspect. To date, even with advancements in computer technology, laboratory research has shown that identification of facial composites is at low levels, when using forensically valid procedures (Frowd et al. 2005a, b).

The effectiveness of facial composite systems

Research on the value of composite production systems has concentrated on the reasons behind the poor performance of composite systems in relation to facial recall ability from memory. Specifically, research has focused on the question of whether it is the systems themselves that are inflexible and ineffective, whether it is the witness with the poor recall ability and descriptive powers or a combination of both of these factors. Early research by psychologists into the effectiveness of composite systems centred around the sketch artist in comparison with the IdentiKit and PhotoFit systems. Davies (1983) refers to the difference between the sketch and composite systems, respectively, as a choice between a realistic and life-like portrait, which may be wrong in detail but readily related to faces stored in memory and a schematic representation which conveys only what is known and, because of its simplicity will trigger recognition.

The Photo-Fit system was launched in 1970 and was based largely on IdentiKit but included features from photographs rather than sketch-like features as in the IdentiKit system. A survey was conducted in 1970 - 1971 to collect information relating to the initial experiences of police forces using the system (King, 1971). The data analysis provided information that the average time between offence and interview was four days, with a range of six hours to over two weeks. The preferred order in which faces were constructed was: hair, chin, eyes, nose and mouth and the average time taken to construct an acceptable composite was around fifty minutes, with an average of six attempts before an acceptable composite was shown to a witness. Operators were questioned as to the weak areas of the system and answers included: a lack of age-defining lines resulting in an ageless face, hair styles becoming outdated and eyes and eyebrows in combination (the kit only allowed for viewing in combination). Success rates were analysed and it was found that of the 96 occasions on which a composite had been produced, 16 cases of tracing a suspect resulted and 13 people were charged with an offence. Initial results indicated a great deal of success in using the PhotoFit system.

Performance of the PhotoFit system was further assessed in a study by Ellis, Davies and Shepherd (1978). It was recognised that police officers in some forces were expected to remember composite ‘PhotoFit’ faces as part of their everyday duties. Ellis et al. proposed that composite faces would be less easily remembered than real faces due to their inherent unrealistic appearance, brought about by either: lines at the boundaries of the five facial components; the absence of colour and skin texture information in the images, or that in construction of the kit, only one feature was taken from a particular face. Ellis et al. sought to determine whether composite faces were less easily remembered than real faces. Results from three studies indicated that lines on the face produced impairment in memory. The presence of lines was found to disrupt either the encoding or storage of faces, indicating that faces are processed in a holistic manner and that lines or boundaries interfere with this. It was suggested that composite systems should avoid the presence of lines or boundaries on the face. Ellis et al. also suggested that the IdentiKit system may, in this way, be superior to PhotoFit as IdentiKit uses transparent layers to build up the features of the face rather than construction of the face from several different components, which would create boundaries.

In a subsequent study, Davies, Ellis and Shepherd (1978) studied the effects of identification of well-known faces from photographs, detailed line drawings and outline drawings. Photographs gave superior performance. A second study involved either photographs or detailed line drawings and participants were asked to recognise the persons depicted in a photographic recognition task. Recognition was found to be better with the photographic stimuli. However, subsequent research by Bruce et al. (1992) showed that line drawings that contain elements of ‘mass’ or convey shape from shading information and so on may be as good for facial recognition as are photographic reconstructions. Laughery and Smith (1978) acknowledged the importance of the type and quality of facial images in an identification task where identification was based upon a constructed image. They studied the sketch against the IdentiKit composite and predicted that, as sketches contain more detail than IdentiKits, recognition performance with sketches would be superior. Results were in line with predictions and sketches were found to have an advantage over IdentiKits in a recognition task. Laughery and Smith put forward the view that sketch artists should be used in preference to IdentiKits wherever possible.

Following on from earlier discussions by researchers on the performance of face recall systems such as PhotoFit and IdentiKit, Davies and Christie (1982) studied the reasons behind their poor performance, in comparison to the overall superior levels of performance in face recognition studies in general. Two suggestions were assessed; that the process of composite construction may produce memory interference and that the design of the systems may be incompatible with the way in which participants store and retrieve facial information. Davies and Christie proposed that the process of continuous exposure and comparison involved in selecting the components of the composite may interfere with, and degrade the subject's internal representation of the target's face. In addition, systems that require a face to be built from individual components may be incompatible with the way faces are stored in memory. Evidence suggests that faces are stored and recognised as wholes (e.g. Tanaka & Sengko, 1997). Results relating to the holistic storage of faces were in line with the proposed suggestions; there was support for the view that faces are normally stored as integrated units from which it is difficult to extract feature information. The practical implications of this result were similar to those of Laughery and Fowler (1980) who preferred the sketch artist. Davies and Christie suggested that there was an advantage in selecting whole faces or groups of features rather than individual features.

Davies, Milne and Shepherd (1983) recognised the poor performance of the PhotoFit system but suggested that performance from any other system was not superior and that until future systems were developed, work should be undertaken towards maximising its potential. They were concerned that there appeared to be a skill element involved in the operator producing a recognisable composite but this skill had not yet been identified. In a paper by Bennett (1986), there is further recognition of the limitations of the PhotoFit system and discussion is made surrounding the use of the system within

UK law enforcement since the early 1970s. Bennett noted that the PhotoFit system was rarely able to provide a satisfactory impression and that from a practical perspective there was much apathy towards the use of the system from police officers. He suggested that improvements should be made in several areas if PhotoFits were to achieve their potential. Firstly, that there was a lack of features within the database from which to construct composites. Additionally, that operators needed further training in order to become skilled in the use of the system and in their own artistic abilities. Bennett recognised the need for an understanding of the psychological factors surrounding eyewitness recall (delay, stress and weapon focus, see Loftus, 1979) and the ways in which these factors should be considered alongside the operation of the PhotoFit system.

Bennett's (1986) discussion paper emphasised the practical application of facial composite systems within a law enforcement setting in relation to: witness factors in a psychological context, systems factors such as a lack of features within the database and the role of the operator. Many studies have assessed the importance of the effects of face processing in general, on composite construction accuracy.

Psychological research has recognised that there are two distinct processes involved in the recognition of familiar and unfamiliar faces. Ellis et al. (1979) proposed that the internal and external features of the face are of roughly the same importance in determining its identity in faces seen only once. Further, they put forward the view that not all features were equally memorable but that the hairline would attract more attention than the jawline and the eyes would receive more attention than the nose and mouth. In contrast, for familiar faces, the internal features are more likely to lead to recognition than the external ones. In a later study, Young et al. (1985) also confirmed the differences in the processing of internal and external features involved in the recognition of familiar and unfamiliar faces. They extended this finding to show that it is only evident when stimuli are treated as faces. In summary, Young et al. were able to show that in face-matching tasks, it is easier to match familiar than unfamiliar faces on their internal features and further, that there is no difference between familiar and unfamiliar faces on external feature matching.

Later, Bruce and Young (1986) put forward a functional model for face processing whereby they explained the differences between the structural coding of familiar and unfamiliar faces. They argued that the structural differences arise because stored structural codes for known faces become elaborated through repeated exposure and are represented in recognition units, which are not present for unfamiliar faces. So, to recognise a familiar face, there must be a match between the encoded representation and the stored structural code.

In addition to the effects of familiarity and non-familiarity upon memory performance in face recognition, is the issue surrounding whether individual feature processing or holistic processing resulted in superior recall and therefore, composite construction and identification. Much research has suggested that face recognition favours holistic processing; faces are recognised in terms of the whole that emerges from the features and not on their individual facial features (Tanaka & Sengko, 1997). Early studies pointed to the effects of feature configuration, or their arrangement in respect to each other as being an important factor in face recognition (Matthews, 1978) and on holistic processing. In fact, since Galton's time, researchers have suggested that the spatial relations of the facial features might be as important to recognition as the features themselves (Tanaka & Sengko, 1997). The holistic hypothesis of face recognition proposes that the facial features and their configuration interact in such a way that changes in one source of information should disrupt the other source of information. Therefore, if modifications are made to the spatial locations of the facial features, the face may not be recognised as readily. The results of four experiments in which Tanaka and Sengko manipulated spatial locations of the facial features showed that configuration affected the holistic recognition of features from normal faces. Specifically, that modifying the spatial location of one feature impaired the recognition of other features; these features remained in the same positions. This result provided supportive evidence that spatial information about a feature is not defined by its absolute position but by its position relative to other features. In summary, the results of the study showed that changes in face configuration affected the recognition of its features.

Research by Yount and Laughery (1982) defined visual familiarity with the face as dealing with a large unit of visual information. They suggested that the higher the level of familiarity we have with a face, the more we process it holistically. Results from a more recent study by Schwaninger et al. (2002) have challenged the assumption that faces are processed only holistically. Their research results showed that we process familiar and unfamiliar faces by encoding and storing configural information as well as the local information contained in facial parts. They put forward an integrative model for familiar and unfamiliar face recognition based on these results.

The manufacturers of composites systems have attempted to take into account these findings in relation to holistic face processing and configurational relationships and to modify the systems to make them more effective in these areas. The earlier composite systems required that a composite was constructed on the basis of individual feature components; later computerised systems allowed for a more holistic strategy of composite production and greater ranges of features. It was therefore considered that the contemporary systems would be more flexible and provide for more effective composite construction and increased identification. In a study using the Mac-a-Mug Pro system, Koehn and Fisher (1997) revealed that even the more modern computerised systems were lacking in terms of their ability to promote holistic processing at retrieval. Further, that the Mac-a-Mug Pro system produced composites that were of a poor quality and rarely produced correct identifications. They suggested that the poor performance of the system resulted from the fact that composites in their study were produced from memory.

A study by Davies and Oldman (1999) also recognised that the contemporary systems did not give superior results compared to the older systems, when composites were constructed from memory. In order to explore one of the reasons for the poor performance, Davies and Oldman replicated the main features of an earlier study by Shepherd et al. (1978) where it was suggested that where the demands of the recall procedure exceeded the subject's recall abilities, participants would rely on their expectations and stereotypes to construct the face. Results were that positive and negative attitudes did have an effect on the construction of the composites and also on the identification of the composites. Composites made by those who disliked the targets were ranked higher overall and elicited more spontaneous correct identifications, than those who liked them. The reason put forward for this occurrence was that positive judgements encourage global evaluation of faces, which improves recognition but impairs composite construction. This result was reassuring in terms of the fact that a negative attitude towards a suspect was likely to have a positive effect on the construction process rather than hinder it.

A study by Brace, Pike and Kemp (2000) gave more promising results in terms of memory performance. They found that the E-FIT system was a useful tool for identification although there were problems with verbal translation of the witness memory. Mean identification scores were 35% with an operator constructing and 25% where a describer was used alongside the operator. The suggestion was that the quality of the E-FITs was limited by the verbal translation of the memory. Memory performance differed depending whether the composites were constructed by the operator or with a describer; for composites constructed by the operator, those constructed from memory were equally likely to be selected as the best likeness as those constructed with a photo in view. With a describer, more E-FITs with a photograph were chosen to be superior to those constructed from memory. These findings are in line with those from Wogalter and Marwitz (1991) who found it was the verbal description that added ‘noise’ to the composite when constructing using the Mac-a-Mug Pro system. Their results showed that participants could quickly learn how to operate the system and could produce competent composites with minimum assistance. Where performance was poor, was when a second person was involved to whom the witness provided a description. However, early research by Laughery et al. (1980, cited in Wogalter & Marwitz, 1991), using the Field Identification System (FIS) which required minimal operator assistance, found that when compared to composites constructed using the IdentiKit system and a sketch artists composites, performance was worst from the FIS composites. Laughery et al. suggested one reason for this poor performance was in the lack of an expert familiar with face recall procedures in the composite construction process.

Davies et al. (2000) also provided positive results in terms of the effectiveness of the E-FIT system. However, performance was no better over the older, mechanical composite systems when studied under more realistic conditions, with recall from memory. As for the Mac-a-Mug Pro studies, performance was found to be good only for cases where the target was present in the construction phase. These results were disappointing given that E-FIT (as well as other similar systems such as PROfit) was designed specifically for retrieval of configurational facial feature information and therefore differed in this way from the older systems, that were criticised for their lack of flexibility in terms of feature by feature retrieval.

In summary, the newer composite systems did not appear to perform better than the older systems when assessed under realistic conditions of recall from memory and they continued to be ineffective in terms of promoting holistic processing at retrieval. Alongside all of the factors surrounding the construction and identification performance of facial composites from facial composite systems, this thesis will concentrate specifically on the effects of: operator performance and artistic enhancement; distinctiveness of the facial features and retention interval, and detailed discussion will now be given to each of these factors.

Operator performance, artistic enhancement, target distinctiveness and retention interval

To a great extent, it is difficult to make the distinction between operator performance and artistic enhancement as the two appear to go hand in hand. Intuitively, the more experienced the operator is, the greater skill they will have at producing enhanced composites that have more identifiable qualities. However, research findings in this area have produced mixed results in relation to the level of skill required from the operator to produce identifiable composites. Early research by Ellis, Davies and Shepherd (1978) highlighted the effects of operator skills on the accuracy of Photo-FIT constructions. They found no difference in composite quality whether there was a professional police operator or an inexperienced novice operator constructing. Results showed that the experienced operator employed more pencil work to modify the features of the kit but this had no discernable effect upon identification.

Disappointingly, they also found that judges rated participants sketches of target faces as being better likenesses than the Photo-FITs themselves.

In a later study examining the effects of operator skill, Laughery and Fowler (1980) compared composites produced by a sketch artist with those produced by an IdentiKit technician. Results indicated that sketch artists produced better images than the IdentiKit. Several reasons were put forward for the superiority of the sketch artist including the infinite set of facial features that they were able to produce compared to the IdentiKit, with its limited amount of facial features. Also, sketch artists were able to add greater detail to their composites in terms of shading and other artistic effects that are not available in the IdentiKit. A further reason was that the sketch artist spent a longer time in drawing the features of the face, which in turn led to the witness having more time to devote to thinking about the face of the target, and in turn may lead to a more accurate facial representation. As mentioned previously, the feature-oriented approach of the IdentiKit may have been inferior to a more holistic approach allowed by the sketch artist, whereby the witness was able to move around the face, rather than concentrating on each individual feature sequentially. Results of the study provided evidence that it was the inflexibility of the system itself rather than the ability of the technician that affected composite quality; there was little or no difference between the performance of the technicians but there were differences between the sketch artists. From a policing perspective it was suggested that where there is a choice, sketches should be chosen over IdentiKits. Artistic skills were found to be important and it was suggested that selection and training in these areas should be maximised. It was further suggested that any system that eliminated the artist or technician from the construction process would be advantageous over current systems.

Davies, Milne and Shepherd (1983) directly compared novice and experienced operators. From informal discussion with police officers, they noted it was the experience of many of these officers, that to produce a good PhotoFit involved skills that not all operators possessed. In a study, a target face was viewed by two witnesses and subsequently, a composite constructed by both a novice and an experienced operator, independently. Composites were assessed on two measures of composite quality; likeness ratings and a sorting task. Results clearly supported the view that operator experience influences the quality of likeness achieved by a witness subject. Davies et al. identified three areas where the experienced operator may have excelled: firstly, in the initial questioning of the witness to obtain a facial description; secondly, in relating the description to the features in the kit and thirdly, in applying technical skills to produce a realistic and acceptable composite. Results of a further study suggested that it is in the initial selection of the features where the experienced operator excels and not necessarily in the technical modification with the use of pencil work.

Gibling and Bennett (1994) conducted a study to further assess the effects of operator enhancement and artistic skills on the quality of Photo-FITs. They put forward the hypothesis that skilled and semi-skilled operators would have the ability to enhance the Photo-FITs using various enhancement techniques, to a level where they would be significantly better recognised than those that had not been enhanced. Results confirmed the value of enhancement; experienced operators produced significantly better likenesses. These results were obtained from operators undergoing training and it was therefore hypothesised that results would be even more promising from skilled and experienced operators. Gibling & Bennett's results were in agreement with the views of operators in the field, that the limitations of the composite systems can be overcome with proper training. From a practical perspective, this study illustrates the importance of: "knowledge of and skill in enhancement techniques, as well as artistic ability" and that these are "critical…if the PhotoFit system is to be employed to its full potential and with any real value or advantage (pg.99)."

Feature saliency or distinctiveness is a further factor that may have an impact upon the operator's skill and ability to produce a recognisable composite. The effects of saliency or distinctiveness on facial recognition have been studied for a number of years with theorists recognising the importance of salient features within facial recognition. Winograd (1981) studied the effects of distinctiveness in terms of the elaboration hypothesis, which states that it is the amount of information encoded rather than the type that is important. Winograd's results were in line with the elaboration hypothesis; it was found that faces searched for their most distinctive feature were remembered as well as those evaluated with respect to a trait. Further, that a judgement about a single feature can be as beneficial to memory as a trait question, if the feature is distinctive. Trait judgements enhance memory more for faces of high distinctiveness than low distinctiveness and the optimal condition for encoding faces is a scanning strategy applied to a face that has a distinctive feature. Courtois and Mueller (1981) were in agreement with the suggestion that salient or uncommon faces are better remembered in terms of descriptions and identifications than non-salient or common faces.

In their meta-analysis of facial identification studies, Shapiro and Penrod (1986) recognised target distinctiveness as one of the variables studied by a number of researchers. Findings from various studies showed that distinctive targets are more easily remembered than ordinary looking targets and that distinctiveness has a definite role at the retrieval stage and may also have a role at the encoding stage of face recognition. At the encoding stage, distinctive faces may contain more information and elicit more extreme judgements, thereby increasing the level of processing. Valentine and Bruce (1986) also report on the effects of distinctiveness on face recognition: "a distinctive or unusual face is generally better remembered than a typical face in a recognition memory task." Valentine and Bruce (1986a, as cited in Valentine & Bruce, 1986, pg. 525) reported that faces that were rated as very distinctive or very familiar were recognised faster than faces rated as typical or less familiar. They put forward the hypothesis that a general face prototype was abstracted from faces previously encountered and used as a basis for encoding faces in the future. Vokey and Read (1992) decomposed typicality in face recognition into two components: familiarity and memorability. They proposed that general familiarity and memorability work in opposition in face recognition and that general familiarity would reduce facial discrimination whereas memorability would enhance it. Results favoured their proposals and they further explained the differences in the two typicality components in terms of an atypical face: an atypical face is low in structurally induced familiarity (it resembles few other faces in the recogniser's experience) and high in memorability.

Green and Geiselman (1989) assessed the effects of feature saliency in relation to composites constructed using the IdentiKit system. They noted that in terms of practical application, the presence of salient features negatively affected the quality of the composites; the limitations of the composite system meant that salient features could not be represented. Therefore, where composite quality and subsequent identification should have been at its greatest potential, with the presence of salient features on the target face, this was not the case and, in fact, the composites were less likely to be identified. Green and Geiselman suggested that future systems could benefit from accommodating 'exaggerated' features within their databases.

Much of the earlier research on composites concentrated on construction with the target face in view or with immediate construction (Shepherd et al., 1978; Ellis et al., 1978; Laughery and Fowler, 1980). Psychological research on memory recognises that recall ability deteriorates with time and consequently, contemporary studies have progressed towards more realistic circumstances including construction of the target face from memory, although many of these studies point to the lack of effectiveness of composites produced from memory (Wogalter & Marwitz, 1991; Brace, Pike & Kemp, 2000). Results from a study by Cutler et al. (1988) were promising in terms of the effectiveness of composites produced from the Mac-a-Mug Pro system that gave favourable identification results. However, these results were not borne out in a later study by Koehn and Fisher (1997) who found that composites in their study were of very poor quality and that their utility value was extremely low, when implementing a two day retention interval from viewing of the target to construction. They suggested that the poor performance of the Mac-a-Mug Pro system, in contrast to the Cutler et al. (1988) study, was in the fact that composites in their study were produced from memory, whereas this was not the case in the earlier Cutler et al. study.

Results from a further study by Kovera et al. (1997) also revealed the poor performance of the Mac-a-Mug Pro system for producing recognisable composites from memory. They found that problems with identification arose at the composite generation phase and that composites made with photographs in view were more recognisable than ones made from memory. This led Kovera et al. to question the applicability of the Mac-a-Mug Pro system to real-life situations. The most recent studies on delay interval have implemented delays of a few hours to two days (Frowd et al. 2005a, b) in an attempt to reflect the average time taken for police to approach a witness for the facial imaging interview. The Association of Chief Police Officers (Scotland) (ACPO(S)) recommends that the Facial Imaging Operator or Police Artist contacts the witness within 24 – 36 hours of an incident occurring, to assess the level of recall.

A study by Mauldin and Laughery (1981) assessed the effects of exposure time to the target and the delay time from exposure to the target to composite production on recognition. Two delay intervals were imposed: thirty minutes and two days. The procedure used for the thirty minute delay interval was for participants to either: produce an IdentiKit composite; to complete an Introvert Extrovert scale, or to complete a facial feature description questionnaire. They then returned to complete a recognition task. The participants in the two day delay condition either completed the composite activity and then returned after two days to do the recognition task or left immediately and returned two days later to complete the composite activity and the recognition task. One important finding of the study was that time delay from viewing to construction had no significant effect on recognition. In addition, findings clearly indicated that participants constructing an IdentiKit composite were more likely to recognise the target face in a recognition task. Suggested reasons for this were the verbal processes involved in describing the target face to an operator and also the development of retrieval processes for obtaining the information from memory.

Results did not indicate that improved recognition resulted from composite production due to increased elaboration shortly after target exposure, but that the improved recognition more than likely arose from the development of retrieval processes. The suggestion was that when participants constructed a composite they searched the target face for information, this results in retrieval processes being developed and learned which are more available for the subsequent recognition task. Mauldin and Laughery also found that the exposure duration variable was not significant. In practical terms, Mauldin and Laughery's results indicate that IdentiKit composites may have a more optimal role in enhancing a witness's memory than in being used as facial representations given their poor performance in identifying suspects as reported by earlier researchers.

Green and Geiselman (1989) also studied the effects of time delay and found that a significant effect was only found where faces had non-salient features. In their study, they implemented an immediate and a one week retention interval from viewing to construction and hypothesised that salient features would command attention with brief exposure and that these features could not be adequately represented in composites. The immediate delay condition gave superior results to the week delay but only with non-salient faces. Green and Geiselman put forward the view that their test was a strong one given that the purpose of composites created with the IdentiKit is to eliminate sections of the population rather than to identify specific suspects.

In a study by Bruce et al. (2002) (experiment 2), composites were constructed from memory of unfamiliar faces, in an attempt to reflect a more ecologically valid procedure. The experiment involved using composites from multiple witnesses, whereby four composites of each target were constructed. Participants were then shown all four composites, a combination of all four targets (4-morph), and the best and worst composites. Identification rates were low (greatest identification was 38% where all four composites were shown).

More recent studies on the effects of retention interval by Frowd and his colleagues have attempted to reflect as forensically relevant a procedure as possible, in order to maximise the ecological validity of their findings. In a study (2004) using the recently developed EvoFIT system alongside PROfit and E-FIT, witnesses worked from memory and selected a target face of a person who was unfamiliar to them. In this way, the procedure reflected real-life, where the witness constructing the composite does not know the *crime perpetrator*. A retention interval of two days was implemented to reflect a typical police interview time interval. Naming rates were extremely low and were only 3.6% for EvoFIT, 1.3% for PROfit and no correct identifications recalled for E-FIT. It was considered that the target set were too old and therefore contained salient facial features which could not be replicated with the use of the composite systems (as reported by Green and Geiselman, 1989).

A further study was undertaken, implementing a new target set, with younger males and a single operator to control both the PROfit and EvoFIT systems (to avoid differences between operators). Again, a two day retention interval was imposed and naming results were low but were significantly higher for EvoFIT (8.5%), compared to PROfit (3.7%). The two day retention interval and the fact that witnesses were unfamiliar with the target faces they constructed appeared to have a major impact upon recall ability and subsequent composite quality and identification. A reduction in composite quality would be expected given that memory traces deteriorate rapidly after retention intervals of a few hours and therefore recall abilities at the retrieval stage will be poorer. Subsequently, the resulting composite may not involve as much detail or this detail may be inaccurate when compared to composites constructed with no retention interval and identification levels may decrease.

In a further study, Frowd et al. (2005a) added to the forensically relevant procedure adopted earlier and assessed five composite methods: PhotoFit, EvoFIT, PROfit, E-FIT and Sketch. In this study, witnesses were unfamiliar with the target set (who were young males of both high and low distinctiveness) and a three to four hour retention interval was imposed. Different experienced operators used each of the systems to produce composites through the use of the cognitive interview procedure and were blind to the targets. Target faces were chosen to require minimal artwork and to be sufficiently familiar to a second sample of participants who were asked to evaluate the composites through a naming procedure. Results were that composites constructed using the PROfit and E-FIT systems had higher naming than the other systems. An elevated naming rate was revealed for the faces rated as being highly distinctive, for all systems. The mean naming rate for E-FIT and PROfit of around 20% was found to be similar to naming rates from other research in this area (eg. Bruce et al., 2002; Frowd et al., 2004).

Following on from this, Frowd et al. (2005b) again adopted a similar procedure when they assessed five composite systems (they used the same four systems as in the earlier study but added FACES rather than PhotoFit). This study employed a two day delay interval from viewing the target to construction. Naming results were extremely low (3% overall) in comparison to the previous study and similar research (around 20%). The E-FIT and PROfit composites were correctly named on only two occasions out of around three hundred attempts, FACES were better named at 3.2% and EvoFIT were very slightly better named still at 3.6%. Sketches were found to have the highest naming results of 8.1%, in contrast to the earlier study where sketches did not fair well in comparison with the E-FIT and PROfit composites.

Limitations in verbal descriptions were offered as one reason for the poor naming results recorded. Further, the results were explained in terms of a shift in cognitive processing due to the two day delay interval; after the two day delay, witnesses would have more of an overall impression of the target face and therefore systems that allow for a more holistic approach to composite construction would be favoured (i.e. sketch) over one's that are more feature-based (i.e. PROfit and E-FIT). This is therefore one explanation for the superiority of sketches in this study. Frowd et al. recognised that the use of different operators may have been a limitation of the study and that operator skill is likely to vary. Finally, Frowd and his colleagues suggested that law enforcement agencies should attempt to interview witnesses sooner and that where they are unable to do so, the use of the sketch artist over other composite systems may be beneficial.

This thesis will consider in detail some of the factors discussed above in relation to the construction and resulting identification of facial composites. In particular, this thesis will focus upon the effects of operator experience, target distinctiveness, retention interval and artistic enhancement. As mentioned previously in this chapter, even with advances in composite construction systems, composites still remain poor in terms of identification levels. Psychologists have recognised the role of the operator in the construction process but further research is required to determine the precise part that experience plays. Previous research has suggested that the experienced operator may excel in terms of their level of artistic skill and enhancement techniques, the ability to relate the witness description to the features of the system, or in the initial selection of the features (Davies et al., 1983; Bennett, 1986; Gibling & Bennett, 1994). As Davies et al. (1983) suggest it is likely that experience enables the operator to become more proficient in selecting features which closely resemble the target at the stage of producing the initial composite.

The literature on memory would predict that a superior, more alike composite would result where there were fewer modifications made to the initial composite at the construction process. In this way there is less scope for interference with the original memory to occur. Where the process involves the operator scrolling through a number of features before selecting the 'correct' feature, it is more likely that the witness's original memory for the feature will become displaced and therefore the resulting composite is less alike the target image.

A further suggestion is that the experienced operator is more adept in the use of the cognitive interview procedure to elicit optimal recall from the witness. The literature on memory would predict that optimal recall would result from the inclusion of good retrieval strategies within the recall process, something which the basic cognitive interview (Geiselman et al., 1985) aims to promote. An experienced operator may be able to elicit more information from the witness by the manner in which he approaches the witness and asks them to report what they have seen; for example, in different orders, from different perspectives and by aiding with the mental recreation of the image. The experienced operator is also likely to avoid the use of leading questions so that the witness does not reconstruct the memory to fit with their expectations;

something which psychological research has recognised as a problem since Bartlett (1932) first carried out research in the area.

This thesis aims to determine whether an experienced operator who has undergone training at a recognised police facial imaging course and who has research experience in producing facial composites and in delivering the cognitive interview and skill in the use of the composite system, can produce more readily identified composites when compared to novice operators. It is predicted that skill and experience in the areas mentioned above will be advantageous in producing identifiable composites. It is also of interest to determine whether an experienced operator is able to promote superior retrieval and therefore produce superior composites or whether difficulties in gaining access to the original memory will mean that composite identification levels will remain poor as in previous composite research.

Target distinctiveness is a further area of interest. As discussed previously, the face recognition literature has shown that distinctive faces are more likely to be recognised than are common faces (Winograd, 1981; Courtois & Mueller, 1981; Valentine & Bruce, 1986; Green & Geiselman, 1989; Vokey & Read, 1992). This seems sensible given that the literature on memory in general also recognises the importance of distinctiveness in the retrieval process. Eysenck (1979) put forward the view that long term memory is affected by distinctiveness of processing as well as by depth and elaboration. Eysenck's suggestion was that unique or distinctive memory traces will be more readily retrieved than traces that closely resemble a number of other memory traces. Frowd et al., (2005a) were the first to show a distinctiveness effect with composites. They found an elevated naming rate for composites previously rated as having high levels of distinctiveness. This thesis aims to replicate the findings of Frowd et al., (2005a) in relation to distinctiveness.

It is well recognised in the memory literature that memory traces decay or are displaced over time, whether due to physiological processes or the introduction of new information. It has long been known that forgetting increases with time and that the longer the retention interval, the more scope there is for decay or displacement to occur. Ebbinghaus (1885) was the first to recognise the effects of retention interval and plotted a forgetting curve. He learned nonsense syllables and plotted the amount of effort required to relearn these syllables after various retention intervals. He found that more effort had to be expended to relearn the syllables, the longer the retention interval, as one might expect. The decline in the amount of effort required to relearn the syllables was sharpest immediately after learning and then became more gradual. In fact, the requirement for relearning was dramatically increased after only a few hours but there was little difference between the amount of relearning required between eight and forty-eight hours.

Although participants of facial composite studies are generally aware they are to construct a composite at a later time and are therefore able to adopt rehearsal strategies to increase the likelihood they will remember a target facial image, recall still remains poor. This is especially so with the introduction of a considerable retention interval.

Recent composite studies have implemented retention intervals of a few hours to two days in an attempt to be closer to the length of time taken for a facial imaging interview to take place in a police investigation (Frowd et al., 2004; Frowd et al., 2005a, b). Identification was shown to be at low levels in these three studies and poor verbal descriptions were suggested as one of the reasons for the poor performance.

In contrast with what the general memory literature would predict to happen (ie. that there would be greater recall immediately and this would decrease dramatically after a few hours), there appeared to be little advantage to composites constructed with immediate construction (naming rates of around 20%, Brace, Pike & kemp, 2000; Davies et al., 2000) when compared with composites constructed after a retention interval of a few hours (naming rates of around 20%, Frowd et al., 2005a). Naming was at lower levels with a two day delay (Frowd et al., 2005b) compared to the shorter delays, however, the use of different operators across retention interval are problematic. Given the poor identification levels in these studies with forensically realistic retention intervals it was of interest, in this thesis, to determine whether the use of the same operator across conditions would result in similarly poor naming levels to the Frowd et al. studies. Additionally, whether there would be even poorer naming with a longer retention interval of one week. According to the Ebbinghaus theory of forgetting and retention interval one may expect significantly superior identification of composites constructed after a few hours compared to two days, given the expected superior recall abilities after this time.

A further area of interest within this thesis that has not been studied previously is the area of artistic enhancement of the composite from the witness's memory. Research on facial composites has shown that artistic ability and enhancement of composites is important and composites that have been enhanced are more alike the target image than unenhanced composites (for example, Gibling & Bennett, 1994). The use of artistic enhancement may mean the composite moves closer to the mental image of the target, thereby further aiding the retrieval process through the use of optimal retrieval cues and subsequently may trigger identification or recognition. However, given what is known about retrieval of information from memory, it is of interest here to determine whether artistic enhancement can still be of value when composites are constructed from memory. One would predict that the use of artistic enhancement (ie. the addition of marks and scars, eyebags, shadow etc.) would mean a composite was more life-like in appearance thereby aiding identification. However, with the introduction of a retention interval it will be more difficult to accurately recall facial appearance and therefore problems with reconstruction of the original memory to fit with expectations are likely to occur. This may mean too much enhancement of the composite takes place or enhancement of the incorrect areas of the face occurs, subsequently composite quality will decrease.

The four factors detailed here are studied in greater detail in the next three chapters of this thesis. Chapter 2 discusses operator performance (the extent to which operator experience and skill is required to produce a good quality, identifiable composite alongside the tools of the composite system) and target distinctiveness (the extent to which the target face as a whole, or indeed, the individual features are considered unique and the effect this has upon the quality and identification of the composites). Chapter 3 focuses on delay (retention interval) from viewing of a ‘target’ face to construction of the composite and artistic enhancement (the effects of the use of tools and techniques for elaboration of the composite). Finally, chapter 4 will address the overall findings of the three studies in the context of the wider literature on facial identification from composite production systems, with particular reference to their impact or meaning in relation to practical policing.

2

Composite quality: Can novice operators out-perform an experienced operator?

Introduction

It is current police practice to employ a facial imaging expert to produce a likeness of an offender from an eyewitness account, wherever possible. As the majority of modern composite systems are computerised (E-FIT, PROfit and Mac-a-Mug Pro), there is the view that, as with many computerised products, the result will be achieved more quickly and will be superior to the more traditional methods of composite construction, such as IdentiKit and PhotoFit. Early research into facial composites using these systems reported that composites were not very human in appearance and therefore effectiveness was limited. These systems were also criticised for their lack of available features (e.g. Laughery et al, 1980; Bennett, 1986). Mechanical systems such as IdentiKit and PhotoFit involved the selection of individual features and the construction of these features into a face. Selection of individual features led to lines being formed at the boundaries of features, resulting in limited resemblance to a human face and impairment of recognition (Ellis, Davies & Shepherd, 1978).

Developers of modern systems have attempted to move away from such criticism by including drawing tools and paint packages within the software. These packages allow for the use of various techniques, such as feature blending to provide a more human appearance through the development of skin texture information. Significantly greater facial features with more sophisticated descriptors (for example, eye depth, heavy lines, brows overhanging and so on) and the implementation of editing tools may be advantageous for replication of specific features and the production of a superior likeness. They do, however, require greater input from the operator. It is considered that a greater level of operator experience would therefore be advantageous to the requirement for greater operator intervention. The level of skill and experience required by the operator to use these systems remains undetermined.

Gibling and Bennett (1994) criticised earlier research into operator experience by Ellis, Davies and Shepherd (1978) which reported that recognition accuracy did not vary whether a professional operator or a novice operator were constructing the faces. Gibling and Bennett tested the view that skilled and semi-skilled operators would have the ability to produce significantly better recognised composites through the use of artistic enhancement, when compared to composites that were not enhanced. Results indicated that operators of the PhotoFit system required knowledge and skill in enhancement techniques and they also required artistic ability. They further reported that although the main selling point of computer-generated systems such as Photo-FIT was that they could produce accurate likenesses from unskilled operators, operators actually required a high level of skill to use these systems effectively. This research was in agreement with the views of operators in the field who suggested that the limitations of the composite systems could be overcome with proper training. Earlier research by Laughery et al. (1980, cited in Wogalter & Marwitz, 1991) pointed to the lack of an expert familiar with face recall procedures, as being one of the reasons for the poor performance of composites produced using the Field Identification System (FIS). Bennett (1986) also recognised the advantages of using skilled operators who had an understanding of the psychological factors involved in eyewitness recall.

Laughery and Fowler (1980) also assessed the effects of operator experience and found that sketch artists produced better composites than those constructed using the IdentiKit system. The sketch artist excelled in several areas: infinite numbers of features could be produced and these included as much detail as necessary, a longer time was spent producing the sketches so the witness had a greater amount of time to spend thinking about the target which may have led to a superior representation and further, the sketch artist adopted a more holistic oriented approach, believed to promote superior retrieval. The study also showed that it was the inflexibility of the system itself rather than the operators' ability that affected composite quality. There was little difference between the results from the technicians whereas there was a great deal of difference between the sketch artists. Artistic skills were found to be important and it was suggested that training in this area should be maximised.

A later study by Davies, Milne and Shepherd (1983) also indicated the value of operator experience on the quality of the composite likeness achieved, when using

PhotoFit. Their results supported the view that there was an element of skill involved in the production of a composite and that this skill could be categorised into three areas: eliciting feature information from the ‘witness’ in the form of a verbal description, relating this to the features of the system and, finally, the use of technical and artistic skill to produce the composite. Further, Davies et al. found that it was in the initial selection of the features where the experienced operator excelled and not necessarily in the use of artistic enhancement using pencil work.

Facial research acknowledges the importance of facial feature saliency or distinctiveness in the recognition of faces (Shapiro & Penrod, 1986, Valentine & Bruce, 1986). Courtois and Mueller pointed to the fact that salient or uncommon faces are better remembered in terms of descriptions and identifications than are non-salient or common faces. In effect, composites produced of target images with distinctive features should be more recognisable than those without these features. In their prototype theory of face recognition and the role of feature distinctiveness, Valentine and Bruce (1986) proposed that faces are encoded by comparison to a single prototypical face. They went on to explain that this prototype arises from ‘averaging’ all faces encountered in everyday life; a distinctive face differs from an average face in the manner in which it is encoded. Vokey and Read (1992) further explain facial distinctiveness as ‘atypicality’. A typical face is defined as such due to its general level of familiarity and memorability. An atypical face is low in familiarity because it resembles few other faces and is therefore generally highly memorable.

On a practical level, Green and Geiselman (1989) discussed that replication of salient features was not always possible given the limitations of some facial composite systems such as IdentiKit. Therefore, where salient features were present on the target face, this negatively affected the quality of the resulting composites. This doesn't appear to be the case for more modern composite systems that provide drawing tools for replication of particularly distinctive features. In their study across five composite production systems, Frowd et al. (2005a) reported that faces classified as having high distinctiveness were better recognised than those with low distinctiveness.

A further area of interest in this study was the nature of encoding processes in composite construction. Poor memory recall has been put forward as an explanation for low rates of composite identification. Koehn and Fisher (1997) found that in contrast to Cutler et al's (1998) study, composites constructed using the Mac-a-Mug Pro system were of poor quality. This was explained by a difference in construction process: Cutler et al's participants produced composites with the target in view, whilst the Koehn and Fisher composites were constructed from memory. A later experiment by Davies, van der Willik and Morrison (2000) gave a similar result. They compared participants' performance using E-FIT and PhotoFit and found that, realistic likenesses in the form of composites resulted only when constructing the composite with the target person present, when utilising the computerised system. Composite quality was inferior when participants relied upon their memory. Naming rates of around 49% resulted when constructing composites with the target present (Davies et al., 2000). However, the preferred method with greater ecological validity is to construct from a 'witness's' memory. Unfortunately, naming rates of around only 20% have resulted when constructing from memory (Frowd, Hancock & Carson, 2004; Bruce et al., 2002).

Analysis of encoding strategies employed where a participant is instructed to construct from memory, has shown that feature-based encoding is the preferred method for feature-based construction (for systems such as PROfit) but that holistic encoding is best when a sketch is made (e.g. Davies & Little, 1990). In their study, Ellis, Davies and Shepherd (1978) found that lines or obvious boundaries on a face produced impairment in memory for faces using the PhotoFit system. It was hypothesised that faces are processed in a holistic manner and that lines interfere with this.

Tanaka and Sengco (1997) explain holistic recognition as involving information about the features of the face, and information about their configuration, together. They found that the modification of spatial location of one facial feature impaired the recognition of the other facial features. They concluded that spatial information was not defined by absolute feature position but by position relative to other features. This is highly relevant to the recognition of facial composites where it is perhaps difficult for those constructing to replicate the exact locations of the target facial features in relation to the other features.

Much of the current literature on facial recognition points to the distinct processing strategies involved in the encoding of familiar and unfamiliar faces. Hancock, Bruce and Burton (2000) explained that we recognise familiar faces easily, even those of poor quality, but that our ability to match or recognise unfamiliar faces is poor. Yount and Laughery (1982) suggested that faces are processed as large units and that the more familiar we are with a face, higher levels of holistic processing will result and less individual feature processing. Results from a more recent study by Schwaninger et al. (2002) have challenged the assumption that faces are processed only holistically. Their research results showed that we process familiar and unfamiliar faces by encoding and storing configural information as well as the local information contained in facial parts.

Character attribution has been shown to have an effect on the way in which faces are perceived and remembered. Davies and Oldman (1999) found that, in their study, composites made by those who disliked the target images were ranked higher overall and elicited more correct identifications than those made by individuals who liked them. It was suggested that this occurred because positive judgements encourage a more global evaluation of faces, which is known to improve recognition of faces but impair composite construction.

These various encoding issues and strategies were considered alongside operator experience and target distinctiveness. It was considered that detailed information pertaining to the nature of the encoding that was undertaken at the construction phase would provide an insight into the ability to produce a composite that could be spontaneously named by a further sample of evaluators. A further consideration was the role the operator played in the production of this ‘named’ composite. Perhaps enhanced ability to visualise the target or familiarity with the target would require greater skill and experience from the operator but would result in a superior quality composite that was identified to a good extent.

The main area of investigation of this study was operator performance; the comparison of the results from three novice operators with an experienced operator. The hypothesis was that the experienced operator would out-perform the novice operators in terms of naming scores; greater naming would result from composites constructed by the experienced operator. These results were also compared to the results of earlier researchers whose naming rates were around 20% (Frowd, Hancock & Carson, 2004; Bruce et al., 2002).

A further area of investigation was whether distinctiveness of the targets would affect the identification of the composites. Distinctiveness here refers to the extent to which an individual would differ from others if he were in a crowd of young, white males. Accordingly, what may have caused an individual to be rated as being distinctive could be a distinctive facial feature or a distinctive configuration of facial features to give an unusual appearance overall, or could be an unusual marking or scar on his face. The prediction was that those targets rated as being highly distinctive would receive greater identification results than those that were rated as having lower levels of distinctiveness (as found by Frowd et al., 2005a).

To assess encoding strategies and their impact upon naming, information pertaining to witness participants' visualisation of the target face was gathered. An attempt was made to gain an insight into whether faces were processed holistically or in their individual features and what effect this had upon naming scores. Witness participant's perceptions of their level of familiarity with the target and whether this was related to the naming of the target were also assessed (similar to the scale employed by Davies et al., 2000). Evaluator's levels of familiarity with the targets were also assessed. An attempt was made to determine whether participants' attitudes towards the target person, and their perceptions of the likeness of the target to the composite were related to naming of the targets.

Experiment 1: Comparing operator performance

Method

Composites of famous faces were constructed from memory by a sample of participant witnesses and these were evaluated by a second sample of volunteers, through a naming procedure. The facial composite systems E-FIT and PROfit were used to assess the effects of operator experience for four operators. The results from three novice operators were compared with those of an experienced operator, whose composites were constructed for a previous study, see Frowd et al., (2005a) and therefore differed in target familiarity (familiar for the novice operators and unfamiliar for the experienced operator) and delay interval to construction (immediate for the novice operators and 3 - 4 hour delay for the experienced operator). Targets were famous faces chosen on the basis that they were to be highly familiar to a group of Open University students. This was to allow the main dependant variable to be naming of the composites.

Composite construction

Operators

Three of the operators (SP, KW [females] and RC [male]) were inexperienced and had only limited knowledge and use of the PROfit and E-FIT systems prior to the experiment. Additionally, they had no previous experience in conducting cognitive interviews. These 'novice' operators had some limited practice at composite construction and the use of the composite software and had been tutored in the basic procedure employed in a Cognitive Interview used by police facial imaging operators. Results from the novice operators would be compared to those of a fourth operator (HN, female) with experience in conducting facial composite research. This individual was skilled in producing computerised facial composites and at delivering a cognitive interview. Mean naming rates of the novice operators could be compared with mean naming of other researchers in this field implementing a similar design and procedures (Frowd et al., 2005a).

Facial composite production systems

The computerised facial composite production system PROfit was used to construct composites for two of the novice operators (KW and SP) and for HN. RC used E-FIT, which is of a similar specification (i.e. similar interface and operating functions) to PROfit. Frowd et al (2005a) found that both systems performed equivalently when assessing identification performance across five systems. Both PROfit and E-FIT consist of databases of facial features. They allow the user to scroll through all features and visually assess, or limit the number of features available by selecting various descriptors e.g. straight nose ridge, arched eyebrow etc. Re-sizing and re-positioning of features is possible, as well as altering brightness and contrast and adding marks, scars and other accessories. Both systems have a paint package available for the artistic enhancement of features; shadowing, feature blending and other functions allow the selected feature to be modified to have the appearance of the specific target feature.

Targets

Ten famous male faces were used as 'target' faces and these were rated according to distinctiveness (Frowd et al., 2005a). Targets were popular music artists, footballers, television celebrities, tennis players and movie stars. The targets were: Damon Albarn, Noel Gallagher, David Beckham, Michael Owen, Andre Agassi, Robbie Williams, Noah Wyle, Stephen Gately, Brad Pitt and Craig Phillips.

The procedure used to rate the distinctiveness of the targets was for individuals to be presented sequentially with good quality images of the targets and asked to imagine meeting each person at a railway station in amongst their peers (young, white males) and rate from 1 to 7 (1 = *average, blend into the crowd* and 7 = *very distinctive, stand out from the crowd*). These targets were rated in earlier research by Frowd et al. (2005a). From these ratings, targets fell naturally into two categories; high and low distinctiveness. Target faces rated as having low levels of distinctiveness were: Damon Albarn, Stephen Gately, Michael Owen, Craig Phillips and Noah Wyle. Targets rated as being of high distinctiveness were: Robbie Williams, David Beckham, Brad Pitt, Andre Agassi and Noel Gallagher.

The faces were printed in colour (sized 80 x 150mm on A4 paper) to a high quality and were standardised where possible to be of a neutral expression, front facing and with good lighting conditions. The celebrities had a mean age of 29.2 years (SD = 4.9 years), in an attempt to mirror crime statistics on the age at which the majority of crimes are committed within the UK population (although the peak offending age is known to be much lower than this at 18!). It should be noted that target faces were chosen to limit operator differences by using targets that didn't require a great amount of artistic enhancement; faces were relatively young. They were also chosen to be familiar to a group of Open University students with an average age of thirty years; this sample was to be used to name the resulting composites.

Participants

Witness participants were drawn from a large population: students and staff at Stirling University, students of Aberdeen University and members of the public. Student volunteers from Stirling University were each paid £5 for participation in the experiment.

The participants for Operator RC were 2 males and 8 females and were aged between 20 and 42, with a mean age of 31.7 years (SD = 7.74). Participants were from Aberdeen University and were each given a course credit as motivation to participate in the experiment.

Operator SP's participants were 3 males and 7 females and were aged between 21 and 56 with a mean age of 29.8 years. Operator KW's participants were 5 males and 5 females with a younger mean age of 22.9 years. Operator HN's participants were staff and students of Stirling University and members of the public. Ages ranged from 23 to 53 with a mean age of 37.4 years. They were 6 females and 4 males.

Participants were chosen if they fitted the selected age requirements of between 18 and around 55, had no previous experience in composite construction, with the use of facial composite software and one of the available targets was familiar to them. Individuals with detailed knowledge of the cognitive interview were not selected as participants. Participants were recruited to be involved in an experiment on famous faces and instructed that they would be asked to construct a facial composite. They were debriefed as to the purposes of the experiment following completion of the composite.

Procedure

In order to minimise operator effects during construction of the composites, operators remained *blind* to the identity of the 'target' faces until all composites had been constructed. It was possible for the operator to remain blind as the targets were selected by another person. No information was provided to any of the operators regarding the identities of the targets.

Participants were also asked not to reveal the identity of the target face to the operator at any point during the experiment. The participant was handed an envelope containing the ten pictures of the famous males and instructed to select one at random. If the first target face selected was unfamiliar to the participant they were required to return the picture to the envelope and choose another at random until a target face was selected that was familiar to them (a code was noted pertaining to the particular target). In the absence of familiarity with any of the targets, it was explained to volunteers that they would not be required to take further part in the research and they were also debriefed as to the purpose of the experiment and the reasons they were no longer required. Participants were not made aware that exclusion from the experiment would occur if all targets were unfamiliar to them. In this way, the chances of a participant selecting a target face that was not familiar to them in order that they be included in the experiment were minimised. Target images were not returned to the envelope if they were used by a participant. In this way, they could not be selected by further participants.

The participant was then given one minute to study the picture. A brief description of PROfit or E-FIT was given by the operator lasting approximately three minutes. This covered the basic elements of the software (how to select features with certain descriptors, re-positioning, re-sizing etc. and a basic description of the effects available in the paint package). The participant was then asked to recall as much detail as possible of the facial features of the target face. The procedure used in the recall of the facial features was as near as possible to that employed by police operators in a cognitive interview: two cycles of free recall followed by cued recall. Participants were asked to form a mental image of the target face they had previously viewed. They were then asked to provide the operator with as many details as they could remember of the facial features of the target, in any order and at their own pace – free recall. Interruptions from the operator were minimal and they noted everything pertaining to the facial detail of the face the participant described. This stage was repeated a further time, without participants' awareness that this would happen. The next stage was cued recall, where the operator would repeat back the information given for each feature. The witness is asked to visualise each feature in turn and attempt to add any extra detail they are able to recall.

To attempt to gain an insight into the ways in which the target images were encoded, participants were then asked to visualise the target person and to rate on a scale of 1 to 7 (1 being '*not at all well'* and 7 being '*extremely well'*) the extent to which they could visualise the person in the image. These assessments were made after the subject had undergone the verbal description stage of the procedure and had fully described (to as great an extent as their memory permitted), the facial features of the target person. This way, interference with the memory was kept at a minimum and the memory was still relatively recent. Further questions concerned the amount of facial features that could be visualised: the question posed was 'what amount of the facial features can you clearly visualise?' (scale of 1 to 7, 1 being '*none'* and 7 being '*all'*). A further question concerned the extent to which the relationships between the facial features could be visualised; 'how well can you visualise the relationship between the facial features?' (1 '*poorly*' and 7 '*extremely well'*).

In order to assess the effects of character attribution upon the identification of the resulting composite, participants were asked to rate their attitude towards the person in the picture, on a scale of 1 to 7 (1 being ‘*negative*’, 4 ‘*neutral*’ and 7 ‘*positive*’).

Familiarity, or the extent of the subject’s knowledge of the target person was assessed using a 5 level Likert Familiarity scale (as employed by Frowd et al., 2005b). Participants were asked to indicate whether the face they observed could be classified as one of the following:

1.) not known to them at all

2.) seen the face only and didn’t know anything about the person

3.) say something about why the person was famous but not very confidently

4.) definitely knew the face and could say who he was with confidence

5.) knew lots about the person and the reasons for his fame

These familiarity ratings served as a manipulation check to determine whether the target set was familiar to the witness participants.

The operator then produced an “initial” composite based on the participant’s verbal description, out of sight of the participant. Where the participant was unable to give a detailed feature description for a particular feature, an ‘average’ feature was selected by the operator. An average feature is one which has average descriptors, for example, average nose length and width with a straight ridge etc. Adopting this approach meant that the subject was not distracted by being shown a feature which was very distinctive on initial viewing of the composite. Selecting an average feature did not occur with a high degree of frequency; on only one occasion for each operator.

The participant was then shown the composite and worked with the operator to create and elaborate the initial composite and produce the best likeness of the target face from memory. This meant using any of the composite systems effects (re-sizing, repositioning, changing contrast/ brightness, erasing or adding to the feature etc.) in order to produce features that were a close match to the target person's features.

Following completion of the composite, the subject's opinion regarding the similarity of their composite to the target image and their confidence in this judgement was then elicited. The assessment of likeness was again based on a scale of 1 to 7 (1 being '*poor*' and 7 being '*excellent*'). A likeness-confidence scale was also employed to assess the level of confidence participants had in their likeness ratings (1 being '*not at all*' and 7 being '*extremely confident*').

The procedure was repeated in the same manner for the additional 29 participants across all three novice operators. The fourth operator's (HN) methodology differed in that participants were unfamiliar (all others were familiar) with the targets and they had a 3 - 4 hour delay (all others had immediate construction) from viewing the target to construction of the composites. Ratings pertaining to visualisation etc. were not collected from Operator HN: HN's composites were constructed for a previous study (see Frowd et al., 2005a), prior to this experiment being undertaken.

Composite evaluation

Design

The experimental design was of two factors: operator (between subjects factor with 4 levels, Operators 1 - 4) and distinctiveness (within subjects factor with 2 levels, high and low). Composite quality or identification rate was assessed by asking participants to name composites presented to them serially (random order across participants)– an important test of composite quality. Naming was uncued i.e. participants were not shown the target image or given any other information that would alert them to the identity of the person prior to being shown the composite. The role of feature distinctiveness was assessed by presenting participants with five composites which were highly distinctive and five which were of low distinctiveness.

Participants

Each operator's group of participants were of the same size and had similar demographics. Operator RC evaluators were 9 males and 9 females with a mean age of 29.2 years; 9 males and 9 females with a mean age of 30.5 for operator KW; 10 males and 8 females with a mean age of 31 for Operator SP. Operators SP, KW and RC drew their participants from the various companies around Stirling University's Innovation Park. Operator HN's evaluators were Stirling University Open University students. They were eighteen males and were aged 18 to 42 with a mean age of 33.4 years (SD = 9.5).

Procedure

Participants were tested individually. They were shown composites from only one operator (i.e. the composites were not randomised across operator) and were told that they would be presented with composites constructed from images of ten famous males. Specifically, these famous males were pop stars, sports personalities etc. The order in which composites were presented was randomised across participants. They were asked to try to name each composite as it was shown to them. In an attempt to help participants answer freely, they were told that it was not their ability that was being tested but the quality of the composites and that they should not be afraid to voice any answer that came to them no matter how incorrect they considered it to be.

All answers were recorded. A correct answer was one where a composite was named correctly or where information regarding the profession or any other information that would allow that person to be identified as a specific individual was given. Following naming of the composite, the subject was shown the target image and again asked to identify. This naming of the target image ensured that where participants could not name the composite image, this was not simply because they did not know that particular person. Only evaluators who were able to correctly name 5 or more target images were included in the experiment. Without knowing at least half of the targets, participants would not be able to name many of the composites. Evaluators' perceptions of the extent to which the target set was familiar to them were also collected. The same scale was employed as for the witness participants.

Results

Composite naming across operators

Uncued composite naming was used to test for identification of the facial composites across all four operators. The composite with the highest naming was David Beckham for all three novice operators (conditional naming rate by items: 33.3% for SP; 55.5% for KW; 77.7% for RC); this was not so for the experienced operator (Operator HN) whose most recognised composite was Robbie Williams (conditional naming rate of 66.6%). Figure 2.1 provides the composites constructed of David Beckham for all four operators.

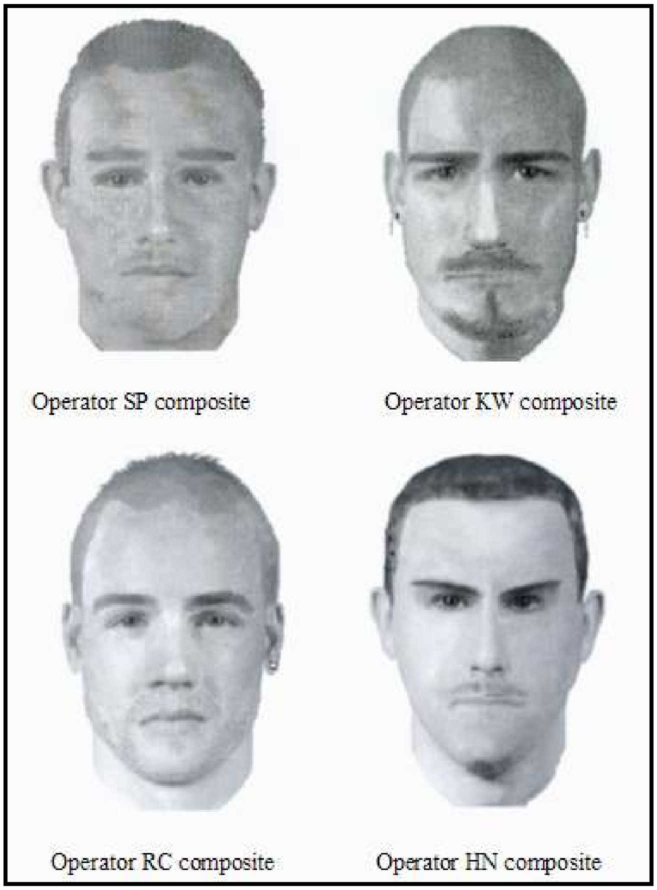

Figure 2.1: Composites of David Beckham for each operator

What is notable is that both targets (David Beckham and Robbie Williams) are rated as being highly distinctive. Additionally, those composites that were named at least once across all four operators were: Noel Gallagher, Michael Owen, Robbie Williams, David Beckham and Brad Pitt. Only one of these targets was rated as being of low distinctiveness; Michael Owen. All others were rated as being highly distinctive.

Mean subject naming rates for the three novice operators were low when compared to previous researchers mean subject naming rates of around 20%: Operator SP (mean = 11.2%, SD = 10.4), Operator KW (mean = 9.6%, SD = 8.2) and Operator RC (mean = 8.8%, SD = 6.5). Mean subject naming rate for the experienced operator (Operator HN) was 17.9% (SD = 12.1). Mean subject naming scores were conditional; they take into account the number of targets correctly recognised (a measure adopted by Frowd et al., 2005a to allow for differences in target familiarity between evaluators). Conditional naming rates were calculated as subject naming score (out of a possible 10 composites) divided by target naming score (out of a possible 10 targets) and expressed as a percentage. Figure 2.2 shows mean subject naming rates across all four operators. Initial viewing indicated an obvious difference for the experienced operator (Operator HN) when compared to the novice operators.

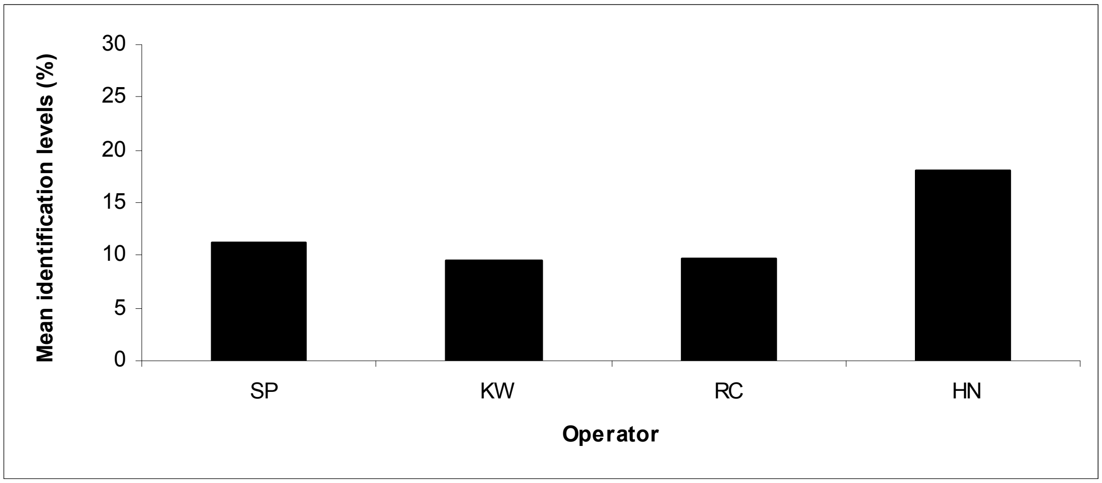

Figure 2.2: Mean identification levels (naming) across operators

Further analysis of Operator HN's results suggested that there was an 'outlier' composite in the form of Michael Owen. This composite was recognised particularly well given the fact that it was of low distinctiveness; 8 out of 18 participants correctly named this composite. It was considered that Michael Owen was particularly familiar to the pool of evaluators used for naming of Operator HN's composites: Open University students. In order to determine whether this was the case, the composites were tested for naming a second time with participants from the same pool of volunteers used for naming for the composites constructed by the three novice operators; from companies at Stirling University Innovation Park. Operator HN's second group of evaluators were 8 females and 10 males with a mean age of 34 (SD = 8.8) years and a range of 20 to 50 years. Mean subject naming rate for the new pool of volunteers was 9.2% (SD = 9.5). The effect size for the experienced and novice operators naming rates was calculated to be –0.08.

Target naming rates were Operator SP (M = 74.4, SD = 16.5), Operator KW (M = 83.3, SD = 15.0), Operator RC (M = 86.1, SD = 19.7) and Operator HN (M = 81.6, SD = 17.9). Target faces were therefore highly familiar to the participant evaluators.

Composite Naming: Target distinctiveness

All data analysis from this point has been undertaken using data from HN's new pool of evaluators. The effects of distinctiveness on naming rates were investigated. Figure 2.3 below illustrates the effects of low and high distinctiveness on composite naming for all four operators, by subjects. There was a major difference between the number of composites recognised, that were previously rated as being highly distinctive and those that were rated as being of low distinctiveness. Mean conditional naming rates were: Operator SP (low distinctive targets M = 7.0%, SD = 11.3; high distinctive targets = 15.6%, SD = 18.6); Operator KW (low distinctive targets M = 1.47%, SD = 5.89; high distinctive targets = 17.1%, SD = 14.6); Operator RC (low distinctive targets M = 0%, SD = 0; high distinctive targets = 15.5%, SD = 11.2) and Operator HN (low distinctive targets M = 1.5%, SD = 4.7; high distinctive targets = 16%, SD = 16.8). Effect size for the naming rates for the high and low distinctiveness conditions was calculated to be 1.07.

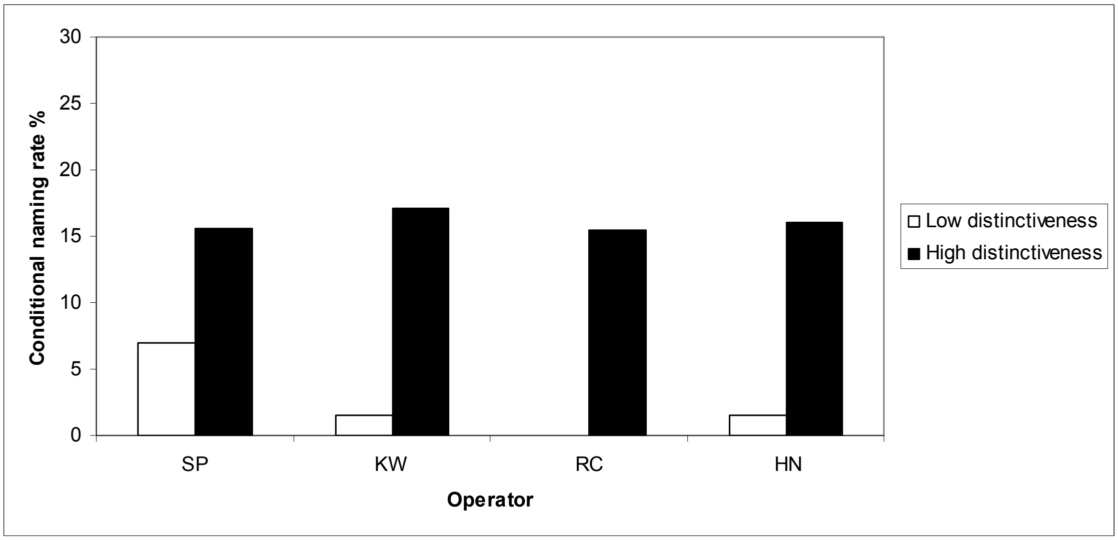

Figure 2.3: The effects of distinctiveness on composite naming; conditional naming rates across operators

A two way, repeated measures, mixed ANOVA by subjects was undertaken. The two factors of the ANOVA were: operator (between subjects factor, with 4 levels; Operators SP, KW, RC and HN) and distinctiveness (within subjects factor, with 2 levels; high and low distinctiveness). Conditional naming rates were used. The main effect of distinctiveness was highly significant ($F_{1, 68} = 48.275$, $p = < 0.01$), with those target faces rated as highly distinctive having greater naming accuracy. The operator by distinctiveness interaction was not significant ($F_{3, 68} = 0.464$, $p = 0.708$). The main effect of operator was not significant ($F_{3, 68} = 0.358$, $p = 0.783$).

A two way, repeated measures, mixed ANOVA, by items, was also undertaken. The two factors of the ANOVA were: operator (between subjects factor, with 4 levels; Operators SP, KW, RC and HN) and distinctiveness (within subjects factor, with 2 levels; high and low distinctiveness). Conditional naming rates were used. The main effect of distinctiveness was significant ($F_{1, 16} = 5.274$, $p = 0.035$), with those target faces rated as highly distinctive having greater naming accuracy. The operator by distinctiveness interaction was not significant ($F_{3, 59} = 0.062$, $p = 0.979$). The main effect of operator was not significant ($F_{3, 16} = 0.068$, $p = 0.976$).

Participant witness/ evaluator ratings

As mentioned previously, a number of further ratings were obtained from participants employed by the three novice operators, to attempt to study some encoding issues at the time of construction of the composite. The level of familiarity the witness participant (at construction phase) had with the target image and the effect this had on the overall naming scores for each target was an area of interest. Mean familiarity rating was 4.4 (SD = 0.73). In order to study this effect, a Pearson's R correlation was undertaken between familiarity rating and overall composite naming rate ($r = 0.073$, $n = 30$, $p > 0.05$). A significant correlation did not result.

Further analysis of naming scores and target familiarity involved the "stripping away" of familiarity ratings for those composites that were not recognised; familiarity ratings for only those composites that were named were employed. This time familiarity ratings from those who were asked to 'recognise' the composites were used. Of the 44 occurrences of naming, across the three novice operators, only four evaluators gave a familiarity score of 3; every score other than that was a 4 or a 5. No ratings below 3 on the familiarity scale were given.

In order to perform a Chi-Square test between naming and familiarity, the data was analysed according to whether the composites were recognised or not and whether they were given high or low familiarity ratings. The data was collapsed into those scoring 4 and 5 on the scale (high familiarity) and those scoring below 4 (low familiarity); expected frequencies of less than 5 resulted prior to collapsing the data. The familiarity scale allows for this natural separation; categories 4 and 5 represent those instances where the target was definitely known to the witness participant, all other categories represent only partial knowledge of the target. A Chi-square test suggested there was a significant relationship between naming and familiarity ($X^2 = 7.511$, df = 1, $p < 0.01$). For those composites that were recognised, only 2.3% were rated as being of low familiarity and 97.7% were rated as being highly familiar. For those composites that were not recognised, 18.9% were of low familiarity and 81.1% were of high familiarity. Overall mean familiarity ratings were $M = 3.75$ ($SD = 1.3$) for the participant evaluators.

An interesting result was found when assessing the relationship between how well a participant could visualise a target face (after viewing for one minute and giving a verbal description) and composite naming rate. Results were not available for one of Operator KW's participants. A Pearson correlation was approaching significance towards a negative relationship ($p = 0.094$) between visualisation ratings and composite naming rate ($r = - 0.316$, $N = 29$, $p > 0.05$). When attempting to establish a relationship between the amount of features that participants were able to visualise and naming rates, no significant relationship was obtained ($r = 0.150$, $N = 29$, $p > 0.05$).

Participants' perception of their ability to visualise the relationships between the target's facial features and naming rates, did not result in a significant correlation ($r = -0.111$, $N = 29$, $p > 0.05$).

Participants' ratings of how alike they considered their resulting composites to the target were collected but no significant correlation was found between this aspect and composite naming rates ($r = 0.256$, $N = 29$, $p > 0.05$).

A correlation was not obtained between participants' attitude towards the target and composite naming rates ($r = - 0.121$, $N = 29$, $p > 0.05$) and therefore attitude toward the target did not appear to affect the resulting composite.

Discussion

This experiment aimed to assess the effects of operator experience upon facial composite naming, using PROfit and E-FIT. Three operators with little or no experience in facial composite construction created composites of famous male faces, where participants were familiar with the target and these composites were tested for identification via a naming procedure. The same targets were constructed by an experienced operator and the results compared. The procedure used to create the composites aimed to be as realistic as possible (minimal target viewing period, composite produced from memory, no previous experience in composite construction, uncued or spontaneous naming of highly familiar targets) and closely followed a technique used by police forces throughout the UK (for example, the cognitive interview technique with no time limit on construction of the composite) and therefore practically applicable to current policing procedures employed.

Initial results indicated a strong difference between naming rates for those composites constructed by the experienced operator and all other operators. Further investigation and additional naming data suggested there was no significant difference between the performances of any of the four operators. Interpretation of the results is problematic when considering the differences in construction methodology for the experienced operator; greater retention interval (3 - 4 hours rather than immediate) and witness participants unfamiliar with targets. These methodological differences may have been disadvantageous to the experienced operator, resulting in naming scores which were not significantly different to those of the novice operators. The experienced operators' witnesses were expected to recall information from unfamiliar faces and with a much greater retention interval.

It may be considered that participants constructing composites after a four hour delay would have greater difficulty in recalling facial feature detail than those with immediate construction. However, the advantage the novice operators' participants had over the experienced operators in retention interval may have been minimised by the fact they were constructing composites of familiar persons. Constructing a composite of a familiar person is perhaps more difficult than constructing an unfamiliar person; one would perhaps have a more detailed memory of a familiar person but this detail is maybe more difficult to replicate within the confines of a facial composite. The findings here, however, in relation to the level of familiarity witness participants had with the target and naming results, suggest this is not the case. There was a significant association between these two variables showing that the majority (97.7%) of those that were named were rated as being highly familiar to the witness participants. The differences in procedure for the experienced and novice operators make it difficult to draw conclusions surrounding operator experience and the data should be interpreted with caution in light of these differences.

Although problematic to interpret, these results may still allow for a conclusion towards no significant differences between the novice and experienced operators, in the light of recent research findings concerning target familiarity. Recent research by Davies et al. (2000) and Frowd et al. (in press) has put forward the view that composite quality is not affected by target familiarity. If this was the case, one would not expect the differences in procedure relating to target familiarity adopted by the novice and experienced operators to have had a major impact upon the results. Much of the facial recognition literature does stress the distinct processes involved in familiar and unfamiliar recognition and therefore Frowd et al. and Davies findings require further analysis. The tests of association that were undertaken here did reveal that named composites were rated by those constructing them as being highly familiar 97% of the time. The differences in retention interval across the two studies remains and the effect of this is unknown.

One similarity that can be drawn from the results is that the novice operators' performance did not differ significantly. Naming results across all three novice operators was very similar. This result seems sensible given that each of the operators had very little practical experience in using the composite system and no previous experience administering the cognitive interview. One may have expected to observe slight differences in the quality of the composites (in terms of artistic enhancements and the ability to choose appropriate features on initial selection) produced by each of the operators due to differences in level of artistic skill. Differences may have been apparent but may have been too subtle to be noticeable at the evaluation stage of the experiment, or may not have been apparent due to problems with replication of features using the composite system.

The results obtained here, in relation to the performance of the novice operators in comparison with the experienced operator, may be in contradiction to those reported by Gibling and Bennett (1994) who found that operators needed to be highly trained and skilled if they were to be effective at producing facial composites. Here, the novice operators produced composites with similar naming to the experienced operator. The experienced operator may, however, have had superior naming had procedures remained the same across all operators. The present results support the finding by Ellis, Davies and Shepherd (1978) that recognition accuracy did not vary between a professional and a novice operator. These results are also in agreement with those of Laughery and Fowler (1980) who found little difference between the performance of operators of the IdentiKit system in their study. Again, comparison with previous research is problematic given procedural differences. Additionally, the results from Ellis et al. and Laughery & Fowler are based on older composite systems which did not allow a great deal of elaboration to be implemented. This could be one explanation for the lack of difference in performance between operators in these two studies. In contrast, one would expect a difference in performance using the E-FIT and PROfit systems as they do allow for feature elaboration.

Although not a professional operator, Operator HN had previous experience in facial composite production (lab-based researcher who had constructed composites for past research projects) and had undergone training at an accredited facial composite course provided for police operators, covering the practical aspects of the facial composite software and the cognitive interview procedure. It is difficult to draw conclusions concerning the effects of experience in composite production in general, and, specifically, the effects of training in the use of the software and the cognitive interview given the procedural – related problems discussed previously. Additionally, one could argue that the design used here did not specifically address the question of whether the experienced operator was more adept at implementing the cognitive interview or in the use of the software. Formal assessment of these factors was not undertaken and therefore any naming differences that did exist between the novice and experienced operators could not be specifically attributed to their effectiveness in either of these two areas. One conclusion that can be drawn, given the alarmingly low levels of naming for the experienced operator, is that training does not necessarily lead to high composite identification levels. This exemplifies that there are factors other than operator training involved in producing well recognisable composites and that training may not be a necessity.

Again as in previous research, target distinctiveness played a critical role in the naming of the composites (Courtois & Mueller, 1981; Valentine & Bruce, 1986; Shapiro & Penrod, 1986; Green & Geiselman, 1989; Vokey & Read, 1992). Distinctiveness effects were studied here by asking witness participants to construct faces that were previously rated for distinctiveness. These target faces fell into two categories: high and low distinctiveness. Faces rated as being highly distinctive were better recognised than those rated as being of low distinctiveness. This finding was consistent across all four operators and is line with Vokey and Read's (1992) finding that atypical or distinctive faces are high in memorability which enhances identification. The results here replicated those of Frowd et al. (2005a) who first found a significant effect for distinctiveness with composites. The effect was apparent across five systems: E-FIT, PROfit, Sketch, Photo-Fit and EvoFIT, with faces of high distinctiveness being better recognised than those of low distinctiveness.

These results are in contrast to those obtained by Green and Geiselman (1989) who suggested that there was increased recognition performance with non-salient faces. This difference may be explained by the use of the older, less flexible IdentiKit system. Limited facial features were available and exaggeration of a specific feature using a pencil was not satisfactory for increasing subject's satisfaction towards the resulting feature, likeness ratings or more accurate identification. Modern composite systems provide the user with greater ability to reproduce distinctive features; they include greatly increased numbers of features in the database and the use of editing tools to specifically modify features.

Ratings were also collected on the ‘witnesses’ opinions towards various aspects of the composite construction phase. It was hoped these ratings would give an insight into the reasons some composites are more recognisable than others. The level of familiarity participants had with the targets was an area of interest; for both witness participants and evaluators. High levels of target familiarity did not produce high naming scores for composites and furthermore there was no significant difference between the average naming rates obtained for the novice operators (where constructors were familiar with the target set) and the experienced operator (where constructors were unfamiliar with the target set). Importantly though there was a significant association between named composites and high levels of familiarity (from those constructing), as mentioned previously. A significant correlation was not obtained between naming and constructors familiarity ratings.

The fact that composite identification scores are low in general (around 20% when compared across a number of studies, Brace, Pike & Kemp, 2000; Davies et al., 2000, Frowd et al., 2004), however, suggests that familiarity with the target is only one of the factors leading to superior composite quality. In their study using the E-FIT system, Davies et al. (2000) found that performance was no better with this modern system compared to studies using the older, mechanical systems, when composites were constructed from memory, even where participants were familiar with the targets. The naming scores here are similar to the average naming scores of 10% found by Frowd et al. (2005a) in their analysis of five composite construction systems, where delay interval was 3 - 4 hours. Even lower naming scores (3% overall) were found in Frowd et al. (2005b) where delay interval from viewing of the target to construction was manipulated (a delay of 2 days was implemented compared to immediate construction as employed here).

One suggestion for the low naming scores here could be that this sample of targets were simply not very familiar to this particular pool of evaluator volunteers. However, this was not the case as mean familiarity rating for the novice operators evaluators was 4.4; the targets were highly familiar to this sample. Alternatively, perhaps it was simply that high levels of witness and evaluator familiarity with the targets resulted in the poor naming results. A witness may have found it difficult to replicate the facial appearance of a target with whom they were highly familiar. Poor composites, where target faces were not replicated to a good extent, would not lend themselves to naming by an evaluator who had a high level of familiarity.

A further suggestion for these low naming rates may be explained by the work of Tanaka and Sengco (1997). They found that configuration of the facial features affected participants' ability to recognise faces holistically. They suggested that a feature's spatial information is not defined by its absolute position on a face but by its position relative to the other facial features. Facial composites are not an ideal means for replication of the exact configuration of facial features on a face and therefore identification of faces in this context may be limited, especially where participants afford greater attention to holistic processing of the face rather than individual feature processing in the construction phase. The fact that composites in this experiment were constructed using the PROfit and E-FIT systems, which involve selection of the individual features of the face in a serial manner, is likely to have facilitated difficulties in retrieval of the features. These features are likely to have been encoded in a more holistic manner, as part of the face as a whole rather than as individual units.

The ability to visualise a target image and composite naming was a further area of interest in this experiment. It was considered that participants own ratings of how well they could visualise the target's face may provide further insight into how the visual memory of the target can aid composite quality and subsequent naming. Perhaps where participants had accurate perceptions of their visual memory and could visualise the target to a good extent, a superior composite resulted which was more recognisable. Results were that there was no correlation between participants' ratings of the extent to which they could visualise the target face and composite naming. However, there was a trend towards significance ($p = 0.094$) but as a negative correlation.

This finding is perhaps counter-intuitive since we would expect to find greater naming of the composites with higher visualisation ratings (i.e. a positive correlation). However, these results do seem to fit with the explanation that participants here were processing the faces in a holistic manner. Therefore, they appear to be paying attention to the face as a whole rather than to the individual features, meaning that attention was not given to the distinctive elements of the face. The result of this could be that participants could produce an overall likeness in the form of a composite but this composite may not necessarily have replicated the distinctive aspects of the target face, thereby not aiding subsequent naming.

The fact that average familiarity ratings were high for the sample of constructors (average rating was 4.4 on the scale), is further evidence that participants may have been employing a holistic encoding strategy. Previous research by Yount and Laughery (1982) defines visual familiarity with the face as dealing with a large unit of visual information. They suggest that the higher the level of familiarity we have with a face, the more we process it holistically. Again, naming will not be aided as attention is not paid to the feature elements of the face. The findings here in relation to naming and the level of familiarity witness participants had with the targets do not correspond well with this explanation. Where composites were named, witness participants had high levels of familiarity with the targets, for the most part; familiarity appears to have aided identification.

An alternative explanation for the results may be that it is extremely difficult to create a composite of someone with whom we have a high level of familiarity simply because the composite just doesn't appear to reflect the facial appearance of that person. Composites, by nature, do not fit with our perception of the human facial form and we may never be satisfied that we have a good quality representation. This would explain the trend towards the negative correlation between naming levels and the extent to which participants could visualise the targets face. Although participants could visualise the target face to a good extent, this did not mean that they were able to represent that person well in the form of a composite.

Further evidence that participants may have been employing a holistic encoding strategy comes from the fact that no correlations resulted on any other visualisation measures. When assessing the relationship between the amount of features that participants could visualise and naming rates, no relationship resulted. Similarly, no correlation was evident between participants' perceptions of how well they could visualise the relationships between the targets features and naming rates. Visualisation of the individual features on the face and the relationships between them did not appear to provoke a great deal of attention; it was the face as a whole that participants concentrated on.

The experiment here may have benefited from a different methodology for looking at the effects of visualisation. The design employed was not the most suitable for gaining insight of visual memory. Ratings scores are extremely subjective; one person's consideration of what the number three on a ratings scale represents may differ from that of another individual.

This experiment did not replicate previous research on character attribution. In particular, the findings of Davies and Oldman (1999), who found that composites made by those who disliked the targets were ranked higher overall and elicited more correct identifications than those constructed by people who liked them. In the present experiment, no correlation resulted between participants' attitude towards the target and composite naming rates. Therefore, attitude toward the target was not a good predictor of naming of that target. Again, the reason for the lack of correlation may be due to the design of the experiment; the manner of assessment may not have been the most suitable. Participants were asked to rate whether they had a negative, positive or neutral attitude towards the target person. The explanation given to participants for assessing positivity or negativity was to decide whether they viewed the target person to be friendly or unfriendly, intelligent or unintelligent, or to assess on any other measure they thought an appropriate representation of positive, negative or neutral. This could have caused some level of confusion among participants and it may have been easier for them to categorise their attitudes into positive, negative or neutral.

It was considered that participants own ratings on how alike they thought their composite was to the target picture may be related to the quality of the composite and resulting naming rates. This was not the case and no correlation resulted between naming rates and participants' likeness ratings. These results seem sensible given that composites judged to have a high level of similarity to a target by the witness participant constructing will not always be identified by an independent evaluator. Additionally, even where a witness participant judges a composite to have a low level of similarity to a target, an independent evaluator may be able to identify the composite.

This experiment could be considered to lack forensic validity in several respects; one being that the targets chosen here were familiar to the witness participant for the novice operators. Where an offender is known to the victim of a crime, it would not be necessary to construct a facial composite; composites are used where the identity of the offender is not known to the witness. A further confounding factor is that the average age of the target set was 29 years. The peak age of offenders is in fact much younger than this at around eighteen years for males in England and Wales (Home Office, 2001). In addition, the retention interval between viewing of the target and composite construction for those witness participants employed by the novice operators was around only five minutes, or immediate construction. Although the Association of Chief Police Officers (ACPO(S)) guidelines on facial imaging suggest that a facial likeness should be elicited from the witness within 24- 36 hours, this is not always practical and it would be rare that immediate construction would take place.

As with similar research in this area, naming of the composites was at a low level. The experienced and inexperienced operators shared similar naming results, although it is difficult to draw any firm conclusions surrounding the effects of experience given the procedural differences detailed in these two conditions. Formal training in the use of the PROfit system and the cognitive interview do not appear to be advantageous to producing recognisable composites, a serious problem in practical terms. As in previous research by Frowd et al. (2005a), similar levels of identification were found from both the PROfit and E-FIT systems. Distinctiveness of the target features emerged as a major factor in the naming of the composites.

There are a number of ways this experiment could be improved upon to provide a more insightful understanding of operator experience in facial composite construction. Firstly, one could implement a target set requiring a high level of artistic enhancement at the composite construction stage; these targets would probably be older and have age-defining features such as eye bags, heavy jowls, nasiolabial folds etc. or would have unique markings that would require skill in elaboration techniques to replicate. A further improvement would be to analyse the content of the cognitive interview to assess the experienced operator's ability to extract detailed feature descriptions from the witness. It would be of interest to determine whether the experienced operator can gain superior descriptions from the witness participant and the retrieval techniques used to do so. Additionally, one could analyse the extent to which the experienced operator is able to produce an initial composite that closely resembles the target; perhaps the experienced operator excels here as suggested by Davies, Milne & Shepherd (1983). One could analyse the number of changes made, or the time taken, from the initial composite to the final version of the composite in order to assess this. The obvious advantage of fewer modifications to the initial composite comes from the fact that there is likely to be less interference with the witness participant's memory of the target face.

Procedural modifications that one might implement to allow for enhanced understanding would be to compare novices and operators using the same composite system; differences in system design may mean that results cannot be attributed to differences in operator experience but in the systems themselves. The use of unfamiliar targets would allow greater ecological validity than using familiar targets and it would be advantageous from a theoretical perspective to use immediate construction. Decay or interference with the visual memory are more likely to result with the introduction of a retention interval and therefore one would expect more detailed and accurate descriptions with immediate construction. With increased verbal description content, there may be greater opportunity for the experienced operator to excel with the inclusion of enhancement techniques etc.

Assessment of the various forms of enhancement techniques application methods, appearance and overall identification value to the composite would be of value. There may be some techniques that are simple to apply but add a great deal to the overall appearance of the composite. Alternatively, some techniques may require a great deal of effort on the part of the operator and do not aid identification to a great extent. Identification results may benefit from the additional measure of an alternate forced choice task alongside the naming task. This would allow for direct comparison of composites produced by the novice and experienced operators and would not be reliant upon evaluators' identification ability as is composite naming.

The issues raised in this chapter will be discussed later, in more detail, (chapter 4) with reference to their practical application to policing.

3

The effects of artistic enhancement and retention interval on resulting composite quality

Introduction

Police investigations are often heavily reliant on an eyewitness's account of a perpetrator of a crime. Facial imaging techniques employed by trained personnel such as the facial imaging specialist or the Sketch Artist can play a vital role in the detection of a crime perpetrator, by helping a witness to extract a memory to recall an individual's facial appearance.

Current UK police practice endorses the production of a facial likeness of the crime perpetrator wherever possible. The two main methods in use are: a 'composite' from computerised facial composite systems or, a sketch of the individual's facial form. Trained artists or operators can help to elicit more 'memory' information of the appearance of the perpetrator through the process of a cognitive interview (Geiselman et al., 1986). Research findings vary as to which of the two methods for producing a facial likeness is superior. However, there is agreement from research examining composite quality that composite construction techniques do not produce accurate representations of target faces (Laughery & Fowler, 1980; Wogalter & Marwitz, 1991; Koehn & Fisher, 1997).

One explanation for this lack of quality given in the literature is that of inflexibility of the facial composite system itself (Brace, Pike & Kemp, 2000; Frowd et al. 2005a). The earliest system used alongside the traditional sketch artist was the IdentiKit system. This was followed by among others, the PhotoFit system. Studies have suggested that limited databases of features (and therefore feature combinations) (Yount & Laughery, 1982) and poor representations of real faces are problematic (Laughery and Fowler, 1980) for composite systems. Yount and Laughery (1982) point to the fact that participants involved in their research, using the Field Identification System (a book-like device for selecting the various features by turning pages) were frustrated by the fact that they knew what the target looked like but could not find features that fitted to an acceptable extent.

It was hoped that with the development of computerised facial composite systems, such as E-FIT, PROfit and FACES, some of the criticisms of the older non-computerised systems would be alleviated to some extent. The Mac-a-Mug Pro was one of the earliest developed computerised systems. Amongst the advantages this system had over earlier systems was a greater number of feature combinations (Cutler, Stocklein & Penrod, 1988). The computerised systems, and indeed the PhotoFit system, include databases of features compiled from photographs and should therefore be more human in appearance than the older formats. In addition, the use of paint package functions (editing tools) allow for a more natural appearance as features can be blended to ‘fit’ the background face. Where the available database features are not considered a sufficient match to the specific facial feature described, the paint package can also be used to modify features to reflect uniqueness.

The argument as to the most effective technique for producing a facial likeness should perhaps be discussed in a wider context. It is debatable whether it is the line drawing (in the form of sketches or line drawing systems) or the photographic likeness which offers the best representation of a face. Early research focused on whether a sketch artist or a non-computerised composite production system would produce the most accurate likeness. Davies and Christie (1982) question whether a realistic portrait which may be incorrect in detail but which may be readily related to faces in memory, or whether a schematic representation conveying only what is known, and which is very simple and will therefore trigger recognition should be used to create facial likenesses. Initially it was surmised by Davies, Shepherd and Ellis (1978), that the systems involving line drawings provided the greatest flexibility and would therefore give a more accurate representation. However, with the advancement of computer graphics this is no longer necessarily the case as there is more realism in the sense of photographic likeness that more modern systems provide. A later study by Davies and Christie (1982) offered no support for the view that line drawings would be superior to photographs for conveying likeness information and retaining it over time. Further research by Davies (1983) also provided no support for this view. In fact, Davies found that when the mode of representation was held constant from study to test, photographs were better stimuli than line drawings. Davies suggested that line representations would be ineffective for recognition for persons unknown to the observer but where persons were familiar, identification could be very high from line material.

Laughery and Fowler (1980) put forward the view that a sketch would be superior to a composite produced using IdentiKit. They proposed that the sketch artist would excel at producing an accurate facial representation as they spent more time enhancing the faces, by way of shading to add contour and they could produce an infinite number of facial features. The sketch artist appears to have a more holistic approach whereby shapes and relationships are considered as one unit. In contrast, the IdentiKit and other feature-based systems have a finite set of facial features. Laughery and Fowler's results suggested that more highly regarded facial representations resulted from construction techniques that used holistic information. Davies et al. (1990) also recognised the effectiveness of the sketch artist over the composite technician. They provided the explanation that the advantage of the sketch artist over the technician, may be in the inflexibilities of the composite system itself; the composite system permits the witness to provide information only on feature detail rather than on the relationships between the features. In essence, the sketch artist allows the witness to undertake a holistic approach by providing information on the face as a whole and working on groups of features and their relationships, rather than on its parts, as does the composite system.

To date, even with the evolution of composite systems to include features with a more human appearance, composite quality remains poor. The ability to produce an accurate facial representation therefore seems to require some level of artistic ability or technical skill in order to assess the face as a whole and to reflect its contours, depth and so on. Davies, Milne and Shepherd (1983) suggested a possible reason for the lack of quality in composites as being due to operator ability in terms of technical skill. This skill included operators' ability to assemble the composite; their ability to modify facial features with the use of pencil work and so on. Their study illustrated the influence of operator experience on the likeness achieved by the subject, with sorting accuracy twenty percent better with composites from an experienced operator compared with a novice operator, indicating that experience and skill plays a part in the production of quality composites.

Bennett (1986) discusses the role of the operator in relation to the fact that the PhotoFit system was rarely able to give a witness with a detailed memory for the target face a satisfactory likeness. Bennett explained that there was a general lack of artistic skill among operators of the systems and that it was important that operators were trained to enhance the PhotoFit using whatever artistic abilities they possessed. A later study by Gibling and Bennett (1994) again revealed the value of enhancement of the composite. Participants were shown enhanced and unenhanced composites from photo-spread arrays with target either present or absent and asked to determine whether the PhotoFit they were presented with was present in the photo-array. Their results provided support for the view that operators with experience and knowledge of enhancement techniques produced superior quality likenesses, than those achievable using only basic kit and materials. They stated that "knowledge of and skill in enhancement techniques as well as artistic ability are critical and of paramount importance if PhotoFit is to be employed to full potential and with any real value or advantage" (pg. 99). Gibling and Bennett explained that the limitations of the composite systems found by many researchers and also in a practical setting, could be overcome with proper training. The superior composites in their study were achieved by operators under training at a UK PhotoFit training course. These courses include training in artistic principles and techniques, the cognitive interview for gaining maximum accurate facial recall and an understanding of facial anatomy. Where these findings lack application to a practical forensic setting, is in the lack of a witness. Operators here produced composites with a target in view and therefore did not include translation of the memory from witness to facial composite system operator.

Verbal translation of the facial memory is an additional problem described in composite research. Although a witness may have a reasonable memory representation of a target face, it is often difficult to describe this to another individual (Christie & Ellis, 1981) and the resulting composite is affected due to inadequate communication of facial features (Wogalter & Marwitz, 1991). As Davies, Milne and Shepherd (1983) describe, the operator has an important role to play in eliciting accurate and detailed information from a witness participant, relating this to the available features of the composite system and producing an accurate composite representation, which is enhanced to a good extent. Again, it may be argued that the best composite is achieved where the operator can elicit the type of detailed information about the facial features which allow him to produce a composite which is greatly enhanced; features are edited to reflect the specific appearance of the targets features, and shading and blending etc. are used to create the contour of the face.

Artistic enhancement of the composite is therefore one factor with which training appears to aid resulting composite quality. Some factors affecting composite quality are more difficult to control due to the general nature of policing. The Association of Chief Police Officers National Working Party (2003) offers guidelines and recommendations on facial imaging practices in relation to witnesses providing evidence in the form of facial composites. The current guidelines state that: "whenever possible, a witness should be contacted by a recognised Police Artist or Facial Imaging Operator within 24- 36 hours of the incident to assess the level of recall" (pg. 12). Research has shown that recall is often most accurate within a short period from the time of the incident and that recall of any information is superior after as short a retention interval as possible (Ellis, Shepherd & Davies, 1980). The nature of police investigations is such that contact with a witness within the recommended time-frame is not always realistic and in some cases the witness may not be contacted for up to a week or more after the incident. An early report into the use of PhotoFit (King, 1971) revealed that the average time between the offence and interview was four days, with a range of six hours to two weeks or more. On average, present delay from incident to police interview is believed to be around two days (Frowd et al. 2005 b), although this can be much longer.

Delay interval from the time of exposure to the target image to composite construction is therefore a further area of interest, in relation to composite quality. Early experimentation on facial composites using the PhotoFit system by Davies, Ellis and Shepherd (1978) illustrated that there was no effect of time delay from exposure to the target to construction on the resulting quality of composites in their study, even if composites were constructed with the target in view. This lack of difference in composite quality was attributed to an insensitivity of the system which was caused by a limited selection of features in the kit. Similar results were obtained by Laughery and Fowler (1980) who also found no decline in composite quality for IdentiKits from memory or with the target in view. Conversely, Koehn and Fisher (1997) found that Mac-a-Mug Pro composites constructed from memory after a delay of two days were of low quality and also of no value in selecting a target from a group of similar looking people.

Green and Geiselman (1989) put forward the hypothesis that an effect of time delay should be masked only for faces with salient facial features. Therefore, the limitation of composite production systems to portray exaggerated facial features should have no effect on performance if the target doesn’t have such features. Their participants constructed IdentiKit composites either immediately or with a one week delay. These composites were then independently rated for likeness. Findings were that performance was superior for composites constructed immediately compared to a one week delay but only with composites that achieved higher likeness ratings and had non-salient faces. It was concluded that IdentiKit composites can be extremely useful if the following ideal criteria are met: a short delay interval from target viewing to construction, the use of a witness who is capable of producing a good composite and the suspect having non-salient facial features.

In a recent study, Frowd, Hancock and Carson (2004, experiment 3) attempted to recreate a more realistic forensic setting than had previously been attempted, or one where similar procedures were adopted to those used in police investigations. Adopting the EvoFIT (Principle Components Analysis shape and face texture breeding) system alongside E-FIT and PROfit, they implemented a two day delay interval and target faces were young male celebrities with whom witness participants were unfamiliar. The composite naming rate was extremely low, with between one and four percent correct identifications overall, by system. A later experiment, compared EvoFIT and PROfit under similar conditions (but with a single operator) and mean naming levels were again low; 8.5% for EvoFIT and 3.7% for PROfit. It is well known that human ability to remember or to match unfamiliar faces is poor (e.g. Hancock, Bruce & Burton, 2000) and this in combination with the two day delay interval may explain these poor naming results.

Later, Frowd et al. (2005a) compared composites from five composite production techniques: E-FIT, PROfit, Sketch, PhotoFit and EvoFIT. Again, their study implemented a “forensically friendly format”. This ‘forensically friendly’ model sought to match the practical procedures used in a police interview, by including the following: composites constructed from memory of an unfamiliar face, a three to four hour delay from viewing of the target to composite construction, experienced artists or operators, famous faces as targets, one composite constructed of one target, a cognitive interview with no limit on construction time and artistic elaboration allowed. Naming and sorting results revealed that PROfit and E-FIT were equivalent in performance and were superior to the other techniques. Overall, naming rates were 18%. Target distinctiveness was found to be an important factor, with those targets with highly distinctive facial features receiving greater levels of naming. The three to four hour delay from viewing of the target to construction of a composite was a limitation of this study; it is highly unlikely that in a police investigation a witness would undergo the cognitive interview procedure within three to four hours of the incident occurring.

Following on from this study, Frowd et al. (2005b) compared composites constructed with a two day delay interval. Given the limitation in retention interval of three to four hours in their previous study, Frowd et al. implemented this more realistic retention interval. It was expected that less facial information would be recalled than with previous delay intervals and therefore deterioration in composite quality and overall naming rates would result. Naming rates were very poor for PROfit and E-FIT (only two correct identifications from over three hundred attempts) and although better, only 8% for the Sketch Artist. These poor results may be explained in terms of the two day delay interval from target exposure to composite construction. Frowd et al. explain the poor naming performance after the two day delay interval as a decrease in verbal descriptions and a shift from feature-based encoding to holistic encoding. They explain this in terms of feature-based encoding being best for feature-based construction (as E-FIT and PROfit), whereas holistic encoding is most appropriate for face recognition and faces constructed by a sketch artist. A two day delay may result in a reduced witness memory of the target image and a shift from feature-based to holistic construction, therefore facilitating increased performance for composites from a sketch artist. This seems a sensible suggestion given that after a retention interval of two days the memory trace may be decaying or displacement may have taken place and therefore little detail can be retrieved about the individual features of the face and it is more likely the mental image will have moved towards an impression of the whole face.

The current study sought to expand the findings of earlier studies by Frowd et al. (2004; 2005a, b) whereby an ecologically valid approach to composite construction and a realistic retention interval were implemented. Here, two delay intervals were employed: 2 days and 1 week. Although difficult to implement in a laboratory setting, it was felt that these delays were more realistic and forensically valid than the three to four hour delay imposed by Frowd et al. A further concern with the Frowd et al. (2005 a, b) studies was the use of different target sets across the two delay intervals. To minimise the effects of a different target set across retention intervals, the same targets were constructed in both delay intervals and by the same operator. Target images were selected to be highly recognisable (famous people: sports personalities, politicians, film stars etc.), in general, although participants were pre-screened so that they were unfamiliar with the targets in an attempt to reflect a real-world paradigm.

Retention interval was one area of investigation in this study. This was the manipulation of the length of time participants were given from viewing a target image to recall and subsequent construction of a facial composite of that person. It was expected that there would be a difference in the nature of the resulting composites constructed at the 2 day and the week delay intervals and more specifically, that the composites constructed with the week retention interval would be less well identified than those constructed with the 2 day retention interval. As discussed earlier, research has shown that recall is often most accurate within a short period from the time of the incident and that recall of any information is superior after as short a retention interval as possible (Ellis, Shepherd & Davies, 1980).

A further variable was artistic enhancement; the extent to which artistic editing tools such as feature blending could be used to elaborate the composite in order that it represented the target image to an optimal extent. This variable was manipulated through the use of three stages of artistic enhancement, whereby the composite was saved at three stages of construction: with no artistic elaboration; with some elaboration where features were edited to resemble the specific facial features of the target and with the use of editing tools to blend the individual features to the background face to give a more natural appearance. The composites were saved at these stages in order to determine whether it was advantageous to the quality of the resulting composite to use the editing features within modern computerised composite systems. Specifically, whether the use of editing tools would mean that the overall appearance of the composite would better resemble the face being depicted, for example, by providing a more natural, blended appearance of the individual features to the background face (i.e. the feature blended stage or the resulting composite), than earlier composite systems that were criticised for their lack of human appearance. Additionally, whether editing of individual features would provide a greater overall likeness than the use of unedited features as they appear in the PROfit database (i.e. elaboration versus no elaboration). Previous research in this area had concentrated on the effects of artistic enhancement whilst composites were in view (Gibling & Bennett, 1994) or with immediate construction (Davies, Milne & Shepherd, 1983); this study was the first to concentrate on the effects of artistic elaboration from the witness's memory.

It was expected that composites saved at the feature blended stage would be identified to a greater extent than those composites saved at the elaboration stage as they would have a more human appearance. Also, that the composites saved at the elaboration stage would produce more correct identifications than those produced at the no elaboration stage. Past research has criticised composite systems for their lack of ability to produce facial features that would better resemble the specific features being depicted (lack of features in the database) and for the inclusion of lines or obvious boundaries on the face (these would be present to a certain extent without the use of tools to blend the features to the background face).

The main area of investigation in this study was retention interval from viewing of a target image to construction of a composite. Two delay intervals were imposed (2 days and 1 week) which were more forensically valid than those from previous research with similar procedures (Frowd et al. 2005a, b). Artistic enhancement was a further area of investigation; composites were saved at three stages, with differing levels of editing.

Naming and likeness ratings were obtained to assess composite quality. Following Experiment 2, a further experiment was undertaken using shorter retention intervals of 3 – 4 hours and 2 days.

Experiment 2: 2 day and 1 week retention intervals

Method

Composites were constructed of six famous faces that were unfamiliar to those constructing; they were later tested for identification through a naming procedure, by a second sample of volunteers. Target images were constructed over two delay conditions: after 2 days and after 1 week, by two different samples of witness participants. The operator and the facial composite system employed (PROfit) remained the same in both delay conditions to minimise operator and system effects. Composites were saved at three stages to study the effects of artistic enhancement. These stages aimed to assess whether editing of the initial features selected (elaboration condition) and using an editing tool to blend the facial features to the background face (feature blended condition) allowed the composites to be named correctly on more occasions than with no editing or feature blending at the initial selection stage (no elaboration condition). Likeness ratings of the extent to which the composite represented the target image were also collected. These were used as a supplementary measure to naming results, which were expected to be at low levels based on previous research.

Composite construction

Operator

The same operator was involved in the production of composites from both delay conditions in order that any significant differences in naming over the two delay intervals were not confounded by operator differences (level of experience and skill etc.). The operator had previous experience in composite production and the use of the cognitive interview procedure and had been involved in a research experiment (Experiment 1; chapter 2) prior to the present experiment, involving the construction of composites of ten famous males. The operator remained *blind* to the identity of the target images and also to the retention interval that each participant had been given. In this way, operator effects relating to familiarity with the target and the effects of delay were minimised. Target images were chosen by a researcher in the field to allow the operator to remain blind to their identity. Additionally, for viewing of the target images, participants were asked to visit the researcher who then directed them to the operator for the composite construction stage of the procedure. The operator was familiar with the PROfit system and implementing the various artistic features within the paint package.

Facial composite production system

The PROfit computerised facial composite system was used to construct composites in both delay conditions to limit system differences. The databases of features in the PROfit system are compiled from photographs of faces, thereby providing the user with the opportunity to produce composites with faces which have greater resemblance to a target face, when compared with sketch-like systems. The editing tools allow for each individual feature to be modified to the specific requirements of the witness participants. One of the editing tools, 'smudge' allows the user to smooth or blend the features of the face in order that they fit the background face, producing a face which is more natural in appearance and possibly a more identifiable representation.

Targets

Target images were six famous people: Russell Crowe, Robin Cook, Michael Schumacher, Robert Carlyle, Nicholas Cage and Ewan McGregor (actors, film stars, politicians, sportsperson). They were selected on the basis that they were generally well known but were unfamiliar to those constructing them. This was in an attempt to reflect a forensic setting, with increased realism as witnesses generally construct composites of individuals who are unknown. Therefore, it was necessary for targets to be famous but not so famous that they would be identified by all those asked to participate in the construction phase of the experiment.

A sample of 29 volunteers were asked to assess 15 target faces according to familiarity and distinctiveness. The procedure used to rate the distinctiveness of the targets was for individuals to be presented sequentially with good quality images of the targets and asked to imagine meeting each person at a railway station in amongst their peers (young, white males) and rate from 1 to 7 (1 = *average, blend into the crowd* and 7 = *very distinctive, stand out from the crowd*). Familiarity ratings were assessed using a familiarity ratings scale (as described in Chapter 2). Images were selected as targets if they had high mean familiarity ratings (of around 3 or above on the familiarity scale) and were also not known by two or more individuals in the sample who were rating the target images; these people would be asked to participate in the next phase of the experiment. Targets had the following mean familiarity and distinctiveness ratings (Table 3.1 below):

Table 3.1: Mean target familiarity and distinctiveness

Target	**Distinctiveness** **(1 -7)**	**Familiarity** **(1 -5)**
Michael Schumacher	2.9	3.2
Ewan McGregor	3.6	3.9
Robert Carlyle	3.6	3.9
Nicholas Cage	3.7	3.7
Robin Cook	4.0	3.3
Russell Crowe	4.2	3.8

Target images were reproduced as high quality prints, in full colour and were front facing with a neutral expression. Targets were chosen to keep artistic enhancements to a minimum.

Witness participants

Twelve adults were recruited from around Stirling University and were each paid £5 for their participation. They were aged between 23 and 55 and had a mean age of 37.7

years (SD = 11.6) and were 4 males and 8 females. Witness participants were chosen on the basis that they had no experience of composite construction or the cognitive interview technique.

Procedure

Participants were primarily recruited to be involved in the rating of the 6 target images for familiarity and firstly attended an additional researcher to allow the operator to remain blind to target identity and retention interval. Where participants rated any of the target images as rank one on the familiarity scale (i.e. not known at all), they were asked if they would further participate in an experiment of famous faces. In this way, participants were pre-screened so as to be unfamiliar with the target image of whom they would later produce a composite. Two participants were required to be unfamiliar with each target face; one of whom was randomly assigned a retention interval and by default, the remaining participant was given the remaining retention interval. In this way, a target image was constructed by different participants but at two retention intervals. Witness participants were instructed that they would be asked to construct a facial composite and were given one minute to study a target image. They were then asked to return to complete the experiment after a period of either 2 days or 1 week (half of the participants were given a 2 day delay and half a 1 week delay). The retention interval given to each of the witness participants was randomised.

After the prescribed retention interval of either 2 days or 1 week, witness participants returned for the next stage of the experiment and constructed a composite from memory.

The procedure remained largely the same as for Experiment 1 apart from those aspects relating to artistic enhancement; the composite was saved at three stages of construction. Firstly, the witness participant was given the opportunity to work on the composite to improve it on a feature-by-feature basis, beginning with whichever feature he or she wished to modify in the first instance. When the participant was satisfied that he or she had chosen the facial features that most closely matched those of the target face, the composite was saved; stage one (no elaboration condition). Next, any amendments were made to the facial features themselves, so for example, where the hairstyle of the composite was closely matching on initial selection of the features but was not matching the target's hair to the participant's requirements, the operator used the editing tools within the paint package to make the change. Again the composite was saved; stage two (elaboration condition). Finally, when the witness participant had made all modifications to the facial features and the resulting composite was a good resemblance to the target face, the facial features were blended to 'fit' the background face i.e. the face was given a more natural appearance. This was achieved by using the 'smudge' tool within the PROfit paint package. The composite was again saved; stage three (feature blended condition).

The operator attempted to follow the procedure as closely as possible and save the composite at the prescribed three stages. The procedure described above was not always possible to follow; participants requested to modify their composites at various times throughout the construction phase which did not necessarily follow the three stages described above. On a few occasions at the elaboration stage, witness participants requested to change one of the features that had been previously saved as part of the no elaboration composite. In police practice, in the cognitive interview, the witness is able to make modifications at any time and can construct the composite in any order. The procedure adopted here aimed to be as realistic as possible and therefore the operator saved the composite in three stages which were as closely associated with the no elaboration, elaboration or feature blending conditions as was possible. Following completion of the composite, participants were asked to make a judgement as to whether they preferred the composite saved at the elaboration stage or at the feature blended stage.

Composite evaluation

Design

There were 6 composites in each of the delay conditions and these were each saved at 3 levels of artistic enhancement and were tested for identification by a naming procedure. The composites were divided into 6 sets and counterbalanced so that half of the composites in a booklet were constructed after a 2 day delay and half after a week delay, with one of each of the no elaboration, elaboration and feature blended composites for each of the delay conditions. Each participant was shown one of the sets.

Composite naming was the primary measure of composite identification; this has previously been shown to be the most ecologically valid measure of composite quality. Likeness ratings were used as a supplementary measure to the low levels of naming expected, following results from previous research implementing similar procedures.

Evaluator participants

Naming of the composites was undertaken by 72 volunteers who were either students of Stirling University or were recruited from around the general area of the University. Demographics relating to age and gender are not available due to oversight on behalf of the author.

Procedure

Participants were informed they would be shown six composites of famous people and if possible, they should try to provide the identity of each of the composites, even if they were unsure as to correctness. Composites were shown for as long as the participant required and the response recorded. Where participants were unable to provide a name for the identity portrayed in the composite, a description of the person's occupation or any other details regarding the specific identity of that person (which would mean he could be identified as a particular individual) were accepted by the operator.

Following the composite naming procedure, participants were shown the target images alongside the composites and asked to rate how alike they considered the composites to the targets. A scale of 1 to 10 was employed, where 1 was not at all alike and 10 was extremely similar.

Finally, information relating to the nature of the experiment was provided to participants, any questions they had were answered and they were thanked for their participation.

Results

Composite naming: retention interval and artistic enhancement conditions

Composite naming was very poor overall with only 9 correct identifications out of a possible 216 in the 2 day delay condition and only 12 correct identifications out of a possible 216 in the week delay condition. The composite with the highest naming score in the 2 day delay condition was Robin Cook, who was correctly named 5 times (2.3%). This composite was constructed in the no elaboration condition. The composite with the highest naming score in the week delay condition was Nicholas Cage who was named correctly on 8 occasions (3.7%). This composite was constructed in the feature blended condition and was the only composite that received any correct naming in the feature blended condition with the week delay. Figure 3.1 shows composites constructed of Nicholas Cage as saved at the no elaboration stage, the elaboration stage and the feature blended stage respectively.

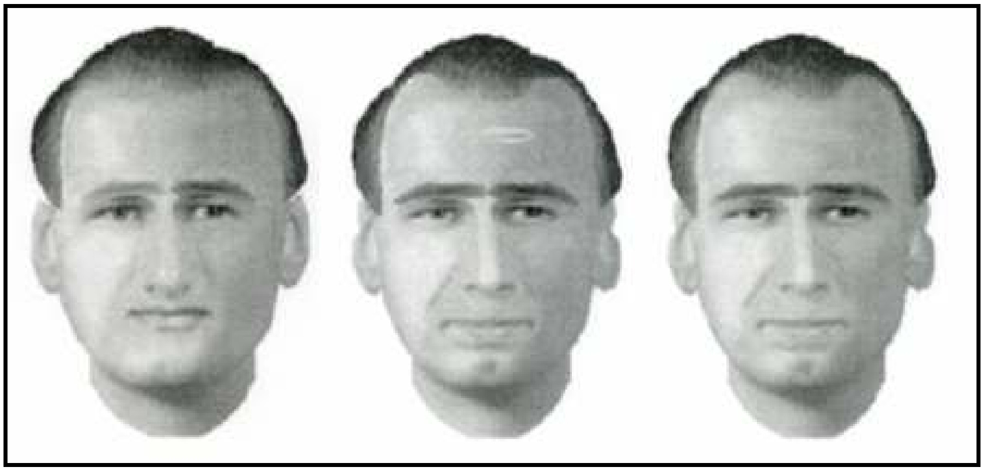

Figure 3.1: Composites constructed of Nicholas Cage (from left to right: no elaboration, elaboration and feature blended composites)

Mean participant naming rates for the 2 day delay condition were 4.2% (SD = 0.2) and for the week delay condition were 5.6% (SD = 0.2). An effect size of -7.00 was calculated for the two day and week conditions. These are extremely low levels of naming when compared to mean naming rates from other studies of around 20% (Brace, Pike & Kemp, 2000; Davies et al., 2000; Bruce et al. 2002; Frowd et al. 2005a), although these studies did not have retention intervals of the length used here. The naming rates do not take into account target naming. However, targets were selected to be highly familiar, based on familiarity ratings, prior to the experiment. It was therefore expected that targets would be named correctly to a good extent.

Performance was extremely low overall. The naming results should be interpreted with caution, given the ‘floor effects’ that would result in terms of low values from the naming scores. Figure 3.2 illustrates the number of correct recognitions by participant, for the 2 day and 1 week delay conditions by artistic enhancement. There is greater naming in the ‘no elaboration’ condition with the 2 day delay and greater naming in the ‘feature blended’ and ‘elaboration’ conditions for the 1 week delay.

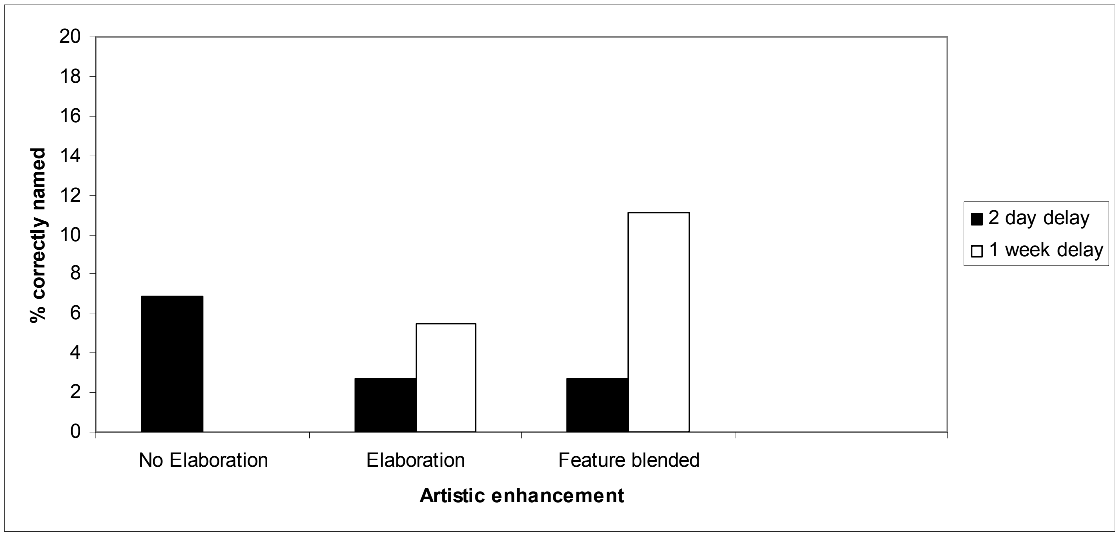

Figure 3.2: Percentage of correctly named composites by delay and artistic enhancement

A two way repeated-measures mixed ANOVA by subjects was undertaken using naming rates. The two factors were: delay (between subjects factor, with 2 levels: 2 day and 1 week delay) and artistic enhancement (within subjects factor, with 3 levels: no elaboration, elaboration and feature blended). The main effects of artistic enhancement ($F_{2, 142} = 1.107$, $p = 0.333$) and delay ($F_{1, 71} = 0.357$, $p = 0.552$) were not significant. The artistic enhancement by delay interaction was highly significant ($F_{2, 142} = 5.179$, $p = 0.007$). Post hoc analysis of the interaction between delay and artistic enhancement using the Bonferroni test revealed that those composites produced in the week delay condition, at the feature blended stage had significantly higher naming than those produced at the no elaboration stage ($p = 0.012$). Also, for composites produced at the no elaboration stage, naming was significantly higher for the 2 day delay compared to the week delay (p = 0.024).

A further two way repeated-measures, mixed ANOVA by items was undertaken using naming rates. Again, the two factors of the ANOVA were: delay and artistic enhancement. The main effects of artistic enhancement ($F_{2, 284} = 1.137$, p = 0.322) and delay were not significant ($F_{1, 142} = 0.406$, p = 0.525) as for the subjects analysis. Simple main effects analysis of the artistic enhancement by delay interaction revealed similar results when comparing by subjects and by items. The artistic enhancement by delay interaction was highly significant ($F_{2, 284} = 5.037$, p = 0.007). Post hoc analysis of the results using the Bonferroni test revealed that in the week delay condition, naming was significantly higher in the feature blended stage than the no elaboration stage (p = 0.009). Also, for composites produced in the no elaboration stage, naming was significantly higher for the 2 day delay compared to the week delay condition (p = 0.023). In addition, the feature blended composites obtained significantly higher naming for the week delay than the 2 day delay conditions (p = 0.050).

Likeness ratings: retention interval and artistic enhancement conditions

As for the naming scores, likeness ratings were also relatively low overall, with a highest mean likeness rating of 4 for composites saved at the no elaboration stage with a 2 day retention interval. The composite with the highest mean likeness rating was Nicholas Cage (M = 4.7, SD = 2.27) and the composite with the lowest mean likeness rating was Russell Crowe (M = 1.36, SD = 0.54). The mean overall likeness rating for the 2 day delay condition were (M = 3.7, SD = 1.97) and for the 1 week delay condition was (M = 3.3, SD = 2.2). An effect size of 0.19 was calculated for the two day and one week delay conditions. Figure 3.3 illustrates mean likeness ratings by delay and artistic enhancement.

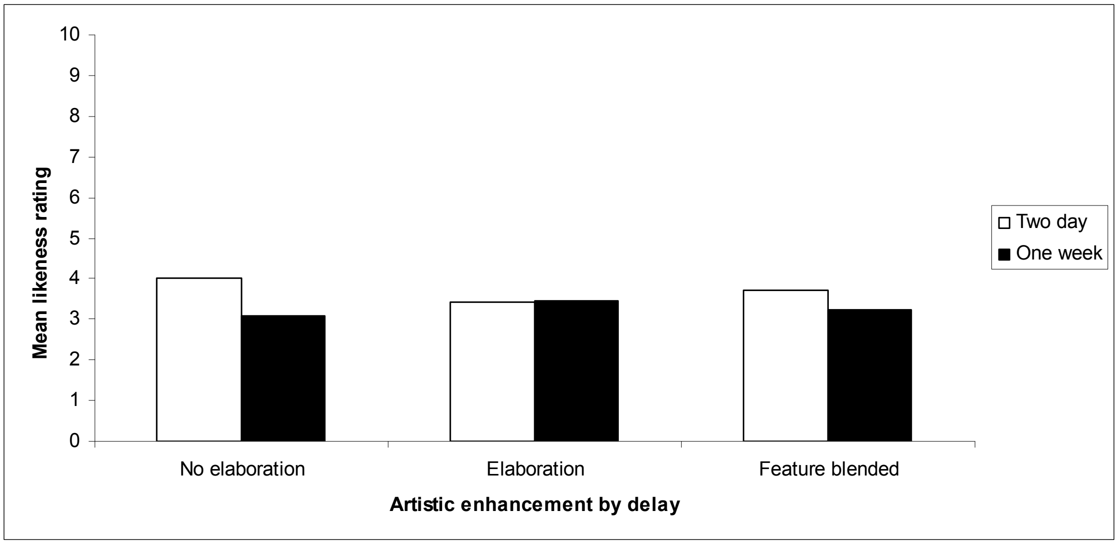

Figure 3.3: Mean participant likeness ratings for the three stages of artistic enhancement for the 2 day and 1 week delay conditions

An ANOVA on likeness ratings was undertaken. This was a two-way, repeated measures, within subjects ANOVA, by subjects. The two factors were artistic enhancement and delay. As for the naming scores, the main effect of artistic enhancement was not significant ($F_{2, 142} = 0.226$, $p = 0.798$). However, unlike naming, the main effect of delay was highly significant ($F_{1, 71} = 7.837$, $p = 0.007$), with composites constructed with a 2 day delay (M = 3.7) receiving higher likeness ratings than those constructed with the week delay (M = 3.3). The artistic enhancement by delay interaction was not significant ($F_{2, 142} = 1.278$, $p = 0.282$).

A further ANOVA was undertaken with evaluators' likeness ratings. This was a two way, repeated measures, mixed ANOVA, by items. Again the two factors were artistic enhancement and delay. Again, the main effect of artistic enhancement was not significant ($F_{2, 284} = 0.103$, $p = 0.902$). The main effect of delay was again significant ($F_{1, 142} = 4.618$, $p = 0.033$), with those composites constructed at the 2 day stage obtaining higher mean likeness ratings (3.7) than those constructed at the week delay (3.3). The artistic enhancement by delay interaction was not significant ($F_{2, 284} = 1.972$, $p = 0.141$).

Witness participants' judgements

Witness participants were asked to judge whether they preferred the composite saved at the feature blended stage or those saved at the elaboration stage. Nine out of twelve participants expressed a preference for composites saved at the feature blended stage. There was a trend towards significance ($X^2 = 3.000$, df = 1, $p = 0.083$).

Discussion

The aim of this experiment was to assess the effects of delay and artistic enhancement upon composite quality. Witness participants were unfamiliar with target faces in order to implement a real-world paradigm. One operator constructed composites with both a 2 day and a 1 week retention interval from target exposure to composite production. In order to assess the effects of artistic enhancement within the composite construction phase of the experiment, composites were saved at three pre-defined stages: no elaboration, elaboration and feature blended. The stage of ‘no elaboration’ was defined as the initial selection of features by the operator based on the constructor participant’s verbal description of the facial features of the target and also the final selection of features by the witness participant. In the next stage, participants worked with the operator to modify the individual facial features to reflect the specific appearance of the target’s facial features. The final stage of ‘feature blending’ involved using the ‘smudge’ tool to blend the facial features to fit the background face, in an attempt to provide a more natural appearance. Composite quality was assessed by naming and likeness ratings from evaluator participants.

In spite of good target familiarity, naming rates were extremely poor when compared to the 20% naming rate found by previous researchers using the PROfit and E-FIT systems (Brace, Pike & Kemp, 2000; Davies, van der Willik & Morrison, 2000; Frowd et al. 2005a and others). However, a recent study by Frowd et al. (2005b) also revealed low naming rates, of less than 2%, with a two day retention interval. Mean participant naming rates for the 2 day delay condition were 4.2%, and 5.6% for the week delay condition. Two composites accounted for all but one of the instances of naming: Nicholas Cage and Robin Cook. Nicholas Cage was correctly named on eight occasions in the feature blended stage of the 1 week delay condition. He was also named on four occasions in the elaboration stage of the week delay condition but there were no instances of naming in the no elaboration stage. In contrast to this, the composite of Robin Cook was named correctly on five occasions in the no elaboration stage and only once in both the elaboration and feature blended stages, with the 2 day delay interval. In spite of significant differences, the naming data is largely dependant on a few items and as such more weight will be given to results from the likeness ratings, in this discussion.

The main effects of artistic enhancement and retention interval were not significant, when considering the naming data. However, these variables interacted such that for composites constructed with a delay of 1 week, naming was significantly higher for the feature blended composites than the no elaboration composites, and there was significantly greater naming with 2 days retention interval for the composites with no elaboration. Further, composites constructed at the feature blended stage produced significantly higher identification levels at the week delay rather than the 2 day retention interval. Witness participants' views appear to be at odds with the results obtained on the effects of artistic enhancement. Overall, nine out of twelve witness participants expressed a preference for the composites saved at the feature blended stage of the procedure, compared to those in the elaboration stage; there was a trend towards a significant difference.

Evaluator participants' likeness ratings of the composites were also relatively low; mean likeness rating for the 2 day delay was 3.7 and 3.3 for the week delay. As for the naming results, Nicholas Cage also had the highest mean likeness ratings of 4.7. The composite of Russell Crowe obtained the lowest mean likeness ratings of 1.4. Analysis of the likeness ratings revealed similar results to the naming results in terms of artistic enhancement which was not found to have a significant effect. However, unlike for the naming results, delay was shown to produce significant effects, with those composites produced with the 2 day retention interval receiving higher likeness ratings than the week retention interval composites. This result is in line with the prediction that there would be a difference in the nature of the composites produced at the two retention intervals. Further, although those naming the composites did not provide significant results in terms of more correct recognitions with a shorter retention interval, likeness ratings point to superior quality composites with a shorter retention interval. Perhaps, a more appropriate measure to adopt here would have been to do an alternate forced choice task between the composites produced at the two retention intervals, as a direct comparison.

The interaction between the effects of artistic enhancement and delay produced inconsistent results for the naming scores and the likeness ratings. However, it is difficult to assess the validity of the naming data given that almost all of the instances of correct naming were obtained on the basis of two *good* composites; Nicholas Cage and Robin Cook. Only one instance of naming occurred which did not relate to either of these composites. An assessment of naming results revealed significantly higher scores for composites saved at the feature blended stage than the no elaboration stage for the week retention interval and also that composites in the feature blended stage were better recognised after 1 week retention than 2 days. These results appear to contrast with the results of previous researchers (Davies & Christie, 1982), who suggested that following a longer delay interval, less detail is recalled but enough is recalled to construct a basic resemblance of the important elements of the face. They suggested that a perceptually poorer facial likeness would result when a face with incorrectly recalled feature information was elaborated. Therefore, one would expect that composites produced at the no elaboration stage would out-perform those produced at the elaboration and feature blended stages with a week retention interval, as less factually correct information would result but this would not be elaborated upon. Perhaps following a week retention interval, the introduction of the feature blended elements of the face meant that the composite was altogether more aesthetically pleasing with a more human appearance and this outweighed the effects of the lack of correctly recalled facial elements at the no elaboration stage. In contrast, with the shorter retention interval, there was no advantage for elaboration as the information recalled at the no elaboration stage was more accurate.

The results of elaboration obtained here do go some way towards a conclusion towards enhanced composites being of superior quality to unenhanced composites as in previous research (Laughery & Fowler, 1980: Davies, Milne & Shepherd, 1983: Gibling & Bennett, 1994). The problem here is that the results on enhancement are based largely around the identification of two composites. The present study is the first study to directly assess the effects of artistic enhancement with witness participants, from memory. Davies, Milne and Shepherd (1983) assessed the effects of artistic enhancement through the use of a novice or an experienced operator, with immediate construction and Gibling and Bennett (1994) assessed enhanced and unenhanced PhotoFits constructed whilst in view.

The poor results obtained here in relation to artistic enhancement (based largely around one or two good composites) may be explained by the very poor naming rates overall. Perhaps the delay intervals of 2 days or 1 week were too long for artistic enhancement to produce beneficial results; there may have been an insufficient level of facial detail recalled from which to elaborate the composite. An informal assessment of witness participants' verbal descriptions for the 2 day and week delay intervals reveals that there does appear to be some decline in the amount of facial feature detail recalled after 1 week. There were five instances (across the six witness participants) where no information was provided for one of the facial features, in the week delay condition. This did not occur once in the 2 day delay condition. In general, witness participants' descriptions lacked some content in the week delay condition when compared to those in the 2 day delay condition. Although based around only one or two composites, the results here in relation to artistic enhancement do suggest there may be an advantage in the use of enhancement techniques following delays of two days or more.

Experiment 3: 3 - 4 hour and 2 day retention intervals

Introduction

In the previous experiment, a significant difference between the 2 day and the week retention intervals was found for participants' likeness ratings, although naming rates were extremely low (especially when compared to naming rates from previous researchers). The difference in likeness ratings for the two retention intervals was in line with the prediction that more facial feature information would be recalled with the shorter retention interval and therefore, a composite with more recognisable qualities would result. Following this, it was considered that a much shorter retention interval of only a few hours would produce a significant difference in composite quality when compared to a delay interval of 2 days. Although past research by Frowd et al. (2005a) had recognised the 3 - 4 hour retention interval as being a limitation of their study due to its lack of forensic validity, a further attempt at implementing this delay was considered sensible when employing the same procedures used for Experiment 2.

A follow-up experiment was undertaken which again sought to include an ecologically valid approach to composite construction and included a different target set and retention intervals of 3 - 4 hours and 2 days. Previous research by Frowd et al. (2005a, b) also implemented these delay intervals but using different target sets across the two conditions, which was a limitation of the research. To minimise the effects of a different target set across retention intervals, the same targets were constructed in both delay intervals. Target images were selected to be highly recognisable (famous footballers from international and Premier League teams), although participants were pre-screened so that they were unfamiliar with the targets in an attempt to reflect a real-world paradigm.

Given that the effects of artistic enhancement were based largely around only one or two composites and poor recall after considerable retention intervals was suggested as a possible reason for this, the effects of artistic enhancement were not studied further. As would occur in a facial imaging interview, the composite was saved when the participant was satisfied that he or she had produced the best likeness of the target image and had no further modifications to make. Feature blending of composites directly before completion was used as witness participants preferred the use of this technique in the previous experiment.

Mean likeness and naming rates from the 2 day delay composites in this experiment may be informally compared with the 2 day delay composites from Experiment 2, to determine any similarities or differences. Statistical analysis is possible but would be confounded by the use of a different target set.

The main area of investigation in this experiment was retention interval from viewing of a target image to construction of a composite. The two delay intervals imposed had been used in the previous experiment here and by Frowd et al. (2005a, b). It was expected that composites constructed with the 2 day retention interval would be less well identified than those constructed with the 3 - 4 hour retention interval. Naming and likeness ratings were obtained to assess composite quality.

Method

Composites were constructed of 12 famous faces that were unfamiliar to those constructing them; they were later tested for identification through a naming and a likeness rating procedure, by a second sample of volunteers. Target images were constructed over two delay conditions; after 3 - 4 hours or 2 days, by two different samples of participants. As for the previous experiment, the operator and facial composite system employed remained the same in both delay conditions to minimise operator and system effects. Composites were saved at only one stage as the effects of artistic enhancement were of no further interest following the results of the previous experiment.

Composite construction

Targets

Target images were of 12 famous footballers. Targets were selected to be highly familiar in general but were unfamiliar to those constructing them; they played for teams in the Premier League in England or were famous on an international scale. As per Experiments 1 and 2, information relating to the distinctiveness of the target images was collected, in the form of individual subject ratings (see Table 3.2 below). Again, the images were front facing, with neutral expression and were printed to a high quality. As per the previous experiments, targets were chosen to minimise the extent to which artistic enhancement of the features was required. A different target set to that used in Experiment 2 was employed so that the operator was again blind to the identities of the targets.

Table 3.2: Mean distinctiveness rating for targets

Target	Mean distinctiveness rating
Teddy Sheringham	3.3
Paul Scholes	3.5
Wayne Rooney	3.6
Roy Keane	3.7
Ole Gunner Solskjaer	3.9
Emmanuel Petit	3.9
Tony Adams	4
Zenadine Zidane	4.4
Steve McManaman	4.7
Alan Smith	4.7
Peter Beardsley	4.9
Dennis Bergkamp	5.6

Witness Participants

Participants were recruited from Stirling University, staff and students, and each paid £5 for participation in the experiment. The participants had a mean age of 31.7 years (SD = 10.8) and were 4 males and 20 females.

As in Experiments 1 and 2, participants were chosen on the basis that they fitted the selected age requirements of between 18 and around 55 and had no previous experience in composite construction with the use of facial composite software. Individuals with detailed knowledge of the cognitive interview were not selected as participants. All participants were debriefed as to the purposes of the experiment following completion of the composite.

Procedure

The procedure employed here remained largely the same as in Experiment 2 in order to maintain a forensically valid technique (targets drawn from a large population, unfamiliar witness participants for construction of the composites and familiar evaluators, realistic retention interval etc.). Again, in order that the operator remained blind to retention interval and target identity, an additional researcher selected the targets, allowed the witness to view these and directed the witness to the operator following the retention interval.

Participants were recruited to be involved in an experiment on famous faces and instructed that they would be asked to construct a facial composite. They had been previously involved in the rating of the twelve target images for familiarity. Where participants rated any of the target images as rank 1 on the familiarity scale (i.e. not known at all), they were asked if they would further participate in an experiment of famous faces. In this way, participants were pre-screened so as to be unfamiliar with the target whose image they would later be asked to commit to memory, recall and produce a composite of. Firstly, participants studied the target image with whom they were unfamiliar for one minute after being instructed that they would be required to produce a facial composite of that person at a later time. They were then asked to return to complete the experiment after a period of either 3 - 4 hours or 2 days (half of the participants were given a 3 - 4 hour delay and half a 2 day delay). The retention interval given to each of the witness participants was randomised.

The same procedure was adopted as for Experiment 2 but this time with 12 targets in each condition instead of 6, and a 3 - 4 hour retention interval rather than 1 week.

The operator constructed the same targets with both a 3 – 4 hour and a 2 day delay interval, but remained unaware of the identity of the targets and the delay interval imposed throughout the composite construction phase of the experiment.

Composite evaluation

Design

The procedure used here was implemented to determine whether the retention interval imposed on witness participants (of either 3 - 4 hours or 2 days) would have an effect on subsequent composite naming by a further sample. Previous studies have shown composite naming by a third person who may or may not know the target person, to be an ecologically valid methodology and a sensitive performance measure. The primary measure of composite quality was therefore composite naming. Likeness ratings were also used to assess composite quality and were obtained from a separate sample of participants.

<u>Evaluator participants</u>

Naming data

Naming of the composites was undertaken by 24 volunteers who were either students of Stirling University or were recruited from around the general area of the University. They were all males and were chosen on the basis that they were football fans (no female football fans were available as an opportunistic sample). The mean age of the males naming set one was 28.3 years (SD = 10.3) and for set two was 30.5 years (SD = 12.3).

There were 12 composites in each of the delay conditions and these were tested for identification through a naming procedure. The composites were divided into two sets and counterbalanced so that half of the composites in a booklet were constructed after a 3 - 4 hour delay and half after a 2 day delay. Each participant was asked to name composites in only one of the sets.

Participants were informed they would be shown 12 composites of famous people and if possible, they should try to provide the identity of each of the composites. It was implicit that the composites were of famous footballers given that participants had been asked whether they were football fans when asked to participate in the experiment. Participants were encouraged to provide an answer, if possible, for each of the composites, even if they were unsure as to correctness. Composites were shown for as long as the subject required and the response recorded. Where participants were unable to provide a name for the identity portrayed in the composite, a description of the person's occupation or any other details regarding the specific identity of that person (which would mean he could be identified as a particular individual) were accepted by the operator. Naming of the targets was undertaken following naming of all twelve composites. Again, as in Experiment 1, this naming of the target image ensured that where participants could not name the composite image, this was not simply because they did not know that target person. Only evaluators who were able to correctly name six or more target images were included in the experiment.

Likeness ratings

Likeness ratings were collected from a further sample of participants. Participants were shown only one of the sets of composites and asked to rate these. Set one participants were 5 males and 1 female with a mean age of 27.2 years (SD = 8.6). Set two participants were 4 females and 2 males with a mean age of 29.5 years (SD = 9.5). Following rating, participants were provided with target names for each of the targets.

Finally, information relating to the nature of the experiment was provided to participants, any questions they had were answered and they were thanked for their participation.

Results

Composite naming and retention interval

Composite naming data was collated for both the 3 - 4 hour and 2 day delay conditions. Composite naming was very poor overall with only 11 correct identifications out of a possible 144 in the 3 - 4 hour delay condition and only 5 correct identifications out of a possible 144 in the 2 day delay condition. Composites correctly named on the highest number of occasions in the 3 - 4 hour delay condition were Roy Keane and Ole Gunner Solskjaer; both named correctly on 3 occasions (see Figure 3.4 below). The composite with the highest naming in the 2 day delay condition was Emmanuel Petit, who was correctly named on 2 occasions.

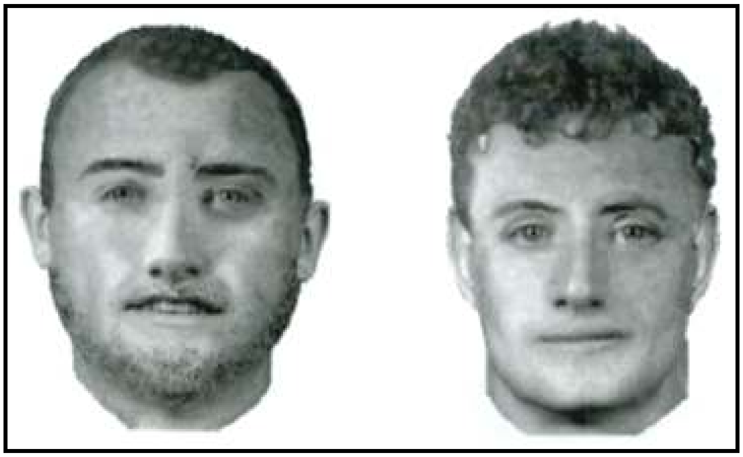

Figure 3.4: Composites of Roy Keane and Ole Gunner Solskjaer (left to right)

Mean subject naming rates for the 3 - 4 hour delay condition were 7.6% (SD = 0.3) and for the 2 day delay condition were 3.5% (SD = 0.2). An effect size of 16.0 was calculated for the 3 – 4 hour and 2 day delay conditions. Again, despite high levels of target familiarity (mean target naming of 98%), these are extremely low levels of naming when compared to mean naming rates from other studies of around 20%. However, the results obtained here for the 2 day delay condition are similar to those obtained in Experiment 2 with the same retention interval of 2 days where mean naming was 4.2%.

Figure 3.5 illustrates the number of correct recognitions by subject, for the 2 day and 1 week delay conditions. There is greater naming in the 3 - 4 hour delay condition than the 2 day delay condition.

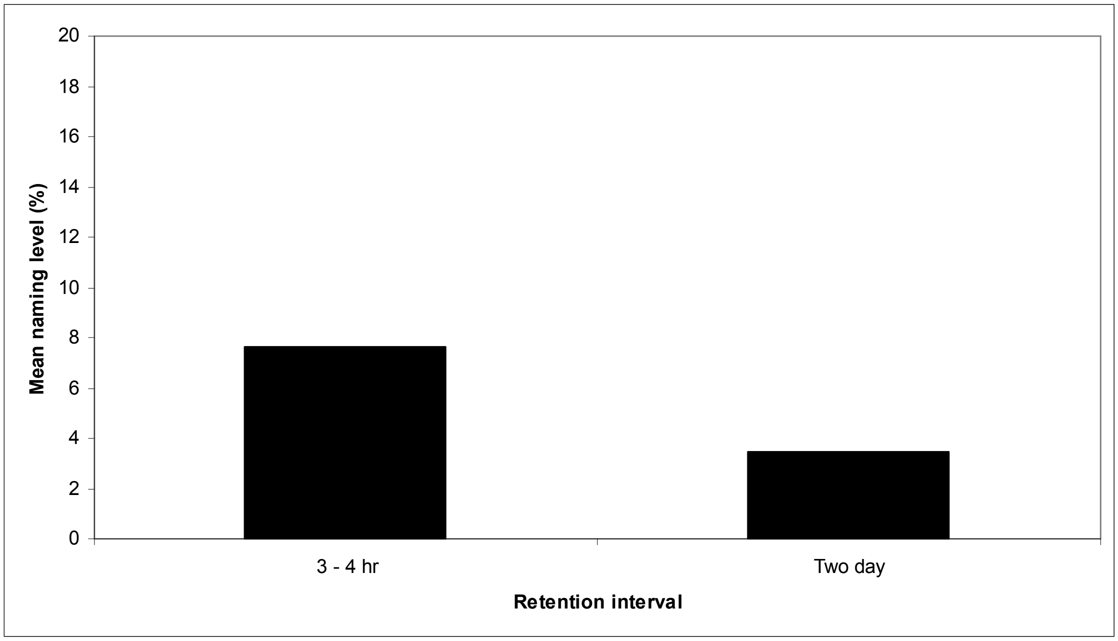

Figure 3.5: Percentage of correctly named composites by retention interval

To determine whether there was a significant difference between composites constructed at 3 – 4 hours and 2 days, a paired t-test by subjects was undertaken and there was a trend in the direction of a significant difference between the two retention intervals (t = -1.466, df = 23, p = 0.081, one tailed). An independent samples t-test by items was also undertaken using participants' naming rates. There was no significant difference between the two retention intervals (t = -1.290, df = 22, p = 0.105, one tailed).

Likeness ratings and retention interval

As for the naming rates, likeness rates were also relatively low with a mean subject likeness rating of 4.7 (SD = 1.1) for the composites produced with the 3 – 4 hour delay. Similarly, 4.8 (SD = 1.1) was the mean likeness rating for composites produced with a 2 day delay. This result was slightly better than that obtained in the previous experiment where mean likeness ratings for composites produced with a 2 day delay of 3.7 resulted. An effect size of -0.09 was calculated for the 3 -4 hour and 2 day delay conditions based on likeness ratings.

A paired t-test by subjects also gave a non-significant result of delay (t = 0.874, df = 11, p = 0.401) on likeness ratings. A further independent samples t-test of delay, by items, was undertaken with participants' likeness ratings. Again, delay did not have a significant effect on likeness ratings (t = -0.347, df = 22, p = 0.732).

Mean naming and likeness: Experiments 2 and 3

Given that the same procedure had been adopted in the construction of the composites produced after 2 days in Experiment 2 and Experiment 3, it was possible to compare the results from both studies, although statistical analysis was not undertaken as the results would be confounded by the use of a different target set across studies. Mean naming rates from both experiments were similarly low in comparison to mean naming of around 20% from researchers in this field: mean naming for Experiment 2 was 4.2% and for Experiment 3 was 3.5%. Again, likeness ratings from both experiments were relatively low although the Experiment 2 composites faired slightly better: mean likeness rating of 3.7 for Experiment 1 and 4.8 for Experiment 2.

Discussion

The aim of this further experiment was to assess the effects of two relatively short retention intervals upon composite quality. Composites were constructed with both a 3 - 4 hour and a 2 day retention interval from target exposure to composite production. The effects of artistic enhancement upon identification were not of interest here; results from Experiment 2 pointed to an advantage in terms of identifiable quality, for composites that were artistically enhanced after a longer retention interval of one week but the results were largely based around two composites. It was considered that studying the effects of artistic enhancement with considerable retention intervals was perhaps not the most sensible approach and therefore further study of the effects of artistic enhancement did not take place here. Composite quality was assessed by naming and likeness ratings from further witness participants.

As for Experiment 2 with retention interval conditions, naming rates were extremely low; for the 3 - 4 hour delay a mean of 7.6% was obtained and 3.5% for the 2 day delay. Analysis of the naming rates did not reveal a significant difference for the primary measure of delay, this result being in agreement with that obtained in the previous experiment. However, there is a trend in the direction of an advantage for the shorter retention interval when a subjects analysis is performed (p = 0.081). Comparison of the results of this experiment with Experiment 2 are problematic although similar procedures were adopted. Differences in the target set used make it difficult to draw firm conclusions across the two experiments. However, it is possible to assess similarities and differences between the two. Naming results from this experiment were similar to those obtained in the previous experiment where mean naming was 4.2% for the 2 day delay interval and 5.5% for the week retention. Composites produced with the 2 day retention interval in Experiment 2 faired slightly better than those in this experiment, in relation to mean naming rates.

As for the previous experiment, naming rates were extremely low when compared to the average naming rates from previous research of around 20%. One might have expected an advantage in the naming scores in this experiment over the previous experiment given that it was implicit to evaluator participants upon asking if they were football fans that they were to be shown images of footballers. Although in the previous experiment evaluator participants were aware that they would be shown images of famous people from several genres, they had a more difficult identification task than the evaluator participants in the present experiment who were football fans, as they were selecting from a wider population.

Likeness ratings were also reasonably poor although very similar mean likeness ratings resulted for the 3 - 4 hour retention interval (4.7) and the 2 day delay (4.8). Likeness ratings for the composites constructed in this experiment were higher than those for the previous experiment; 3.7 for the 2 day retention interval and 3.3 for the week retention interval. In contrast to the analysis of the likeness ratings in the previous experiment, likeness ratings here did not reveal a significant difference across the two retention intervals.

General discussion of experiments 2 and 3

The present experiments have concentrated on the effects of retention interval on recall ability for participants constructing facial composites. Specifically, the delays of 3 - 4 hours, 2 days and 1 week have been implemented. Retention interval from viewing of the target to composite construction resulted in composites receiving significantly higher likeness ratings when constructed after 2 days compared with 1 week. This is further evidence towards a shorter retention interval between incident and facial imaging interview. However, results did not point to an advantage for the even shorter retention interval of 3 - 4 hours over 2 days, as one might expect. There was a trend in this direction from the naming levels.

Although likeness ratings provided significant results, naming results remain very poor overall when compared to the mean naming rates from previous studies of around 20% (e.g. Brace, Pike & Kemp, 2000; Davies et al., 2000). These studies differ from the present study in terms of the retention interval imposed; composites were constructed immediately from memory and also with the target present. The present results are sensible when compared with previous research implementing similar conditions (Frowd et al. 2004, 2005b) who also obtained poor levels of naming following a delay interval of 2 days.

Frowd et al. (2004) found lower naming rates for composites constructed with the PROfit system when a two day delay interval was imposed (mean naming rates of less than 4% from two experiments) and target images were unfamiliar to participants. Similarly, Frowd et al. (2005b) obtained only 2 instances of naming out of 300 attempts for composites constructed using PROfit, with a two day delay interval. The composites constructed of Nicholas Cage and Robin Cook in Experiment 2 appear to be outliers; they are extremely good composites with higher naming levels compared to all other composites. If they are considered as outliers, this leaves only 1 other correct identification out of 432 naming attempts in the experiment as a whole. Frowd et al. explained their poor naming results as a consequence of a two day delay interval and a shift from feature-based to holistic processing. Feature-based systems such as PROfit rely on composites being constructed on a feature by feature basis and therefore do not facilitate holistic processing.

Despite the similar procedures and conditions adopted in this study and the Frowd et al. (2005a) study (results do not appear to be influenced by target unfamiliarity for the evaluator participants, witness participant age or differences in where the pool of evaluators were drawn from in either study), results differed to a great extent in terms of mean naming rates. The PROfit system gave mean naming rates of around 20% in the Frowd et al. study, with a three to four hour retention interval imposed, but mean naming reached only 7.6% for the present study. In contrast, the results from the Frowd et al. (2005b) study were similarly low when compared to the results obtained here at the two day retention interval; mean naming in the Frowd et al. study was only 0.6% using the PROfit system, compared to 3.5% here. It may be possible to explain these results in terms of operator differences and differences in target set, although this may not be so for the 2 day delay experiments where naming results were approaching the same, very poor levels across both operators.

Target images here were unfamiliar to witness participants as in the Frowd et al. (2004; 2005a, b) studies. It is well known that our ability to remember unfamiliar faces is poor although individuals are very good at identifying familiar faces, even from very low quality images (Hancock, Bruce & Burton, 2000). Target unfamiliarity and therefore lower levels of facial feature recall may have contributed to the poor naming rates obtained here. In addition, it is known that individuals are sensitive to the configuration of facial features (Hancock, Bruce & Burton, 2000). It is likely that composites are not an ideal means for replication of the exact configuration of target facial features and as such composite naming may be negatively affected. Furthermore, perhaps configurational relationships are less likely to be replicated to an acceptable extent with a delay in recall of two days or longer.

The results of the present study reveal the discrepancy between ‘good’ and ‘bad’ composites. Good composites may be defined as ones which resemble the facial appearance of a target face to an extent which enables them to be identified to a reasonably consistent degree by a number of individuals. Composites that are bad do not appear to be identifiable to any number of individuals, or at most to very few.

Naming was very low in the present study on retention interval suggesting that composite quality was very poor and that there was little resemblance between the target images and the composites constructed with a retention interval of 3 - 4 hours, 2 days or 1 week. Mean likeness ratings were also low, verifying the poor likenesses achieved and the very low levels of subsequent naming, although composites constructed after 2 days received significantly higher likeness ratings than those constructed after a longer delay. Where there were instances of high levels of naming (Nicholas Cage), mean likeness ratings do not appear to reflect the high level of resemblance of this composite with the target image. A similar effect was found by Green and Geiselman (1989) who suggested that the more sensitive method for evaluating the usefulness of facial composites is through performance measures (i.e. naming) than through likeness ratings, although ratings in their study were from witness participants themselves, rather than from independent evaluators as in the present study.

The present results indicate that memory performance in facial recall is generally poor. In addition, that recall accuracy from a witness who is unfamiliar with a target face is poor and identification results based on the resulting facial composites with a retention interval of only a few hours to a week, are at extremely low levels. Additionally, poor memory performance in terms of recall ability with delays of two days or more may mean that the use of artistic enhancement techniques to 'improve' the composite may be of little value. There is some evidence, from the experiments undertaken here, to suggest artistic enhancement may be beneficial following a retention interval of one week, however further study of the effects of artistic enhancement is required to clarify this. These results are cause for concern within a practical forensic setting; it is not always possible to undertake the facial imaging interview with a witness within a short time-frame, and in fact, research has shown that intervals of two days are average. These results also suggest that there is little value in a composite constructed after two days of an incident occurring. However, composites produced at an interval of two days do appear to be superior in quality to those constructed after two days or more when using likeness ratings as the assessment measure. From a forensic perspective, there does appear to be an advantage to constructing composites at a two day retention interval compared to longer intervals of up to one week. Further, it may be advantageous to undertake the facial imaging process prior to two days; there was a trend in this direction from the naming scores and previous researchers have shown that shorter delays (immediate construction or 3 – 4 hours) result in increased identification.

4

Review, comparisons, forensic validity and the direction of future research

Review of present studies

Facial composite research to date has attempted to determine the reasons behind poor identification levels from facial composites, in spite of improvements in composite construction systems and interviewing techniques. The research undertaken here has sought to provide further explanations for these poor identification levels, in light of previous research results, using forensically valid procedures where possible. The first experiment assessed the main factors of operator performance, the effects of experience and level of skill on resulting composite quality and subsequent identification and, target distinctiveness. The second experiment concentrated on retention interval, from the time of viewing a ‘target’ face to construction of a facial composite, and artistic enhancement, the extent to which artistic tools and techniques can be used to improve the identifiable qualities of the facial composite. A third experiment focused upon retention interval, this time introducing a shorter time-lapse between viewing and composite construction.

Given that modern composite systems provide greater flexibility in terms of the number of features available, the photographic appearance of these features and the use of various tools and techniques to create increased human resemblance, it is considered that the systems require greater operator skill and experience to achieve optimal results. Early research findings on facial composites suggested that there was no difference in composite quality whether a professional police operator or an inexperienced novice operator were constructing the composites (Ellis, Davies & Shepherd, 1978), and that there were inflexibilities in the composite system itself (Laughery & Fowler, 1980). Later results were in contrast and clearly supported the view that operator experience influences the quality of likeness achieved by a witness participant (Davies, Milne & Shepherd, 1983; Gibling & Bennett, 1994).

The first experiment here directly assessed the identification results of three novice operators against those of an experienced operator. Differences in the procedures implemented across the two studies make it difficult to draw any firm conclusions in relation to operator performance (the novice operators' witness participants were familiar with the target faces and composites were constructed immediately following viewing of the target whereas, the experienced operators' witness participants were unfamiliar with the target faces and had a 3 - 4 hour delay prior to construction). There was no significant difference in the performance of the novice operators. Target distinctiveness was studied as an additional factor. Results were in line with those of Frowd et al. (2005a) among others (e.g. Shapiro & Penrod, 1986; Valentine & Bruce, 1986), who reported that faces classified as having high distinctiveness were better recognised than those with low distinctiveness.

Given the practical application of research results in the area, retention interval from viewing of a target face to construction of a composite from memory has been of interest in the area of facial composite research in recent years (Green & Geiselman, 1989; Koehn & Fisher, 1997; Frowd et al. 2004, 2005a, b). The second experiment concentrated on whether there would be a difference in the identifiable quality of composites constructed with a two day delay compared to those constructed after a week's retention interval. It was anticipated that composites constructed with a two day delay interval would include greater recall detail and would therefore be of superior quality and result in increased identification. These predictions were made given that recall is often most accurate within a short period from the time an incident occurs and that recall of any information is superior after as short a retention interval as possible (e.g. Ellis, Shepherd & Davies, 1980). Naming was extremely poor and provided no significant difference between the two retention intervals. Likeness ratings, however, showed that identification of composites was significantly better following a 2 day delay compared to a week delay, thus supporting the main hypothesis. Implementing a shorter retention interval provided no significant advantage for composites constructed after 3 - 4 hours compared to 2 days, in a third experiment. However, naming results did show a trend in this direction.

As an additional factor, the second experiment was the first piece of research to assess the effects of artistic elaboration from a witness's memory. Composites were saved with various levels of enhancement throughout the construction process. As with previous research findings (Gibling & Bennett, 1994), identification results here showed significant differences between composites produced with or without artistic elaboration but only for composites constructed with a weeks retention interval (and these results were based largely around one composite).

This chapter will discuss the main outcomes of all three experiments in relation to one another to determine any emerging trends, with particular emphasis on their practical application within a forensic setting. Comparisons will be made with other research in the areas of: operator performance, artistic enhancement, target distinctiveness and retention interval. In addition, the discussion will identify future areas of research in relation to these factors and the success of facial composite systems.

Comparisons across studies

An assessment of performance across all three experiments is pertinent to determine any similarities or differences in identification levels. Identification performance was assessed here using composite naming levels which, although is a forensically valid measure of assessment, tends to produce results near floor levels (i.e. up to 20%, Davies et al. 2000; Brace et al. 2000; Bruce et al. 2002; Frowd et al. 2005a). The use of naming as the identification measure resulted in no significant differences between most of the variables manipulated across all three experiments. Specifically, operator performance did not differ significantly between an experienced and three novice operators (although this result is problematic given the differences in procedure between operators), naming levels did not differ significantly between novice operators, retention intervals of 3 - 4 hours, 2 days and 1 week did not result in significantly different identification results (although there was a trend in the direction of the 3 - 4 hour delay producing significantly better naming to the 2 day delay). The elaboration of composites did not result in more identifiable composites in general, when compared to those that were not artistically enhanced but significant results were obtained for those composites that were artistically enhanced following a week retention interval. Superior naming resulted from composites rated as highly distinctive in the first experiment but in Experiments 2 and 3 distinctiveness was not manipulated.

The secondary measure of likeness ratings, in Experiments 2 and 3, were used to assess the extent to which composites resembled target faces. Mean likeness ratings were reasonably poor across both experiments, although results for the 3 - 4 hour retention interval compared to 2 days were slightly better than for the earlier experiment involving longer delays (2 days and 1 week). In the experiment assessing the longer retention intervals, likeness ratings were found to be significantly higher for the composites constructed with the 2 day delay.

When comparing results across all three experiments, naming levels were poor in general. The first experiment, which concentrated on operator performance and distinctiveness, provided the best results, with mean naming of around 15% for targets rated as being highly distinctive. This experiment, however, was lacking in forensic validity given that the targets were familiar and composite construction was immediate following target viewing, for the novice operators' witness participants. When a more forensically valid procedure of unfamiliar targets and more realistic retention intervals were implemented in Experiments 2 and 3, naming was at much lower levels (mean naming of between 3.5% and 7.5%). These poor results could not be accounted for by the use of a target set that was not very familiar in general. All target faces were rated by an independent sample of participants prior to the experiment and chosen on the basis that their average rating on a familiarity scale was at least a level three (i.e. 'could say something about why the person is famous but not very confidently'). Good levels of composite identification did not result even though evaluators were generally highly familiar with the target faces.

Naming performance was much poorer in Experiment 3 than in Experiment 1 given that similar procedures were adopted (unfamiliar targets and 3- 4 hour retention interval). By this time, operators from both experiments had been involved in the construction of a good number of composites across several studies and were therefore both well practiced in construction and enhancement techniques and the use of facial composite production systems. It is difficult to draw conclusions as to the reasons for the poorer identification results in Experiment 3 as different operators and targets were used across these two experiments. What can be drawn from these experiments is that it is likely that the operator involved in all three experiments would have gained experience from the first to the third experiment in terms of technical skill (the use of the composite system and the ability to use the tools of the system to enhance the composites). They would also have gained experience in the delivery of the cognitive interview to gain optimal levels of recall. Although operators were experienced in the use of the composite system and in conducting the cognitive interview, poor composite naming resulted in all experiments. Therefore suggesting that even with considerable experience, the poor recall abilities of the witness or the inadequacies of the system itself prevent composites of identifiable quality being produced, rather than the operators own skills and abilities.

Despite choosing targets with high levels of distinctiveness (shown to facilitate recall and recognition, e.g. Winograd, 1981; Valentine & Bruce, 1986a), composite identification remained at low levels across all experiments, particularly in Experiments 2 and 3. Experiment 1 gave a distinctiveness effect, with much greater naming resulting for those composites constructed of targets rated as being highly distinctive. However, given the poorer naming levels in Experiments 2 and 3, the use of distinctive faces did not appear to aid with the identification of target faces when targets were unfamiliar to those constructing and there was a retention interval. Perhaps after significant delays of a few hours to a week, recall of the distinctive features of the face was apparent but recall of additional facial information at this time was minimal. Therefore, the distinctive aspects of the face could be recreated in the composite construction phase but very few descriptors were available from which to represent the other facial features. Further, with a longer retention interval, it is likely there would have been increased disruption in the memory for the configurational relationships of the individual features of the face (Matthews, 1978; Tanaka & Sengko, 1997), meaning that recall may not have been optimal. Taken together with poor recall of the facial features in general, this would produce a poor likeness but one with some distinctive elements that would provide for low levels of identification. Thus, even though some aspects of the composites may have been correct, the presence of inaccuracies lowered naming considerably. Conversely though, it is possible that the use of target faces with low levels of distinctiveness would have produced even poorer levels of naming.

The use of artistic enhancement to elaborate composites was found to produce composites that were significantly more identifiable when a retention interval of one week was imposed. However, these results were based around one identifiable composite and are therefore problematic. Possible reasons for artistic enhancement not yielding significant results for all composites may again be due to the lack of verbal information recalled at the construction phase, when retention intervals of a few hours to a week were imposed. Imposing a retention interval appears to elicit only basic descriptions of the facial features and ones which lack in terms of the types of information required to elaborate features to represent the specific target's facial features. Formal analysis, however, of verbal descriptions was not within the scope of this thesis.

Implications

The results of the research undertaken here will now be further discussed in relation to previous research findings. In general, memory performance for recall of facial information across all experiments was poor. Results from the first experiment provided mean naming of around 15% (for composites rated as being highly distinctive). This finding was not dissimilar to the general finding of other researchers in the area of facial composites who found consistently poor identification rates of around 20% (Brace, Pike & Kemp, 2000; Davies, van der Willik & Morrison, 2000; Bruce et al. 2002; Frowd et al. 2005a). The first experiment here involved witness participants who were familiar with the target faces and immediate construction, similar to the Brace et al. (2000) and Davies et al. (2000) studies. These procedures are not particularly relevant to a forensic setting where the victim of a crime is likely to be unfamiliar with the perpetrator and there will be a typical delay of two days (Frowd et al., 2005b) before a facial imaging interview takes place. In light of this, the Bruce et al. (2002) and Frowd et al. (2005a) studies implemented the viewing of unfamiliar target faces and the Frowd et al. study went further still to include a slightly more realistic retention interval of 3 - 4 hours. Identification results from the second and third experiment reported here were much poorer than in the first experiment. These experiments adopted realistic retention intervals and unfamiliar target faces. In general, identification of composite faces across all three experiments was extremely poor and naming was at much lower levels than in previous research, as discussed earlier.

A further finding of these experiments and one that is in line with previous research, concerns retention interval (from viewing of a target face to composite construction). It is suggested that the retention interval imposed in the second and third experiments affected the level of facial recall from memory and led to poorer naming of the resulting composites. Identification results were poor here in comparison to studies that employed immediate construction. Previous research has shown that composites produced from memory lack effectiveness (Wogalter & Marwitz, 1991; Brace, Pike & Kemp, 2000) and that composites are of a very poor quality and lacking in their utility with delays of two days (Koehn & Fisher, 1997).

Recent research has attempted to maximise forensic validity by adopting retention intervals ranging from a few hours to one week (Koehn & Fisher, 1997; Frowd et al. 2004, 2005a, b). This was the first study to use the same operator to compare the same target faces across two different but realistic and forensically valid retention intervals (other researchers have used immediate construction or very short delays i.e. half an hour compared with a longer delay). Naming results did not show a significant difference between delay intervals of 2 days and 1 week, although likeness ratings provided significantly superior results for composites constructed after 2 days. In addition, there were no significant differences for naming and likeness ratings for composites constructed with retention intervals of 3 - 4 hours compared with 2 days, although there was a trend in the direction of greater naming after 3 - 4 hours. Experiments 2 and 3 here provided similarly poor naming results to the Frowd et al. (2004, 2005b) studies when a 2 day retention interval was imposed. However, results differ markedly between the current experiment using a retention interval of 3 - 4 hours (mean naming of around 8%) and the Frowd et al. (2005a) study implementing the same retention interval (mean naming of around 20%).

In terms of their practical application in relation to policing, these results are of concern. Earlier research where target images were in-view or construction was immediate did not obtain particularly high identification results (e.g. Davies & Oldman, 1999; Brace, Pike & Kemp, 2000; Davies et al., 2000). This poor performance is exacerbated when a retention interval is introduced. Recall levels are extremely low when retention intervals of a few hours to one week are implemented. It is apparent that police interviews do not take place within a short time-frame and that a facial imaging interview does not normally take place until around two days after an incident has occurred (Frowd et al. 2005b). There appears to be little value in composites constructed with a retention interval greater than two days but results point to an advantage for composites constructed with a retention interval of two days. It may be still more beneficial to conduct the facial imaging interview after a delay of only a few hours. These results suggest a change to police practice would be the only possible means to gain the most from facial composites.

A further finding of this research concerns operator performance. It was expected that an experienced operator would out-perform novice operators in terms of composite identification rates. Given that the experienced operator would have greater knowledge and practice in the technical skills required to produce recognisable composites and in the use of the cognitive interview, it was hypothesised that they would have the ability to produce composites that were of superior quality. Results from the first experiment were not in-line with these expectations and composites were of poor quality in terms of naming for both the novice and experienced operators. One must not place too much emphasis on a comparison of naming rates between the experienced and novice operators, however, given the differences in procedure implemented under the two conditions.

Further, results from the novice operators did not differ significantly. One might have expected the novice operators to differ in terms of their level of artistic skill and subsequently that their composites would differ in identifiable quality. This was not the case and possible reasons for this may be that the composite system could simply not be used effectively to illustrate artistic enhancements or that although the novice operators had undergone some training in the use of the system and had some practice, their level of experience with the system was simply not great enough. Alternatively, it may be the case that the poor recall abilities of the participants led to poor facial representations. A further possibility is that there were subtle differences in the level of artistic skill of each novice operator but inadequacies of the composite system itself made it difficult for these differences to be observable in composite quality; perhaps the editing tools of the system were not useful.

Results from all operators were not too dissimilar to previous researchers who found naming of around 20% (Brace, Pike & Kemp, 2000; Davies, van der Willik &

Morrison, 2000; Bruce et al. 2002; Frowd et al. 2005a). However, the limitations of this experiment were that procedures used for the novice and the experienced operator differed, therefore making it difficult to draw firm conclusions in relation to the similar naming performance resulting from the novice and experienced operators' composites. The results should be interpreted with caution given these differences in procedure.

Although it is difficult to draw conclusions between the performance of the novice and experienced operators, one finding of interest is the poor naming results obtained for the experienced operator. This operator had undergone training at an accredited facial composite course provided for police operators, covering the practical aspects of the facial composite software and the cognitive interview procedure. The operator was a lab-based researcher in the area of facial recognition and had previously constructed composites in a number of studies. Despite this level of practical experience with the composite production software and the cognitive interview procedure, overall composite identification remained poor. It may have been the case that there was no discernable difference in the identifiable quality of the composites produced by the novice and experienced operators and that operator experience did not play a significant role. However, this is only one interpretation of the results here and one which could only be drawn where procedures remained the same across both operators.

Although differences in procedure are apparent between the novice and experienced operators (familiar vs. unfamiliar targets and immediate construction vs. 3 – 4 hour delay), recent research has shown that target familiarity does not affect composite quality (Frowd et al., in press), especially where composites are constructed from memory (Davies et al., 2000). Therefore, one might argue that the procedural differences in relation to target familiarity are unlikely to have impacted upon the overall results. However, this is only one interpretation of the results and target familiarity has been shown to effect recall ability in a good number of studies in this area. Additionally, the differences in retention interval still remain.

It is difficult to relate the results of the experiment concerning operator experience to those of other researchers in this area, given the issues surrounding the differences in procedure. Additionally, the work undertaken here in relation to operator experience may have benefited from the inclusion of a greater number of variables under measurement in order to make clearer the areas in which experience is advantageous. An example of one such study that considered some of the factors involved in producing more recognisable composites was that of Laughery and Fowler (1980). They found that differences between composite systems themselves (the IdentiKit system versus a sketch artist) rather than the operator's technical abilities affected composite quality. They highlighted the fact that it was artistic skill that was of greater importance to the construction process, rather than the operator's abilities to use the system. Artistic skills were found to play an important role for several reasons: an infinite number of features could be created and the sketch artist could add greater detail and use more artistic effects in terms of shading etc. Bennett (1986) was also able to recognise the value of artistic skill in composite production. In a discussion paper on the early use of the Photo-FIT system, it was recognised that operators needed on-going training in order to become skilled in the use of the system and in their own artistic abilities. Training in the development of artistic skill was observed to be of value.

Davies, Milne and Shepherd (1983) also looked at the specific role of experience in composite construction. Experienced operators were shown to out-perform novice operators on a number of levels: initial questioning of the witness; relating the witness description to the features of the kit (ie. initial selection of the features); and applying technical skills to produce a good likeness. A subsequent study reported that the application of technical enhancements was not important but the initial selection of features was where the experienced operator excelled. Gibling and Bennett (1994) studied the effects of experience in terms of artistic enhancement. They found that PhotoFits elaborated with pencil work by experienced operators were more alike targets than were the non-elaborated PhotoFits.

Although the operators in the experiments reported here had been involved in formal training and had significant experience and practice in the use of the editing features of the PROfit system for artistic enhancement of the composites, composite quality and subsequent identification remained at low levels. These results appear to contradict Bennett (1986) who suggested that lack of operator skill is one of the reasons accounting for poor system performance. Operators here possessed knowledge of, and had skill in, enhancement techniques critical to a composite system’s potential as defined by Bennett. It may be the case that the system’s potential in this, and in Bennett’s study, had been maximised and that other factors are indeed responsible for the poor identification performance. Perhaps, as Davies et al. (1983) suggested, experience should not be measured in terms of technical skill and the ability to enhance composites but in the initial selection of the features. It is possible that the novice and experienced operators' abilities were lacking in this area, and therefore naming levels were low. Conversely, it may be the witness participants themselves who were unable to provide detailed descriptions of the target's face from memory. The first possibility can be remedied through training and skill enhancement exercises, whereas the latter is unlikely to be improved upon in this manner.

The results obtained here in relation to operator performance are problematic in terms of interpretation of their practical application. They provide little insight into whether an operator experienced in the use of the composite production system and the cognitive interview is able to out-perform a novice operator, when the same systems are used by both. Retention interval (e.g. Mauldin & Laughery, 1981; Green & Geiselman, 1989; Koehn & Fisher, 1997; Frowd et al., 2005a, b) and familiarity with the target (Ellis et al., 1979; Young et al., 1985; Bruce & Young, 1986) are factors known to impact significantly upon recall accuracy in facial recognition. Therefore, the absence of significant differences in the results from the novice and experienced operators may be explained by these two factors, rather than from any actual lack of difference between the operators performance that may have occurred if similar procedures had been followed by both. However, the confounds of the experiment make this a difficult argument. The counter-argument is that composite quality is not affected by target familiarity (Davies et al., 2000; Frowd et al., in press). If this were the case, one might place more emphasis on the overall results obtained here, that there was no significant difference between the performance of an experienced operator in comparison with three novice operators. The results provide further knowledge in terms of identification levels from composite systems in general. Even where good levels of practical experience with the composite production software and the cognitive interview procedure were apparent, overall composite identification remained poor.

The effects of artistic enhancement were investigated in the second experiment. To a certain extent, it is difficult to divide operator skill and artistic enhancement as factors. One would expect a more experienced operator to be more adept in artistic skill and enhancement techniques to produce a good facial likeness. The results obtained here go some way towards the conclusion that artistic enhancement of composites is beneficial after a retention interval of one week. Composites were saved at three pre-defined stages, which included no artistic elaboration or editing of the features; editing of the individual features to reflect the specific appearance of the target's feature; and, the use of feature blending of the 'completed' composite. There was a significant difference between identification of the composites at the feature blended and no elaboration stages with retention intervals of two days and one week.

This was the first study to assess the effects of artistic enhancement from memory. Previous studies have concentrated on the effects of artistic enhancement while composites were in view (Gibling & Bennett, 1994) or with immediate construction (Davies, Milne & Shepherd (1983). One suggestion for the significantly superior enhanced composites produced with the week retention interval is that although little detail may be recalled after such a delay, the effects of artistic enhancement outweighs the fact there may be less factually correct detail. In contrast, there is no advantage for enhancement at the two day retention interval as more correct detail is recalled. As the results relating to artistic enhancement were largely based around one or two ‘good’ composites here, further study is necessary to clarify the position. What is true is that considerable retention intervals of two days and one week are likely to result in poor recall of faces, and perhaps shorter retention intervals would have been more appropriate to use to study the effects of enhancement. Longer retention intervals would lend themselves to poor recall accuracy which would, in turn, impact upon the types of verbal description offered by the witness participant. Prolonged retention intervals and less recall are likely to exacerbate the problems relating to the translation of the facial memory. As Christie and Ellis (1981) point out, a witness may have a reasonable memory of a target face, but it is often difficult to describe this to another person. Further, the resulting composite is likely to be affected due to inadequate communication of the facial features (Wogalter & Marwitz, 1991). Informal assessment of the results for the 2 day and week retention intervals provided some evidence that witness participants’ descriptions lacked some content in terms of the amount of detail provided, in the week delay condition when compared to the 2 day delay condition. Formal analyses were not conducted thereof, not being central to this thesis.

It is ventured that artistic elaboration of the facial features is only likely to occur where the witness participant is able to recall a high level of detail surrounding the particular appearance of the specific feature. In the experiment reported here initial descriptions of the individual features tended to be superficial and did not include information that would necessarily permit artistic elaboration at the ‘elaboration’ stage of the experiment (ie. changes involving the addition of marks and scars, eyebags etc.). Most notably, descriptions at any stage of the cognitive interview did not include information on contour or tone (except in the case of eyes and hair colour); artistic enhancement techniques such as shading and blending would have been necessary to reflect both of these. Laughery and Fowler (1980) found that the use of shading and other techniques for reflecting colour and contour produced better images. Comparatively, sketch artists were able to add greater detail to their composites in terms of shading and other artistic effects that were not available in the IdentiKit system, and produced significantly better composites. It is perhaps the case that recall abilities do not allow such information to be accessed either during the cognitive interview or during the composite construction session. It is notable, however, that the addition of feature blending was advantageous after longer retention intervals. This stage doesn’t involve editing of the features but, simply blending to fit the background face.

From a practical perspective, the results relating to artistic enhancement suggest that when composites are constructed from memory, artistic elaboration of the individual features may be of value for considerable retention intervals of around one week but it is probably not of value to elaborate with retention intervals of around two days. It appears that the final stage of blending the features to fit the background face is of benefit to the overall appearance of the composite.

Feature distinctiveness is a further factor examined in this thesis and in face recognition literature (e.g. Winograd, 1981; Courtois & Mueller, 1981). In their meta-analysis of facial identification studies, Shapiro and Penrod (1986) recognised target distinctiveness as one of the variables involved in face perception. They showed that various studies had pointed to the influence of distinctive facial features in facial identification: distinctive targets were more easily recalled than ordinary looking targets. The first experiment here found a distinctiveness effect. Results revealed a major difference in terms of identification, with faces rated as being highly distinctive gaining significantly superior identification levels than faces rated as having low levels of distinctiveness. These results replicated the effects of distinctiveness found by Frowd et al. (2005a) in their study of five composite production systems, wherein a distinctiveness effect was apparent across all five systems.

Green and Geiselman (1989) were unable to find a superiority effect for salient faces in their study using the IdentiKit system. They found that identification for composites of ‘average’ faces was better than for distinctive faces. However, that the Green and Geiselman study was unable to find positive identification results from distinctive targets may actually be explained by the composite system itself. Unlike the experiment reported here and the Frowd et al. (2005a) study, which used modern computer-based systems, the Green and Geiselman study may have suffered from the use of the older, less flexible system. They reported that replication of salient features was not always possible given the limitations of the system. Therefore, where salient features were present on the target face, this negatively affected the quality of the resulting composites, due to the absence of distinctive features in the IdentiKit system's library of features. However, contemporary composite systems may provide advantages in terms of reproduction of distinctive features: increased numbers of features in the database and the use of editing tools to specifically modify features are thought to be advantageous.

In terms of practical applicability, the results of the present experiment showed that target faces were found to produce superior quality composites that were more identifiable where they were consistently judged as distinctive. Modern composite systems appear equipped to replicate distinctive faces or features to a greater extent and there is a major discrepancy in identification for those faces that are described as typical; identification here is at much lower levels. The current results are positive for occasions where a target face is considered to be distinctive by a number of people; the target faces in this experiment were selected for their use on the basis of average distinctiveness ratings from a sample of independent evaluators. The problem is that there may be only one eyewitness to an incident in a police investigation and factors associated with a stressful experience may induce increased subjectivity in terms of the level of distinctiveness associated with the crime perpetrators facial appearance. Distinctiveness cannot, therefore, be measured objectively.

It is documented in the literature that poor performance in studies of facial composite systems in comparison with performance from face recognition studies in general, may result from poor system design (e.g. Davies & Christie, 1982). Evidence suggests that faces are stored and recognised as wholes and that it is difficult to extract individual feature information regarding faces (e.g. Tanaka & Sengko, 1997). Face composite systems generally require the face to be constructed serially, concentrating on each of the individual features separately, therefore making retrieval difficult. The first experiment attempted to gain insight into the ways in which we encode faces and the effects the method of encoding has upon the retrieval stage of the face composite construction process, by asking witness participants to rate the extent to which they could visualise the target face on a number of measures. The following measures were assessed: the extent to which the target face could be visualised; the number of features that could be visualised; the extent to which the relationships between the features could be visualised; the extent to which a composite resembled a target face; and attitude towards the target face (using a negative - positive scale). No significant correlations were found on any of the measures with naming levels.

The fact that participant ratings on all of the visualisation measures were not found to correlate significantly with identification of the composites may provide some evidence towards a holistic encoding strategy. As Yount and Laughery (1982) suggest, the more familiar we are with a face, the more we process that face holistically. The construction process in this experiment may have been restricted by the fact that participants were constructing faces with which they were familiar. Participants here appear to be paying attention to the face as a whole rather than to the feature elements of the face based on the visualisation ratings provided. Intuitively, one would expect that individuals paid particular attention to the individual features of the face as they were aware of the nature of the experiment, specifically that they would construct a facial composite. However, the participant witness ratings obtained here do not appear to reflect this suggestion. An alternative explanation for the results obtained here is that it is simply extremely difficult to create a composite of someone with whom we are highly familiar. We may never be satisfied that we have a good representation of a person's face due to the fact that composites do not fit with our perception of the human facial form. This explanation is supported by the fact that results were approaching a negative correlation between the extent to which participants were able to visualise the target face and naming levels. Although participants could visualise target faces well, this did not mean they were able to produce a highly identifiable composite of that face. Naming levels, however, were very low in general meaning that this negative correlation may be spurious. Additionally, results in relation to those composites that were correctly named show that the majority were rated as being highly familiar by the witness participants constructing. Therefore, familiarity appears to have aided identification rather than the opposite effect.

Direction of future research

The experiments discussed have highlighted some of the factors affecting facial composite production and the extent to which these factors can facilitate or impede the composite construction process. The factors of operator performance, artistic enhancement target distinctiveness and retention interval have been manipulated in three experiments. It is perhaps worth mentioning that there are a number of variables operating and interacting at the encoding and retrieval stages of face recognition that produce strong effects on recognition performance. These include context reinstatement, transformations in the appearance of faces, depth of processing strategies, target distinctiveness, elaboration at encoding and exposure time (Shapiro & Penrod, 1986). Additionally, there are a number of variables relating to the target faces themselves (i.e. age, mode of presentation, feature distinctiveness etc.) and to the composite systems (i.e. number of features in the database, standard and inclusion of editing tools, line drawings or photographic feature representations etc.) that are likely to impact upon the quality of the composite and subsequent identification levels. The present studies have attempted to keep these variables constant as far as possible and any influence these factors may have had has been kept to a minimum through randomisation procedures. These findings are borne out by the fact that there is a high level of similarity between the identification levels found in these experiments and those of previous researchers in the area using similarly forensically valid procedures: composites constructed from memory of an unfamiliar face; realistic retention intervals; younger, highly familiar targets (famous faces); experienced artists or operators; one composite constructed of one target; a cognitive interview with no limit on construction time and artistic elaboration allowed (Frowd et al., 2004, 2005b).

The consistent finding throughout this research has been that naming is at extremely low levels when composites are constructed using forensically valid procedures. Future composite research could focus upon techniques for increasing composite quality. As found here, one way to do this is to reduce the retention interval from viewing to construction. However, as mentioned previously this is highly impractical in policing. A further method for increasing composite quality would be to adopt the methods of the sketch artist, who uses a more holistic approach to composite construction, within the computerised composite system (Frowd et al. 2005b). The EvoFIT system is an alternative computerised system developed with the holistic approach to construction in mind (Frowd et al., 2004). This system combines 'eigenfaces' (a simplified form of a large data set of faces using Principal Components Analysis) which become closer to the target face as the process moves forward; a number of possible faces are presented to the witness and the witness selects those that are most alike the target. In this way, the system eradicates the need for assembling features individually, found to be disadvantageous to the construction process. In their study assessing the effectiveness of EvoFIT in comparison with other composite systems, Frowd et al. (2005b) found that the manual process used by the sketch artist provided superior results over all of the computerised systems.

Future research on retention interval could focus on composite systems other than PROfit. The adoption of similar procedures and retention intervals in an attempt to determine whether similar results could be obtained across systems would be beneficial to our knowledge of which systems provide superior results in terms of identification. Additionally, further study might benefit from the inclusion of a greater number of targets and witness participants.

The finding reported here in relation to target distinctiveness (that targets rated as being highly distinctive were identified significantly more often than those rated as having low levels of distinctiveness) appears to be a strong one and one that has been found consistently in the face recognition literature (Courtois & Mueller, 1981; Valentine & Bruce, 1986; Shapiro & Penrod, 1986; Green & Geiselman, 1989; Vokey & Read, 1992) and in a similar research study (Frowd et al., 2005a). Although a positive finding for cases where the crime perpetrator can be classified as having a distinctive face, this finding is problematic for instances where a face is not classified as being distinctive and recall may be negatively affected. Future research in the area of target distinctiveness could thus apply detailed focus on the extent to whether it is the composite system or operator that are unable to reproduce the specific features required, especially with typical target faces. It may be simply that the witness has a poor memory for the face because of its typical appearance and recall is consequently poor.

A related area of future work could be to determine the level of distinctiveness the target face has according to the witness's own impression. Current research practice is to test the distinctiveness of target faces by asking an independent sample of volunteers their opinion. A more appropriate method and one which would be of greater practical relevance would be to assess the witness's opinion on target distinctiveness. Where target faces are considered to be of low distinctiveness, perhaps the interviewing process (from the first time police officers make contact with the witness) would benefit from the use of rehearsal techniques.

A related area of interest is that of retention interval and rehearsal of the target face. Rehearsal is likely to produce superior recall and lead to better quality composites, especially where longer retention intervals are involved. Future research could concentrate on whether rehearsal is beneficial for various retention intervals and the methods and techniques that the witness could employ for rehearsal.

It would be of benefit to determine the precise role of the operator in relation to techniques for artistic enhancement. It is unknown whether it is of greater importance that the operator is able to select the initial facial features well in the construction process, or that they can enhance these features to reflect specific target appearance. It is necessary to gain further insight into whether the operator is able to approach the witness in a more appropriate manner to elicit superior descriptions, allowing for increased accuracy in the selection of the features at the first stage of the procedure.

A further area of research would be to attempt to determine whether the use of information relating to contour and tone of the facial form would increase composite quality. This could be determined using cued recall via prompting. Example booklets depicting various tones for the facial features and levels of shading to reflect depth and contour on the face could be used. Further study is required to determine whether the benefit of artistic enhancement following a week retention interval reported here can be replicated. This result is a positive one in terms of its practical application, specifically, that in police investigations it can often be several days before a facial imaging interview takes place.

In summary, this thesis has concentrated on a number of factors relating to composite production, using the PROfit and E-FIT systems, and has discussed their meaning from the results of three experiments, in terms of practical application in a police setting. Of primary importance (and an effect seen in each experiment) was that identification levels were very poor and memory performance in terms of recall ability was poor overall (although distinctive faces were significantly better identified than typical ones). This was especially so where composites were constructed using the more forensically valid procedures of unfamiliar target faces and realistic retention intervals. In terms of generating a correct name, there appears to be little value in constructing a composite after two days of an incident occurring, although composites constructed at an interval of two days may be superior in terms of identification. Further, it may be advantageous to undertake the facial imaging interview after only a few hours, a finding which is likely to be difficult to exploit in a practical setting due to known restrictions in average times from incident to interview.

Of further importance, there does appear to be some value in the use of artistic enhancement techniques when faces are constructed from memory. It may be that low levels of recall from memory mean that artistic enhancement techniques cannot be applied effectively. Results here, however, do suggest that it may be advantageous to use artistic enhancement with composites constructed with longer retention intervals of around one week.

Given the issues surrounding the differences in procedure implemented in relation to operator experience, it cannot be determined whether an experienced operator is able to out-perform a novice operator. Future study would certainly benefit from concentrating on a good number of the factors involved in composite construction (initial feature selection, relating the verbal description to the composite, the manner in which the 'witness' is approached, the number of modifications made to the initial composite or the length of time taken to construct, the ability to elaborate the composite through artistic enhancement and its value etc.) and the definition of the role experience plays in each of these factors. Finally, the results provide some evidence that the target faces were encoded holistically. This may be one reason for the poor naming results obtained; it is well documented in the literature that facial composite systems are unlikely to facilitate holistic retrieval. Systems in future development may benefit the composite construction process by minimising the requirement for a serial approach to recall of the features.

References

ACPO(S) (2003). National Working Practices on Facial Imaging. Association of Chief Police Officers (Scotland) Working Group.

Bartlett, F. C., (1932). Remembering: A study in experimental and social psychology. Cambridge: Cambridge University Press.

Bennett, P. J. (1986). Face recall: A Police Perspective. Human Learning, 5, 197 – 202.

Brace N. A., Pike, G. E. & Kemp, R. I. (2000). Investigating E-FIT using famous faces. In Czerederecka, A., Jaskiewicz-Obydzinska, T. and Wojcikiewicz, J. (Eds.). Forensic Psychology and Law. 272 – 276, Krakow: Institute of Forensic Research Publishers.

Bruce, V., Hanna, E., Dench, P. H. & Burton, M. (1992). The importance of 'Mass' in line drawings of faces. Applied Cognitive Psychology, 6, 619 – 628.

Bruce, V., Ness, H., Hancock, P. J. B., Newman, C. & Rarity, J. (2002). Four heads are better than one: Combining face composites yields improvements in face likeness. Journal of Applied Psychology, 87 (5), 894 – 902.

Bruce, V. & Young, A. (1986). Understanding face recognition. British Journal of Psychology, 77, 305 – 327.

Christie, D. & Ellis, H. (1981). Photofit constructions versus verbal descriptions. Journal of Applied Psychology, 66, 358 – 363.

Courtois, M. R. & Mueller, J. H. (1981). Target and distractor typicality in facial recognition. Journal of Applied Psychology, 66, 619 – 645.

Cutler, B. L., Stocklein, C. J. & Penrod, S. D. (1988). An empirical examination of a computerised facial composite production system. Forensic Reports, 1, 207 – 218.

Davies, G. (1983). The recognition of persons from drawings and photographs. Human Learning, 2, 237 – 249.

Davies, G. & Christie, D. (1982). Face recall: An examination of some factors limiting composite production accuracy. Journal of Applied Psychology, 67, 103 – 109.

Davies, G. M., Ellis, H. G. & Shepherd, J. (1978). Face identification: The influence of delay upon accuracy of PhotoFit construction. Journal of Police Science and Administration, 6, 35 – 42.

Davies, G. & Little, M. (1990). Drawing on memory: exploring the expertise of a police artist. Medical Science and the Law, 30 (4), 345 – 353.

Davies, G., Milne, A. & Shepherd, J. (1983). Searching for operator skills in face composite reproduction. Journal of Police Science and Administration, 11, 405 – 9.

Davies, G. & Oldman, H. (1999). The impact of character attribution on composite production: A real world effect? Current Psychology: Developmental, Learning, Personality, Social, Spring 1999, 18, (1), 128 – 139.

Davies, G. M., Shepherd, J. W. & Ellis, H.G. (1978). Face recognition accuracy as a function of mode of representation. Journal of Applied Psychology, 63, 180 – 187.

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

Ebbinghaus, H. (1885) cited in Eysenck, M. W. & Keane, M. T. (1995). Cognitive Psychology: A students handbook. Psychology Press Ltd, UK.

Ellis, H. D., Davies, G. M. & Shepherd, J. W. (1978). Remembering pictures of real and ‘unreal’ faces: Some practical and theoretical considerations. British Journal of Psychology, 69, 467 – 474.

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

Ellis, H. D., Shepherd, J. W. & Davies, G. M. (1980). The deterioration of verbal descriptions of faces over different delay intervals. Journal of Police Science and Administration, 8, 101 – 106.

Frowd, C. D., Bruce, V., McIntyre, A., & Hancock, P. J. B. (under revision –c). The relative importance of external and internal features of facial composites. British Journal of Psychology.

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

Frowd, C. D. Carson, D., Ness, H., McQuiston, D., Richardson, J., Baldwin, H. & Hancock, P. (2005b). Contemporary composite techniques: The impact of a forensically-relevant target delay. Legal & Criminological Psychology, 10 (1), 63 – 81.

Frowd, C. D., Hancock, P.J.B. & Carson, D. (2004). EvoFIT: A holistic, evolutionary facial-imaging technique for creating composites. ACM Transactions on Applied Psychology (TAP), 1, 1 – 27.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P. & Holland, H. L. (1985). Eyewitness memory enhancement in police interview: Cognitive retrieval mnemonics versus hypnosis. Journal of Applied Psychology, 70, 401 – 412.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P. & Holland, H. L. (1986). Enhancement of eyewitness memory with the cognitive interview. American Journal of Psychology, 99 (3), 385 – 401.

Gibling, F. & Bennett, P. (1994). Artistic enhancement in the production of PhotoFit likeness: An examination of its effectiveness in leading to suspect identification. Psychology, Crime & Law, 1, 93 – 100.

Green, D. L. & Geiselman, E. (1989). Building composite facial images: Effects of feature saliency and delay of construction. Journal of Applied Psychology, 74 (5), 714 – 721.

Hancock, P. J. B. (2000). Evolving faces from principal components. Behaviour Research Methods, Instruments and Computers, 32 (2), 327 – 333.

Hancock, P. J. B., Bruce, V. & Burton, M. (2000). Recognition of unfamiliar faces. Trends in Cognitive Sciences, 4 (9), 330 – 337.

Home Office (2001). Criminal Statistics, England and Wales 2001. HMSO: London, UK.

King, D. (1971). The use of Photo-Fit 1970 -1971: a progress report. Police Research Bulletin, 18, 40– 44.

Klatzky, R. L. & Forrest, F. H. (1984). Recognizing familiar and unfamiliar faces. Memory & Cognition, 12 (1), 60 – 70.

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

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

Laughery, K. & Fowler, R. (1980). Sketch artist and IdentiKit procedures for generating facial images. Journal of Applied Psychology, 65, 307 – 316.

Laughery, K. R. & Smith, V. L. (1978). Suspect identification following exposure to sketches and Identi-kit composites. Proceedings of the Human Factors Society- 22nd annual meeting, 631 – 635.

Laughery, K. R., Smith, V. L. & Yount, M. B. (1980). Visual support devices: evaluation of a new technique for constructing facial images. Proceedings of the Human Factors Society 24th Annual Meeting. Human Factors Society: Santa Monica, CA, 302 – 305 cited in Wogalter, M. & Marwitz, D. (1991). Face composite construction: In view and from memory quality improvement with practice. Ergonomics, 22, 333 – 343.

Loftus, E. F. (1979). Eyewitness testimony, Harvard University Press: Cambridge, MA.

Matthews, M. L. (1978). Discrimination of IdentiKit constructions of faces: Evidence for a dual processing strategy. Perception & Psychophysics, 23 (2), 153 – 161.

Mauldin, M., & Laughery, K. (1981). Composite production effects upon subsequent facial recognition. Journal of Applied Psychology, 66, 351 – 357.

Schwaninger, A., Lobmaier, J. S. & Collishaw, S. M. (2002). Role of featural and configural information in familiar and unfamiliar face recognition. Lecture notes in Computer Science, 2525, 643 – 650.

Shapiro, P. N. & Penrod, S. (1986). Meta-analysis of facial identification studies. Psychological Bulletin, 100 (2), 139 – 156.

Shepherd, J. W, Ellis, H. D., McMurran, M. & Davies, G. M. (1978). Effect of character attribution on PhotoFit construction of a face. European Journal of Social Psychology, 8, 263– 268.

Tanaka, J. W. & Sengko, J. A. (1997). Features and their configuration in face recognition. Memory & Cognition, 25 (5), 583 – 592.

Valentine, T. & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. Perception, 15, 525 – 535.

Vokey, J. R. & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. Memory & Cognition, 20 (3), 291 – 302.

Winograd, E. (1981). Elaboration and distinctiveness in memory for faces. Journal of Experimental Psychology: Human Learning and Memory, 7 (3), 181 -190.

Wogalter, M. & Marwitz, D. (1991). Face composite construction: In view and from memory quality improvement with practice. Ergonomics, 22, 333 – 343.

Young, A. W., Dennis, C. H., McWeeny, K. H., Flude, B. M. & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. Perception, 14, 737 – 746.

Yount, M. B. & Laughery, K. R. (1982). Facial memory: Constructing familiar and unfamiliar faces. Bulletin of the Psychonomic Society, 19 (2), 80 – 82.