Testing Featural and Holistic Composite Technologies under Different Methods of Face Recall

by

Bhavesh Prag

Supervised by Associate Professor Colin Tredoux

Submitted in fulfilment of the requirements for the degree of

Master of Social Science (Research Psychology)

at the

UNIVERSITY OF CAPE TOWN

May 2005

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my sincere gratitude to Colin Tredoux (Associate Professor, University of Cape Town) for his excellent supervision of this thesis, despite us being hundreds of miles away from each other for most of the duration. Colin, thanks for introducing me to the fascinating field of eyewitness research, and for sharing your extensive eyewitness research knowledge and statistical expertise with me throughout the production of this piece of work. I have learned a great deal from you.

I am extremely grateful to the Mellon Foundation for awarding me the A. W. Mellon Fellowship, which gave me the opportunity and privilege to meet and exchange ideas with distinguished academics in both the United States and the United Kingdom. A special thanks to Stephen Palmer (Professor, University of California, Berkeley) and Christian Meissner (Assistant Professor, Florida International University) for hosting me during my time at their prestigious universities.

To Janat Parker (Professor, Florida International University), Nadja Schreiber (Post-Doctoral Research Fellow, Florida International University), Margaret Kovera (Professor, City University of New York), Robert Rios (Director, Faces Training), Graham Pike (Lecturer, Open University), Tim Valentine (Professor, Goldsmith College), and Ray Bull (Professor, University of Leicester), many thanks for making time in your busy schedules to meet with me at relatively short notice to discuss my research. Thank you also to Avraham Levi (Israeli Police Department) for sharing your thoughts on how my research design could be improved; and to Inspector Van Dijk (Facial Identification Department, SAPS) for producing the "expert" Faces reconstructions, and for training me up on how to conduct the SAPS Interview. Huge thanks to Ronald Fisher (Professor, Florida International University) and Christian Meissner (Assistant Professor, Florida International University) for assisting me in formulating a Cognitive Interview Technique for the specific task of facial composite production. All of your input and feedback has been invaluable.

I owe a tremendous amount of thanks to Sylvie Lariviere (President, IQ Biometrix) who generously donated a copy of the Faces (version 3) software to me for evaluation. To Frank Bokhurst (Lecturer, University of Cape Town), your assistance in sourcing and scheduling students to participate in my experiments is much appreciated. Thank you also to the respective authorities at the following tertiary institutions, high schools and shopping malls for giving me consent to collect photographs for the E-Face databases: Cape Technikon, Peninsular Technikon, Athlone College, Habibia High School, Rylands High School, Cravenby High School, Trafalgar High School, Ottery Hypermarket, N1-City and Westgate Mall.

Last, but by no means least, I would like to deeply thank my close friends and family, and especially my fiancée Johanna, for the continued support and encouragement which they have given me from the moment I embarked upon my research journey.

Abstract

This thesis evaluates the effectiveness of two facial composite systems, used to produce facial images of alleged perpetrators of crime, under varying conditions in two experiments. The first system is called E-Face, a configural eigenface-based system originally developed at the University of Cape Town and currently undergoing further revisions; the second is called Faces, a featural system currently being used by several law enforcement agencies around the world.

In Experiment 1, an investigation is made as to whether the E-Face system can be improved by removing Indian faces from the coloured database, and creating a new Indian database. Each participant reconstructed two faces using one of three databases, specially created for the purpose of this experiment, each time they reconstructed a face. Photographs of the target faces were in full view during composite production, thereby assessing the fundamental ability of the E-Face system and three of its databases under ideal conditions for novice operators. Subsequent rating and mugshot evaluation task results disproved the hypothesis of the experiment but revealed that the E-Face system compares quite well against previously conducted studies, where novice operators have used various facial composite systems to produce facial composites under "in view" conditions.

In Experiment 2, a thorough comparative analysis is made between the E-Face and Faces systems. To test the optimality of both systems, expert operators produced "in view" reconstructions of four target faces, which were evaluated against novice "in view" reconstructions. To test the two systems under more realistic and forensically important conditions, "from memory" reconstructions were made by novice operators after being exposed to the targets in a live staged event. Prior to making these "from memory" reconstructions, participants either underwent the SAPS Interview or the Cognitive Interview (CI) devised for the specific task of facial composite production. Independent judges evaluated the quality and utility of the resulting composites through rating, sorting and ranking tasks. The Faces system was found to be marginally better than the E-Face system; expert reconstructions were found to be significantly better than novice reconstructions; "in view" reconstructions were found to be significantly better than "from memory" ones; and the CI was found to be less useful a technique than the SAPS Interview in helping witnesses recall the face of the target. Related findings, confounding variables, and ideas for future research are discussed.

Much optimism can come out of this thesis for E-Face, not only because it was found to have better "from memory" performance on average, but also considering that it is a system still in the process of development, and that further refinements that have been made to improve it, are yet to be evaluated. E-Face was found to perform considerably better from memory than Faces. The fact that the Faces system was found to perform considerably better in view than from memory suggests that the featural method of facial composite construction is not best suited to police investigations. This could point the way forward for a more rigorous comparison of the theoretical approaches with composite systems.

Chapter 1: Introduction and Thesis Overview

In eyewitness identification procedures, witnesses and/or victims of crime who have presumably had a good look at the offender are often called upon to construct a facial composite of the offender. The purpose of producing facial composites is to assist investigators in narrowing down the "suspect population" and thus improve the likelihood of identifying and apprehending the offender. A facial composite is normally used in collaboration with other information like the verbal description of the offender, the accent of the offender, the vehicle used to perpetrate the crime, and so on. An important benefit of the facial composite as a lead to an investigation is that it provides a pictorial representation directly from the eyewitness to the viewer, which is not "imagined" as is normally done by the reader of a written statement (Clark, 2000; Koehn, Fisher & Cutler, 1999). However if these facial composites do not represent accurate likenesses of offenders, then they will not aid investigators in bringing offenders to justice.

Law enforcement agencies depend on a variety of methods to help witnesses and/or victims of crime reconstruct likenesses of offenders, including sketch artistry as well as manual and computer-based facial composite systems. These methods are reviewed in Chapter 2 of this thesis. Extensive empirical examination has revealed that the efficacy of composite systems as tools to promote recognition of suspects in criminal contexts is questionable. The bulk of experiments conducted reveal that composite systems produce poor quality composites, which are difficult to match to target faces (Christie & Ellis, 1981; Duggal, Mickus, Daneker & Kassin, 1992; Davies, Van der Willik & Morrison, 2000; Ellis, Shepherd & Davies, 1975; Koehn et al., 1999; Koehn & Fisher, 1997; Kovera, Penrod, Pappas & Thill, 1997; Laughery & Fowler, 1980; Yount & Laughery, 1982). Facial composite systems have been found to be effective only under certain conditions, such as when the target is in full view whilst participants construct the facial composite or when witnesses are exposed to the target and then anticipate the future task of constructing a facial composite (Cutler, Stocklein & Penrod, 1988; Koehn et al., 1999; Penrod & Stocklein, 1992; Wells & Hyrciw, 1984; Wogalter & Marwitz, 1991).

Research generally has indicated that composite systems are quite ineffective in producing accurate likenesses of faces. It needs to be taken into account that the majority of empirical research examining the effectiveness of facial composite systems thus far has focused on four composite systems, i.e. Photofit, Identikit, Mac-a-Mug Pro and E-Fit. The findings of these studies do not reflect a number of newly developed computer-based composite systems (McQuiston & Malpass, 2000). This is the motivation behind the evaluation of Faces, a newly developed computer-based system used by law enforcement agencies around the world, which has never been put under thorough laboratory experimentation.

The Faces system is compared against the E-Face system, an eigenface-based system originally developed at the University of Cape Town and constantly undergoing further revisions. Chapter 4 elaborates further on these two composite systems. A fundamental difference between them is the theoretical bases upon which they have been designed. Faces is designed upon the "featural" face processing theory, whilst E-Face is designed upon the "holistic / configural" theory. These theories are discussed in more detail in Chapter 3. This key difference is one of the reasons for selecting these two particular systems for the comparative analysis. Another important reason is to see how the E-Face system, which is still in the process of development and refinement, compares against a facial composite system used by police worldwide.

The two systems have been carefully evaluated under a variety of conditions. In Experiment 1, outlined in Chapter 6, an investigation is made into whether the E-Face system can produce satisfactory likenesses of faces from different but related face populations under "in view" conditions for novice operators. In Experiment 2, outlined in Chapter 7, a thorough comparative analysis is made between the E-Face and Faces systems. To test the optimality of both systems, expert operators produced "in view" reconstructions of four target faces, which were evaluated against novice "in view" reconstructions. To test the two systems under more realistic and forensically important conditions, "from memory" reconstructions were made by novice operators after being exposed to the targets in a live staged event. Prior to making these "from memory" reconstructions, participants either underwent the SAPS Interview or the CI. This allowed for a comparison to be made between a standard police interview and an interview technique devised for the specific task of facial composite production. These and other interview techniques used to aid facial composite production are elaborated upon in Chapter 5. The thesis wraps up with concluding thoughts in Chapter 8.

Chapter 2: Methods of Facial Composite Production

A review of various methods of facial composite production is made here, including sketch artistry, manual and computerised facial composite systems. Verbal descriptions provided by witnesses are discussed first as an aid to police investigators in identifying alleged perpetrators of crime.

2.1. Verbal Descriptions

One traditional approach used as a tool to capture a perpetrator of a crime, is the verbal description. The witness to a crime is interviewed to find out a description of the suspect, which is simply transcribed onto paper (Brignull, 1998). According to Shepherd, Davies and Ellis (1978), verbal descriptions are still the primary means of communicating information about appearance. The efficiency with which the receiver will be able to identify the target person will depend in part upon the ability of the informant to imagine or visualise the target, his or her competence in translating details of the image into words, and the receiver's ability to convert the verbal information into some form of image to serve as the basis for recognition. This approach is easy and quick to carry out, and it can be useful when the witness did not get to see the suspect's face properly (Brignull, 1998).

The verbal description approach is however said to be severely limited, since verbal descriptions lack the detail and accuracy required to define a single individual apart from the rest of the general public (Brignull, 1998). As Laughery and Fowler (1980) point out, the assumption underlying the development of facial composite systems is that a verbal description of a suspect's face alone is unlikely to convey sufficient information for purposes of identification. Indeed it is generally assumed that people are simply not good at verbally describing faces because our vocabulary is short of terms to characterise the general physical conformation of the face (Shepherd et al., 1978).

Christie and Ellis (1981) would contest the above argument as they claim that witnesses do have the verbal ability but the problem is in translating this information into an accurate visual likeness. To test this hypothesis, Christie and Ellis (1981) instructed subjects to both verbally describe and construct a Photofit likeness of a target face following a 60-second exposure to the target. They then looked at how well judges identified all six targets from an array of 18 distractors, on the basis of either the verbal description or the Photofit composite. They found that the verbal descriptions were significantly more accurate (48% identification accuracy) than the Photofit composites (23% identification accuracy). Therefore the problem with facial composite systems was, as they hypothesised, not rooted in the witness's verbal ability, but in some other aspect of the system.

2.2. Sketch Artistry

Like verbal descriptions, sketch artistry is an alternative to using facial composite systems to capture a perpetrator of a crime. According the Osterburg and Ward (1997), each sketch artist develops a unique method of working with a witness. In one that has proven successful, the witness describes the perpetrator and answers questions posed by the artist. The artist then goes back to work free from interruption and, most importantly, without being observed by the witness. This prevents the witness from directing and shaping the artist's outlines. Consistent with Gestalt psychology, denying access to the image until its completion enhances viewer or witness perception of what changes are still needed. Subsequent comments made by the witness allow the artist to modify first attempts. This process is repeated until the witness is satisfied or the artist thinks the image cannot be improved (Osterburg & Ward, 1997).

Laughery and Fowler (1980) wanted to see how sketches produced by sketch artists would compare to composites produced by the Identikit system. In the first condition of the experiment, Identikit operators or sketch artists worked in collaboration with 142 participants to construct from description a sketch or an Identikit composite for each of 71 white, male target faces. In the second condition, the artists and Identikit operators produced sketches / composites whilst having photographs of each target face in full view during production. The experiment revealed that, in both conditions, sketches produced by sketch artists were more easily identifiable than composites produced by the Identikit system. One factor, Laughery and Fowler (1980) mention, which may account for this outcome is a limited set of alternative faces one can create with the Identikit system, compared to a sketch artist who can produce an essentially infinite set. A second explanation for sketch superiority is that additional detail (such as shading, age lines, etc.) is typically more predominant in sketches than in composites. A third possible explanation is the total time difference between techniques, i.e. more time is spent generating sketches than composites (Laughery and Fowler, 1980).

It seems that the benefit of the sketch artist approach is that the resulting image is realistic-looking and can be altered to a very fine degree according to the verbal description given by the witness. However, this approach is very much limited by the witness's verbal ability to describe the suspect, since all the features of the suspect's face have to be described "from scratch". Further shortcomings of the sketch artist approach are practical ones, i.e. standard members of the workforce cannot be trained to become police artists, so specialist artists need to be employed, thus introducing major financial constraints. It is also a very time-consuming process (Brignull, 1998).

2.3. Manual Facial Composite Systems

With manual (or non-computer-based) facial composite systems, witnesses pick out an instance of each feature that is most like the feature of the face being reconstructed, and the selected features are fitted together like a jigsaw to form a face (Bruce, 1988; Ellis & Shepherd, 1992). This visual impression can be revised by replacing features, or by moving forehead and chin, until the witness is satisfied that a good likeness has been produced (Barber, 1988).

Two well-known manual systems that have undergone fairly rigorous laboratory testing are the Identikit system and the Photofit system. There are five sets of facial features available in the Identikit system, i.e. forehead and hair; eyes and eyebrows; nose; mouth and lips; and chins. These facial features are printed on acetate transparencies, and acetates are laid on top of each other to create a composite (Ellis et al., 1975). According to Shepherd and Ellis (1996), when P. J. Dunleavy first released Identikit in 1959, features comprised line drawings. These were later replaced in 1975 with photographic elements, which resulted in Identikit being renamed to Identikit II.

Penry (1970, 1974) developed the Photofit system. He first introduced it to the United Kingdom during the early 1970's. Photofit is similar to Identikit II in that it also comprises photographic elements. The main difference is that facial features in Photofit are printed on jigsaw-like pieces that slot into a template. The Photofit system contains 550 facial features whereas the Identikit II system has 470 available. For both systems, facial features are elaborated with the use of a marking pencil (Shepherd & Ellis, 1996).

In an early study aimed at testing the effectiveness of the Photofit system conducted by Ellis et al. (1975), participants were made to compile composites using the Photofit system whilst having full view of the target face or after a 10-second exposure. The composites produced were rated for likeness to their target by independent judges using a 7-point scale. They found that participants experienced difficulties in making up a reconstruction of a face using the Photofit kit even when the target was present and all the necessary features were available. In addition, it was found that when a face was to be recalled from memory, accurate performance was further impaired (Ellis et al., 1975).

Similar low levels of success were found in an experiment conducted by Ellis, Davies and Shepherd (1978). Participants produced composites after either a 15-second exposure or a 2.5-minute exposure to the target and either intentional or non-intentional face learning. The composites produced were rated for likeness to their target by independent judges using a 7-point scale. No significant differences were found in rating scores of Photofits compiled with the targets in view and from memory. The authors hypothesised that more accurate composites should be produced when targets are in view as this allows for attention to detail. Since their hypothesis was disproved, they claim it is the Photofit system that limits performance rather than the participants' memory of the target.

Christie and Ellis (1981), in their study, compared the effectiveness of composites produced by the Photofit system with verbal recall of faces. Their sample of 36 participants worked from memory to provide a verbal description and construct a Photofit likeness of a target face. In the first of two evaluation tasks, each of the descriptions and Photofits was given to 10 judges to identify from an array of 24 faces. In the second evaluation task, 32 judges attempted to sort each description and Photofit according to target. It was found that verbal descriptions given by witnesses were a more efficient means of conveying the impression of the target person than were the subsequent Photofit composites.

As has been mentioned before, Laughery and Fowler (1980) made a comparison between the Identikit system and sketches produced by sketch artists. The experiment revealed that the sketches were of far better quality than composites produced by the Identikit system. Upon examination of the results of their experiment, Laughery and Fowler (1980) started advising law enforcement agencies to choose sketch artists over manual composite systems if they had the choice, as they had good reason to believe that the targets were not too difficult to represent visually.

It can be seen, from the evaluation that these two manual systems have been subjected to, that neither of them is particularly good for achieving a close likeness to a target face. Instead, the composites produced by these systems bear a general resemblance to whoever is being recalled, and therefore they would be of little value to the police (Ellis & Shepherd, 1992; Laughery & Fowler, 1980). Having said that, it should be pointed out that composites produced by these manual systems can serve to eliminate unlikely candidates in mug files or lineups prior to exposing them to witnesses (Laughery & Fowler, 1980). To illustrate this, one can refer to a study, run by Darnborough (1977 in Clifford & Davies, 1989), which asked police officers how valuable they thought Photofits were in solving a total of 140 crimes. The Photofit was found to be "entirely responsible" in only 5% of cases, and "very useful" in 17% of cases. Another study by Bennett (1986) showed that the Photofit was "of assistance to the investigating officer" in less than 3% of reported cases. This emphasises the idea that Photofits are of no great value without the use of supporting evidence.

Based on the above, some may question the purpose of a visual representation of the perpetrator. However, it has been found that recognition performance can be extremely high if the target is known to the person carrying out the identification (e.g. Bruce, 1988; Burton, Wilson, Cowan & Bruce, 1999; Hancock, Bruce & Burton, 2000; Koehn & Fisher, 1997) and is therefore considerably higher than that obtained via a verbal description (Christie & Ellis, 1981). Another factor which stresses the importance of a visual representation is that the visual memory of a face does not decay as fast and is known to be less prone to interference than a verbal memory (e.g. Davies, 1983; Davies, Ellis & Shepherd, 1978; and Ellis, Shepherd & Davies, 1980).

According to Ainsworth (1998), facial recognition can perhaps best be seen as several processes rather than just one. On encountering a face, one is required to decide at least three things, i.e. (1) whether the face is familiar or not, (2) in what way the face is familiar, and (3) the name of the person whom he or she recognises as familiar. Bruce and Young (1986) claim that a familiar face triggers a "face recognition unit". Sometimes, "recognition" does not proceed any further than this and the only thing that the person knows is that the face is one which seems familiar. The majority of times, however, recognising the face as familiar may also trigger "person identity nodes", which indicate how one person knows another, and also where he or she has encountered the other previously. Following these steps, the person may be able to put a name to the face (Bruce & Young, 1986).

Authors of experiments evaluating the effectiveness of manual composite systems have attributed their poor results to limitations of these systems, the first one being the limited number of features available. Hairstyle is a good example to use to illustrate this limitation. Although there is a good variation in hair colouring and style, there are only 130 different hairstyles in the Identikit system and 204 in the original Photofit system. It has been found that hairstyle plays a significant role in the recognition of unfamiliar faces (e.g. Ellis, 1986) and therefore it seems that it would help to have many more hairstyles added to these systems.

Another limitation relates to the lack of deterioration in Photofit performance when there is a long delay between target exposure and composite production. While Davies et al. (1978) found that their participants' recognition ability to a target face declined significantly after three weeks, they could not find a significant difference in the rating scores between Photofit composites made immediately (after a 15-second exposure) or after three weeks. They linked this finding to restraints in the Photofit system. Similarly, McNeil, Wray, Hibler, Foster, Rhyne and Thibault (1987) found no significant decline in performance with Identikit composites constructed after three weeks.

A further limitation of these manual systems is their inability to represent faces that are atypical. There have been a number of studies that have shown that distinctive faces (i.e. a face with uncommon- or unusual-looking facial features) are better remembered than average-looking ones (e.g. Hancock, Burton & Bruce, 1996; Light, Kayra-Stuart & Hollander, 1979; Shapiro & Penrod, 1986; and Valentine & Endo, 1992). However Green and Geiselman's (1989) experiment revealed results contrary to this finding. They did not find higher quality Identikits of distinctive faces, i.e. composites of distinctive faces were not better identified than those constructed from more typical faces. Green and Geiselman (1989) saw this as a deficiency in the Identikit system.

Some researchers have attempted to overcome the afore-mentioned limitations by improving the quality of composites produced by manual systems. Ellis, Davies and Shepherd (1978b), for example, found that the feature demarcation lines present in Photofit constructions reduced the identification rate in recognition tasks, and thus interfered with face processing. Therefore it was thought that the removal of these lines would serve to improve the quality of Photofit constructions. In another study, Gibling and Bennett (1994) asked experienced operators to produce Photofit constructions whilst the target faces were in full view. Some Photofits were further enhanced by removing feature demarcation lines, and adding elaborative detail. Independent judges were then asked to pick out the targets from a 12-person lineup. A hit rate of 54% was found for the enhanced composites, which was significantly higher than the 15% found for those that were not enhanced. These studies show that with the aid of enhancement techniques, there is perhaps potential for manual systems.

2.4. Computer-Based Facial Composite Systems

In the 1980's, the cost of computer hardware dropped considerably, allowing police forces to use them widely for the first time. During this period a number of computer-based facial composite systems entered the market in the hope that they would answer criticisms made of the original manual systems (Brignull, 1999).

Computer-driven systems for constructing composite faces of suspects have largely replaced manual systems in police use. Potentially, they offer clear advantages over manual systems not only through a greatly expanded range of facial features but also through the facility to manipulate these features with the aid of graphic packages. At the click of a mouse, face shape can be altered, partings can be reversed, and hair lengthened, shortened, lightened or darkened (Davies et al, 2000). In addition, computer-based facial composite systems allow operators to effortlessly move features independently of one another, to adjust the dimensions of the face and its individual features, to age the face, and to add accessories (like hats, spectacles bandanas, etc.) to the face. Computer-based systems allow for the manipulation of the facial image in ways that previously would have required a professional sketch artist (Davies et al., 2000).

Manufacturers of the computer-based Mac-a-Mug Pro system boast that with the flexibility to move features independently of each other, their system has the capacity to make nearly 100 times the number of faces offered by manual systems (Davies et al., 2000). Clearly there is evidence to suggest that the introduction of computer-based facial composite systems has helped to make the process of facial composite reconstruction more sophisticated. However, questions needed to be raised about their efficacy as tools to produce accurate representations of faces of criminals.

Several computer-based facial composite systems have been developed since the computer revolution, i.e. Facette, Faces, Comphotofit, Pro-Fit (originally CD-Fit), Phantomas, Compusketch, Suspect-ID to name a few. However, many of these systems have not been thoroughly tested for effectiveness. Two computer-based facial composite systems that have been under particular scrutiny thus far are (1) the Mac-a-Mug Pro and (2) the E-Fit (Electronic Facial Identification Technique).

Cutler et al. (1988) conducted the first evaluation on the Mac-a-Mug Pro system. They had an experienced operator create composites of 10 target faces whilst viewing pictures of them throughout the process. One of the recognition tasks involved participants having to examine the resulting composites and then select the original photographs from a total of 60 distractors. The resultant 49% identification rate was seen as a positive for the potential of the Mac-a-Mug Pro system.

Wogalter and Marwitz (1991) performed the next evaluation of the Mac-a-Mug Pro system. They had 54 participants produce several composites following a brief 8-second exposure to the target. Thereafter, the same participants produced composites of the same targets in view. Ten independent judges evaluated the resulting composites through matching (composite to the target) and rating (for similarity) tasks; five judges per task. The "from memory" composites received an average matching score of 40%, which is not too far off from the matching scores achieved by "in view" composites compiled during Cutler et al.'s (1988) and Gibling and Bennett's (1994) studies. Frowd (2001) points out though that the matching scores in Wogalter and Marwitz's (1991) study were inflated due to the fact that there were only six faces in the target array, thus enabling participants to easily match a few facial features.

Two studies conducted by Kovera et al. (1997) revealed severe limitations with regard to the Mac-a-Mug Pro system and its effectiveness. In the first study, ten college students made "from memory" reconstructions of five faculty members and five students from their former high schools. Composites were created with the use of an experienced operator and there was an opportunity for photographic enhancement with the aid of the MacPaint package. It was found that other students who had attended the same high schools could not recognise ten composites of either students or faculty members when the composites of individuals known to them were mixed with composites of 40 strangers. In the second study, naive witnesses who viewed the composites were not able to select the people depicted in the composites from photo lineups comprising one target and four foils (Kovera et al., 1997).

Similar results were found in a study conducted by Koehn and Fisher (1997), where they had 62 witnesses observe a target person for a few minutes and then two days later either verbally describe or construct a facial composite of the target using the Mac-a-Mug Pro system. Prior to composite production, witnesses were put through either a guided memory or a standard police interview. The Mac-a-Mug Pro composites produced by witnesses, irrespective of which interview they underwent, were not at all similar to the target face and were of no value in selecting the target from a photo lineup. Participants could only identify 4% of the original faces from a 6-person target-present lineup. Koehn and Fisher (1997) have attributed this poor performance to the "from memory" conditions as they found that composites created in-view by the experienced operator resulted in a 77% recognition rate.

The studies evaluating the Mac-a-Mug Pro system indicate that the system is particularly ineffective when composites are created from memory which, as Frowd (2001) quite rightly points out, is the more forensically important condition. The same evaluation has been found with other systems, including CADC and E-fit. The CADC system is basically a digitised version of the manual Photofit system, except that it is a more advanced version. It overcomes certain limitations of the Photofit system like feature delimiting lines, and the limited number of faces that can possibly be produced with Photofit. Christie, Davies, Shepherd and Ellis (1981) made a comparison between the CADC and Photofit systems. Participants were asked to use one of the two systems to create a composite from memory after being exposed to the target for one minute, and then of another target in-view. A matching task with 18 distractors then ensued. A matching accuracy of 28% was found for "from memory" CADC composites, significantly better than 18% found for "from memory" Photofit composites. This finding provides hope for a computer-based system in a more ecologically valid situation. As Frowd (2001) asserts, no additional studies have been conducted on the CADC system, thus the reliability of these results is questionable.

Davies et al. (2000) performed another comparison between a manual and a computer-based composite system. They selected E-fit as the "computer-based" representative and Photofit as the "manual" representative. In their experiment, 24 participants were to construct two of four targets: one familiar and one unfamiliar. The Photofit system was to be used to construct the first face, and the E-fit system to construct the second. Targets were first reconstructed from memory; thereafter reconstructions were made with photographs of the target faces in full view during the reconstruction procedure. The resulting 96 composites were assessed separately for "target-absent" and "target-present" conditions. Three evaluation tasks were performed by 24 judges, all of who were in daily contact with the familiar targets. For the first task, composites were rated for familiarity on a 5-point scale. For the second task, judges attempted to name facial composites that they thought they could identify. For the third and final task, judges were required to sort the facial composites according to all eight targets (Davies et al., 2000).

The thorough assessment of the E-fit and Photofit composites revealed that E-fit was consistently superior when a familiar face was constructed in the presence of photographs (83% matching accuracy). Not only was E-fit more accurate under optimal conditions, it was also more rapid in reproducing faces (Davies et al., 2000). As Frowd (2001) says, this particular finding is of little forensic relevance because in a real case, composites are created of unfamiliar faces from memory. When participants worked from memory, neither system was better than the other. The average naming rate of composites created from memory was a mere 17%. Matching accuracy was 63% from memory, though this was carried out without distractors (Davies et al., 2000). Frowd (2001) notes that compared against Wogalter and Marwitz (1991), who also performed matching without distractors, the E-fit system was found to be about 20% better on this measure than the Mac-a-Mug Pro system. This implies that E-fit is more effective than the Mac-a-Mug Pro system.

Koehn (1995) was the first to make a direct comparison between these two systems. The E-fit and Mac-a-Mug Pro systems are similar to each other in that facial features can be selected, resized and manipulated until a satisfactory likeness has been obtained. The key difference between these two compositors is that E-fit uses more sophisticated methods of composite reconstruction (Koehn, 1995). With the E-fit system, one first inputs the facial description by completing as many description fields as possible. E-fit then filters all the possible matches between the description given and the pre-set descriptors for each of the features in the database. It puts the most likely features first and a set of features is built up for the next stage. Where no description is given, an "average" default feature is used. An image is then displayed which will then be the starting point for the witness to mould the face. Additional features can be added from the computer database (such as glasses, hats, scars, etc). In the final "finishing touches" stage, the operator is taken into a freehand drawing programme where any individual feature, or the image as a whole, can be amended. Clark (2001) points out that the effectiveness of this stage is particularly dependant on the proficiency and skill of the operator.

Since viewing features in isolation has been found to be a more error-prone procedure than viewing features in the context of a face (Davies & Christie, 1982), the designers of E-fit wanted to move away from earlier "first generation" systems, which required witnesses to select the features of the face they were recalling from a printed array of isolated features. This is why E-Fit was designed to always present features in the context of a complete face. Like other computer-based systems, E-fit makes up for the inability of the Photofit and Identikit systems to make adjustments to the configuration of features once selected, which has been said to be one of the principal handicaps of these two manual systems. The facility in E-Fit to move a feature independently of any other feature overcomes this problem. The use of the verbal description given by the witness to select a small range of appropriate features in E-fit reduces the search task that confronts the witness in comparison with using a purely visual index. In addition, a variety of amendments can be carried out on any E-fit feature (Ellis & Shepherd, 1992).

As can be seen, E-Fit was designed to make use of existing knowledge from research into face perception (Turner, Bennett, Pike, Towell & Kemp, 2000). The manufacturers of E-Fit believe that their system is unique since it recognises, and was developed to incorporate the psychological factors that affect human ability to recall and recognise faces. In addition, it was designed to allow novice operators to easily and quickly produce accurate images (Aspley Ltd., 2000).

Koehn (1995) hypothesised that the sophistication of the E-fit compositor would allow it to construct better quality composites than the Mac-a-Mug Pro system. On day 1 of the reconstruction phase of her experiment, participants viewed a 2-minute videotaped lecture on cognitive labels and emotion. They were told to pay close attention to the video-clip, as they would be tested on the material the following day. On day 2, participants returned individually and were told the true reason for exposing them to the video-clip. Some participants were asked to produce a verbal description of the lecturer with enough detail that someone reading the verbal description would be able to pick out the target from a photo lineup. Other participants were asked to construct either an E-fit or a Mac-a-Mug Pro composite of the lecturer, with the help of a research assistant who operated the facial composite systems.

In an evaluation task, evaluators attempted to identify the target from a 6-person lineup. This was done in one of three ways, i.e. (1) whilst viewing an E-Fit composite along with a verbal description, (2) viewing a Mac-a-Mug Pro composite along with a verbal description, or (3) viewing a verbal description alone. In another evaluation task, the composites were rated on a 10-point rating scale according to similarity to the target face. The two evaluation tasks revealed that even though E-Fit composites were found to be of better quality than Mac-a-Mug Pro composites, composites from both systems were found to be useless for identifying the target person from a photo lineup. Further, verbal descriptions were found to be more useful than both E-Fit and Mac-a-Mug Pro composites (Koehn, 1995).

In 1999, another study involving the E-fit system was conducted, which explored the effect of (positive or negative) attitude towards a target on future recognition (Davies & Oldman, 1999). Participants were asked to produce E-fits from memory of four famous faces that they either strongly liked or disliked. Participants then produced E-fits with these faces in full view. The resulting average naming rate was found to be an incredibly low 6% for "from memory" constructions and 10% for "in view" constructions. Frowd (2001) reasons that these floor level results are attributable to a combination of the use of a few targets, and not knowing how many of the original targets were recognisable to the participants.

A year later, Brace, Pike and Kemp (2000) conducted another E-fit study, involving famous faces as targets. Composites were created either by the operator alone (from memory and in-view) or by the normal interaction process involving another person (a "describer"). A 10% increase in naming rate was found for composites created by an operator alone. An average recognition rate of 25% was found for composites constructed via the "describer". It was not possible to extract the recognition rate of composites constructed from memory, but the afore-mentioned average recognition rate is comparable with that found by Davies et al. (2000).

As have been said before, many computer-based facial composite systems on the market have not been thoroughly tested for effectiveness. Apart from the E-fit and Mac-a-Mug Pro systems, the Whatsisface system has been evaluated (Gillenson & Chandrasekeran, 1975). It is a computer-based system with which a non-artist can create sketch-like composites of Caucasian males. It contains pre-stored facial features, an average face that serves as a starting point, and a heuristic strategy that guides the witness through a process of modifying the facial image to produce the desired outcome. In their study, 60 composites were created with the target in full view during reconstruction. These composites were used to select the original photographs (with no distractors), and revealed an 81% matching accuracy, which is quite similar to that found by Davies et al.'s (2000) for E-fits constructed with the target in view. Since there is no other data available, Frowd (2001) believes that it would be safe to assume that Whatsisface is no more effective than E-fit.

Facial composite systems have developed over time, and many designers of these systems have drawn from and based their systems upon knowledge of the ever-expanding research into face perception. As has been explained previously, most composite systems – whether manual or computerised – have been based on the traditional feature-based approach, where the facial composite is put together using pre-defined parts of faces. EvoFIT is a system currently being developed at Stirling Universty and is an example of a facial composite system that has moved away from this traditional feature-based approach to a more holistic-based approach (Frowd, 2001).

The motivation behind the development of EvoFIT was based on the poor performance of manual and computerised composite systems. Frowd (2001) explains that part of the problem with these systems is that they involve a single facial representation that necessitates a verbal interaction. He believes that with a composite system such as EvoFIT, which presents the witness with 18 whole faces seen at the same time (containing random eyes, noses, mouths, etc.), one capitalises on our innate ability to recognise faces and one simultaneously reduces the potentially disruptive effect of the verbal component. The witness's task would be to make a selection out of those multiple faces according to which they think represents the face of the perpetrator. Typically six faces are selected. A complex mathematical approach, known as Genetic Algorithms, begins with a set of random solutions to a problem. The best solutions are selected and "bred" together to produce a new set of 18 faces that more closely resemble the target. This process continues until an acceptable likeness is reached (Frowd, 2001).

As part of the EvoFIT, a utility known as the "feature shifter" can be used to move and resize facial features, e.g. to position the eyes closer together. Other software tools used to enhance the resulting EvoFIT composite include an interface to other computer-based composite systems like E-fit and CD-Fit, which allows a range of hairstyles to be selected; and software which allows additional facial refinements and scars to be added in a number of photographic editing programmes like Adobe Photoshop (Frowd, 2001).

Frowd (2001) compared the effectiveness of the EvoFIT system against the E-fit system. In his experiment, 60 participants were permitted to inspect a photograph of a famous face for one minute. They were then asked to provide a verbal description of the famous face (comprising two cycles of free recall followed by cued recall). Thereafter, they were asked to create a composite of this famous face using either one of the two systems. In a subsequent naming task, a spontaneous naming rate of about 10% was found for the EvoFIT composites and about 17% for the E-fit composites. Frowd (2001) attributed the poor performance of EvoFIT to the upper age limit of its database, which was shown to be around 30 years. This was found to be significantly less that the mean age of the targets in the experiments, which was 47 years. It is interesting to point out that Frowd's (2001) naming rate of about 17% for E-fit composites is almost identical to that found by Davies et al. (2000).

To investigate this further, Frowd (2001) did a follow-up study using more appropriately aged famous faces as targets (10 targets with an average age of roughly 27 years). Spontaneous naming rates of 25% for novice operators of EvoFIT were found for composites created with the target in view. The study proves that the EvoFIT system has the potential to produce recognisable composites when the appropriated aged targets are used. Frowd (2001) concedes that he did not use an ecologically valid method of construction. He is however quite pleased that the same naming rate of 25% was found by Brace et al. (2000) for E-fits of famous faces created in-view.

From the studies outlined above investigating the effectiveness of computer-based facial composite systems, one can deduce that although computer-based systems are more sophisticated versions of manual ones, they do not succeed in generating accurate, recognisable composites of target faces under the forensically important "from memory" conditions. As Davies and his colleagues warn, "until more encouraging findings emerge on the quality of likeness achieved under more realistic memory conditions, such images deserve to be treated with caution by police and public alike" (Davies et al., 2000, p.124). One can also gather from the results of the above experiments that computer-based systems perform just as well as manual systems. It could therefore be argued that the computerisation of the older manual composite systems was of no real benefit to law enforcement agencies, as it did not serve to improve the quality of representations of criminals.

Chapter 3: Face Processing Theories

We have established that both manual and computer-based facial composite systems have thus far failed to perform well under evaluation. Why is this the case?

A plausible speculation made by Barber (1988) to explain the poor performance of facial composite systems is that the range of facial features available in these systems, which although quite extensive, is not wide enough. This therefore hinders witnesses in compiling a facial composite with the precise hairstyles, eyes, nose, mouth, or whatever they are looking for (Barber, 1988).

Another significant argument is that the ineffectiveness of manual and computer-based systems up until now is a result of the fundamental concept underlying them, that being that they are dependent on the featural composition of the face (Kovera et al., 1997). This in turn has served to explain why there is no evidence that computer-based systems lead to better reconstructions than manual systems (Davies et al., 2000; Green and Geiselman, 1989; Kovera et al., 1997).

The argument is that if memory for a face principally involves an integrated representation of the face, then it does not make sense to be heavily dependent on methods of composite construction that analytically consider features in isolation from the global configurations of the face as this is likely to interfere with the memory of the face being constructed (Barber, 1988; Bruce, 1988; Penry, 1971). Therefore it has been said that there exists a flawed theoretical basis upon which most existing composite systems are designed (Koehn & Fisher, 1997; Kovera et al., 1997). As Davies and Christie (1982, p.103) point out, "...judging resemblance from features seen in isolation may be a serious source of distortion in composite production".

Three face processing theories are reviewed here, i.e. (1) the "holistic / configural", (2) the "featural", and (3) the "dual".

3.1. The "Holistic / Configural" Face Processing Theory

There is an abundance of evidence to suggest that individual features are not processed independently from other face information such as the eyes, nose and mouth. In an interesting experiment conducted by Young, Hellawell and Hay (1987), well known faces were divided horizontally into upper and lower halves. Young et al. (1987) requested their participants to make recognition judgments as fast as they possibly could, based solely on the top half of each face presented. Although participants were quite accurate at identifying the isolated top half of the face when it was seen on its own or when the two halves were off-set so that the face outline of the composite was disrupted, when it was combined and properly aligned with the wrong lower half it became extremely difficult to recognise to whom the upper features belonged (Young et al., 1987).

Hole (1994) used a similar procedure in his study to show that the finding made by Young et al. (1987) can be applied to unfamiliar as well as to familiar faces. He explains these results by asserting that, "if a photograph of the top half of one face is aligned with the photograph of the bottom half of a different face, the two halves perceptually "fuse" to produce a strong impression of a complete novel face" (Hole, 1994, p.65).

Tanaka and Farah (1993) provide further evidence to support the claim that faces are processed holistically without separate explicit representation of parts. In their study, they requested their participants to learn the names of six computer-generated faces. Immediately following this learning phase, they were asked to identify from a set of foils the eyes, nose and mouth of each face. Each feature was tested twice, i.e. (1) the "isolated part" test and (2) the "full face" test. In the "isolated part" test, a previously seen feature and a foil feature were presented in isolation (not in the context of a face). The participants would then be asked to identify the feature that belonged to a specific target, e.g. "Which is Larry's nose?" In the "full face" test, the original and foil features were presented in the context of the original face. Participants would then be asked a question like "Which is Larry?" The results of Tanaka and Farah's (1993) study indicate better memory (over 10% higher) for features in their facial contexts than in isolation, which has been taken as further evidence for the configural processing of faces.

Davies and Christie (1982) made a similar finding. In their study, participants rated whether a mouth and a pair of eyes was present in a target. These ratings were made both in isolation and within the context of a face. It was discovered that people rate features for similarity differently when they are seen in isolation than when they are seen in context of the whole face. Davies and Christie (1982) explained their results by asserting that faces are stored in an integrated form, from which it is difficult to extract feature information.

Additional criticism of the featural face processing theory can be drawn from studies conducted by Purcell and Stewart (1986) and Homa, Haver and Schwartz (1976). Purcell and Stewart (1986) found, in their study, that briefly flashed pictures of a face were detected more accurately than was a control pattern with a nose, a mouth, and a pair of eyes positioned arbitrarily so that they did not form a face. Homa et al. (1976) presented their participants with a picture of a face or with the same features rearranged so that they were no longer seen as a face. Subjects were better able to identify a feature on a subsequent trial if it had been seen first as part of a face rather than as part of an arbitrarily arranged pattern of features.

The important role of configural information has been highlighted in a study carried through by Haig (1984). He directly investigated people's sensitivity to detect changes in spatial placement of facial features. He used inward and outward movements of the eyes (compared to an unaltered stimulus) and found that participants quickly noticed changes to the eye region with unfamiliar faces. Haig (1984) notes that in some cases, a small movement of the mouth or eyes within a face can create the impression of a completely new individual. Hosie, Ellis and Haig (1988) reported similar findings using famous faces.

More support for configural processing comes from Thompson's (1980) intriguing "Thatcher Illusion". Thompson (1980) took a picture of Margaret Thatcher, and cut out and inverted the eyes and mouth within the face. When viewed upright the result looked grotesque. However, when turned upside down, the picture looked quite normal, pleasant and very similar to the original version of the picture – hence the "Thatcher Illusion". Bartlett and Searcy (1993) explains this illusion by claiming that the components of an inverted face are processed independently of one another, and that the grotesque appearance of the upright Margaret Thatcher face arises because of the relationship between the arrangements of the eyes and of other parts of the face – which can only be seen with the configural processing mode available for upright faces.

Upon examination of the results of the studies mentioned above, it could not be disputed that there is considerable support for the configural face processing theory. It has been demonstrated how perception of one feature can be influenced by others, showing that the facial pattern is processed as more than the sum of its independent parts. In other words, it has been asserted that faces are not perceived as a collection of individual features, but rather as an integrated whole.

Therefore one can see where experimenters like Baddeley and Woodheard (1983) are coming from when they argue that a detailed analysis of faces into component features is likely to be unhelpful. They believe that in looking at people, it is perhaps the way in which the individual features relate to each other that creates the impression of a particular person. Attempting to go against this "natural" way of looking at faces is not likely to be productive, according to Baddeley and Woodhead (1983).

3.2. The "Featural" Face Processing Theory

Despite the fact that there has been ample support for configural face processing, there has been research that suggests that featural processing is just as important as configural processing. To discover whether or not faces are perceived featurally, Bradshaw and Wallace (1971) conducted a study in which they put their participants through a matching task. During this matching task, pairs of simultaneously presented Identikit faces were to be judged "same" or "different". It was found that as the number of feature differences between two faces increased, the quicker the faces were judged to be different – which in turn led Bradshaw and Wallace (1971) to make the argument that faces are perceived in terms of their features. A year prior to Bradshaw and Wallace's (1971) study, Smith and Nielson (1970) used schematic faces instead of Identikit faces and reported similar results.

Roughly a decade ago, Tanaka and Farah (1993) discovered that individual features could be recognised above chance level when presented in isolation or in the context of a jumbled face. Bruyer and Coget (1987) noticed the same in their study. This serves as further support for featural face processing.

Leder and Bruce (2000) challenge the importance of facial context in face recognition. They argue that isolated configural information plays an important role in face recognition, and that it's processing does not require the context of the whole face. In their study, it was found that participants learned to recognise faces that differed from each other only in respect of one individual configural element, such as the distance between nose and mouth. When these features were presented without any facial context, participants scored recognition rates above 59%. These results support the idea that relational information is processed in a local and possibly independent way. Therefore it could be argued that the relational notion of configuration provides the best account of the critical information used to recognise faces (Leder & Bruce, 2000).

The results of studies reported above provide evidence for featural face processing. During the course of the investigation into featural face processing, some researchers discovered that certain features of the face were more salient than others. Ellis et al. (1975) noted that changes to features in the upper half of the face (i.e. hair and eyes) are recognised better than changes to features in the lower half of the face (i.e. nose, mouth and chin). In an experiment, they constructed two faces using randomly selected Photofit features, presented them to participants for ten seconds each, and then had them reconstruct each face feature by feature using the Photofit system. Although the representations were not all that accurate, higher recall accuracy resulted from forehead and mouth on one of the faces, and for forehead and eyes on the other. Additional evidence for the memory advantage of upper-face features can be found other similar studies (e.g. Davies & Christie, 1982; Goldstein & Mackenberg, 1966; Laughery, Alexander & Lane, 1971; Laughery & Fowler, 1980; McKelvie, 1976; Patterson & Baddeley, 1977; and Walker-Smith, 1978).

Some researchers reasoned that if upper-face features are more important than lower-face features, then perhaps all facial features could be ranked into a feature hierarchy according to their importance in recognition (Reynolds and Pezdek, 1992). However differing results were revealed from various proposed hierarchies of facial feature saliency (e.g. Carey & Diamond, 1977; Friedman, Reed & Carterette, 1971; and McKelvie, 1973). Yet a review by Shepherd, Davies and Ellis (1981) revealed the relative importance of hair, face outline and eyes in face perception and memory, and the relative unimportance of the lower internal features, particularly the nose. They found that different features of the face are perceived and remembered to different degrees, where the general order of saliency is forehead, then eyes, then nose, then mouth, then chin.

Fraser and Parker (1986) examined feature saliency by showing their participants, in rapid sequence, each of the components of a line-drawn face, i.e. outline of the face, eyes, nose and mouth. The participants' task was to detect whether all the components of a face were present. On some trials, one of the components was missing. It was found that participants were best at noticing the absence of the face outline, followed by the eyes, then the mouth, then lastly the nose. Haig (1984; 1986) confirmed these findings, using a computer system that allowed features from one face to be stretched, compressed, deleted or transferred to another face. Bruce (1988) points out that a possible explanation for the relative unimportance of the nose in these studies is the difficulty of its perception in a full-face portrait. She argues that a nose can be perceived better in an angled shot than in a full face.

Besides upper-face versus lower-face features, another combination of facial features has been proposed, i.e. inner-face features (i.e. eyes, nose and mouth) versus outer-face features (i.e. hair and outline). The importance of these categories of features seems to be dependent on familiarity. Ellis, Shepherd and Davies (1979) found that inner-face features are more important than outer-face features in recognising famous faces, while unfamiliar faces are recognised using either inner-face or outer-face features. Endo, Takahashi and Maruyama (1984) replicated this finding by using Japanese faces and subjects. In 1999, Campbell, Coleman, Walker, Benson, Wallace, Michelotti and Baron-Cohen used a blurring technique to defocus different parts of the face image systematically, and confirmed the inner-face advantage for viewing famous faces. They explain that a possible reason for the inner-face advantage is that the inner features capture commonality of different face-views (e.g. full, three-quarter profile and side profile) of an individual better than the outer parts. They add that the inner-face advantage favours familiar faces partly because they have usually been seen and learned under a variety of viewing conditions (Campbell et al., 1999).

In terms of the order of processing in face recognition, some experiments have indicated that facial features are recognised using a top-to-bottom order of processing. For example, in Walker-Smith, Gale and Findley's (1977) experiment, eye movements were monitored while facial pictures were being viewed. A top-to-bottom order of processing was found. Walker-Smith (1978) also found this result when she varied a single facial feature on a same-different matching task. She found that both reaction time and error rates increased in a top-to-bottom direction, with the lowest reaction times and error rates when the facial stimuli differed on the hair and forehead, and increasing reaction times and error rates for the eyes, nose, mouth and chin. Using a same-different matching task, Hines and Braun (1990) examined the order of face recognition for line drawings of familiar and unfamiliar faces. They found a significant top-to-bottom order of processing for unfamiliar faces but not for familiar faces, the processing of which may involve a parallel process, according to Hines and Braun (1990).

3.3. The "Dual" Face Processing Theory

There is evidence for a dual processing strategy of face recognition, i.e. that faces are perceived both featurally and configurally. Matthews (1978) was one of the first researchers to arrive at this conclusion. In a study with Identikit faces in a simultaneous matching task, he found that participants' reaction times increased linearly for detecting differences in eyebrows, nose and mouth features, thus indicating a top-to-bottom serial comparison approach. He also found that reaction times for detecting changes in hair, eyes and chin were virtually the same across features, thus indicating a holistic or parallel comparison approach. Matthews (1978) interpreted from these results that features are checked both in serial and in parallel.

Although the eyes are not part of the facial outline, Matthews (1978) believes that eyes are initially processed in a Gestalt (i.e. overall, holistic) manner. According to this argument, the eyes are where an observer initially fixates to examine the overall facial shape established by the two other peripheral features. Matthews' (1978) results are in line with previous research asserting that upper-face features are more important in face memory than lower-face features.

In 2000, Collishaw and Hole (2000) conducted a study in which they explored the relationship between these two modes of processing, i.e. featural and configural. They presented participants with blurred, scrambled and inverted faces – both singly and in various combinations. Unfamiliar faces (seen once before) and highly familiar faces (celebrities) were presented to participants for recognition. They noticed that either featural or configural processing was disrupted when faces were blurred, scrambled, inverted, or simultaneously scrambled and inverted, and that the mode of processing which was not disrupted was the one used as a route to recognition.

It was further discovered that faces were recognised, above chance level, by the mode of processing that was not disrupted. Blurred/inverted and blurred/scrambled faces were however recognised at or near chance levels. Collishaw and Hole (2000) suspect that the reason for this is that both configural and featural modes of processing gets disrupted in this instance. Hence Collishaw and Hole (2000) emphasise that a distinction should be made between featural and configural processing. Their study supports a "two-process" theory of face recognition where featural processing and configural processing are believed to be dissociable routes to the identification of a face. Comparison of the results for famous and unfamiliar face recognition tasks suggests that these findings are likely to be generalisable across the spectrum of familiarity (Collishaw & Hole, 2000).

More support for the "dual" face processing argument comes from Sergent (1984). She used a matching task with Photofit faces and found that participants detected changes in chin contours and "internal spacings" (i.e. distance between the nose and mouth) faster than they detected changes in chin contours only. Sergent (1984) thus concluded that some features (e.g. eyes and chin contour) seemed to be processed independently of each other whereas other features (e.g. chin contour and "internal spacings") interact and are processed more holistically.

Tanaka and Sengco (1997) conducted four experiments in a study aimed at examining the effect of configuration on feature recognition. They created two configurations of a face, one with eyes close together and one with eyes far apart. After participants studied faces presented in one of the two configurations (i.e. "eyes-close" or "eyes-far"), they were tested for their recognition of features shown in isolation, in a new face configuration and in the old face configuration. It was found that participants recognised features best when presented in the old configuration, next best in the new configuration, and poorest in isolation. In addition, it was found that altering the spatial location of the eyes not only impaired participants' recognition of the eye features but also impaired their recognition of the unaltered nose and mouth features. Tanaka and Sengco (1997) thus concluded that there appears to be an interdependency of featural and configural face information.

In an effort to measure contributions of featural and configural processing to face recognition, Cabeza and Kato (2000) constructed "configural prototypical faces" and "featural prototypical faces" based upon four other faces. To form a "configural prototype" of these faces, the shape and skin colour information was morphed into one composite. To form a "feature prototype" for the four faces, a single composite face was made out of the outline and cheeks of the first face, the eyes and eyebrows from the second face, the nose of the third face, and the mouth of the fourth face. Cabezo and Kato (2000) found that participants falsely recognised non-studied prototypes, and this tendency was equivalent for both featural and configural prototypes. Moreover, participants that viewed inverted faces committed fewer false alarms to configural prototypes than to featural prototypes. On the basis of these results, Cabezo and Kato (2000) find support for the dual-code view that both featural and configural processing make important contributions to face recognition.

The outcomes of the above studies suggest that the face is not perceived strictly and exclusively as features or as a whole, but that there appears to be a link between the two modes of processing. Bruce (1988, p.39) purports that "an initial holistic (or configural) processing may occur so that overall information about a face can be obtained quickly. This holistic processing then directs the subsequent featural processing of more detailed aspects of a face, and the two processes continue in parallel to yield more information about a face".

A review by Farah, Wilson, Drain and Tanaka (1998) revealed that there is general agreement that face recognition involves processing information about individual facial features as well as processing information about the spatial layout or configuration of these features.

Chapter 4: The Two Facial Composite Systems under Evaluation

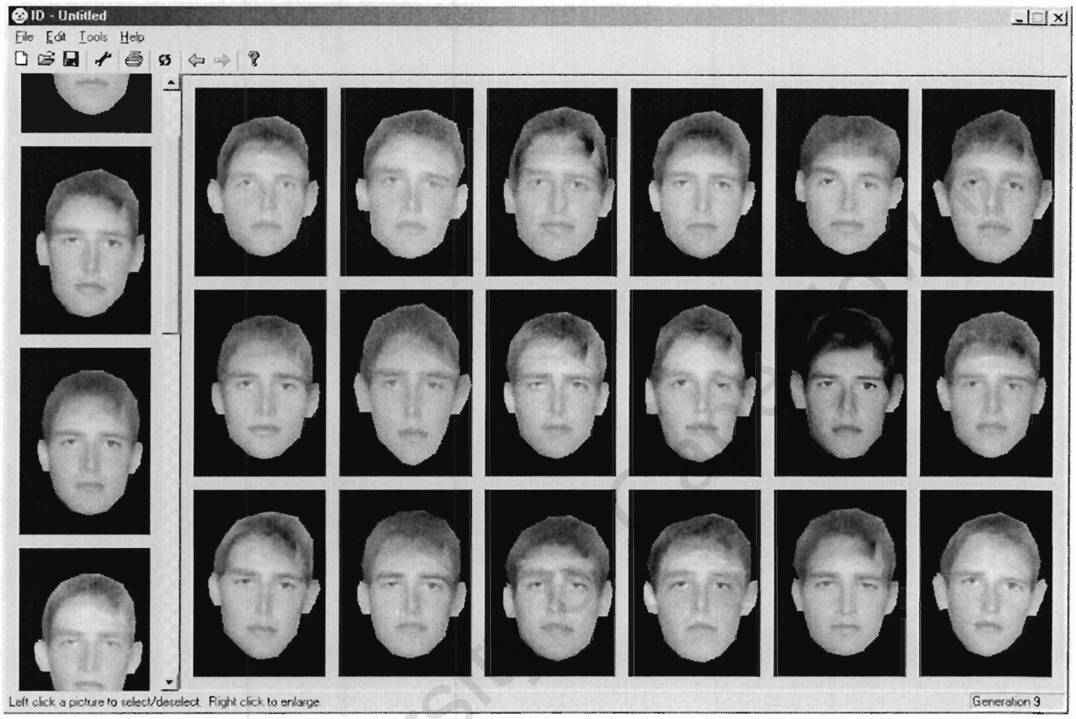
Two facial composite systems are evaluated in this study, i.e. (1) E-Face and (2) Faces. The E-Face system is an eigenface-based system that has been developed at the University of Cape Town. It has been through several versions and is still in the process of refinement. The other facial composite system under evaluation, the Faces system, has been fairly widely used worldwide, particularly in the United States. In 2000, a survey conducted by McQuiston and Malpass (2000) in the United States of 163 law enforcement agencies revealed that of those agencies that use computer-based composite systems, 46% use the Faces system. This figure cannot be generalised to all law enforcement agencies in the United States since there was only a 10% return rate on the surveys distributed to 1637 law enforcement agencies in the United States (McQuiston & Malpass, 2000). Nevertheless, one can get an idea of how many law enforcement agencies in the United States use one of the two composite systems that will be examined in this study.

Although one of the two composite systems under evaluation is quite well represented in terms of utilisation by law enforcement agencies in United States, they are not at all represented in South African law enforcement agencies. According to Van Dijk (personal communication, October 1, 2001), the SAPS have used two systems, i.e. (1) Identikit-2000 and (2) Comphotofit.

A fundamental difference between the two composite systems under evaluation is the theoretical bases upon which they have been designed. Faces v.3 is a featural composite system, which means that composites are produced by having witnesses browse through a library of various facial features (e.g. eyes, nose, chin, etc.) and then have them individually select the most appropriate ones to make up the face. E-Face, on the other hand, is a holistic system that permits configural reconstruction of faces. It constructs a facial image without paying attention to any particular feature, through the use of a linear combination of eigenfaces, i.e. a set of standardised face-like images (Rosenthal, de Jager & Green, 1998; Tredoux, Rosenthal, da Costa & Nunez, 1999).

4.1. E-Face

Figure 1: The E-Face system



With the E-Face system, the face recall process is started with the witness or victim of the crime being presented with a random sample of 18 faces corresponding with the sex, race and approximate age group of the offender (as specified by the witness). The number of faces presented to the witness in each generation can be altered, but was set to 18 faces for this study. As can be seen in Figure 1 above, E-Face is in full colour. Please note that all other copies of images produced by E-Face, which are not printed in colour in this thesis, are in fact colour images. At present, the E-Face system comprises five databases, i.e. (1) black, (2) white, (3) coloured (mixed race), (4) Indian and (5) Indian/coloured. The age range of these databases is 18 to 21. In addition to having limited ability to produce facial composites outside this age range, E-Face does not yet have the ability to produce female facial composites, as the databases compiled thus far comprise only male faces.

Either one of two optimisation techniques can be used to operate E-Face, i.e. (1) Population Based Incremental Learning (PBIL) or (2) M-Choice (an algorithm devised by Tredoux in 2000). With PBIL, the optimisation technique selected for this study, the witness uses a mouse to click on one face out of the 18 that looks most similar to that of the offender. A new set of 18 faces is then generated on the screen that looks more similar to the previously selected face, whilst still ensuring that the spread of different-looking faces is still wide. This process is repeated with each generation resulting in the following set of faces looking more and more similar. Once the witness is satisfied that one of the faces is a "good enough" likeness to the perpetrator, the process is stopped and the face can be printed onto paper or saved as a digital/computer image.

The more recently-developed M-Choice algorithm works in much the same way as PBIL. The key difference is that with the M-Choice algorithm, the witness is not restricted to selecting just one face out of the initial array of 18 faces and/or during subsequent generations. Instead, the witness can select any face or number of faces which they believe resemble the perpetrator's face. Once the selections are made, a new set of faces is generated which look more similar to the previously selected faces, whilst still ensuring a good spread of different-looking faces. This process is repeated until the witness finally finds a face that represents a "good enough" likeness to the offender. As with PBIL, the finally chosen face can be printed onto paper or saved as a digital/computer image.

The E-Face system allows witnesses to go backwards (and forwards) whilst progressing through the task in order to view previous generations. A history of each face selected in each generation is stored on the left-hand side of the screen, which can be viewed by witnesses throughout the reconstruction process. Witnesses are advised to refer to this history before making their final choice as to which image represents the best possible likeness to the offender.

Essentially, the E-Face system operates on the principle of "selection of the fittest" from each generation of faces, until such point when the witness cannot identify a "more fit" face. E-Face is capable of efficiently searching an extremely large sample space of alternative faces, and finding an accurate likeness in a relatively short period of time. The computational basis upon which E-Face has been designed is Principal Component Analysis (PCA). The main idea is to decompose face images into a small set of characteristic feature images called eigenfaces, which may be thought of as the principal components of the original images. These eigenfaces function as the orthogonal basis vectors of a linear subspace called "face space". Recognition is performed by projecting a new face image into this face space, and then comparing its position in the face space with those of known faces (Hancock & Frowd, 2002). A detailed and technical description of this computational basis can be found in Kirby and Sirovich (1990), and Turk and Pentland (1991).

The eigenface approach to face recognition was motivated by information theory, leading to the idea of basing face recognition on a small set of image features that best approximates the set of known face images, without requiring that they correspond to our intuitive notions of facial parts and features. If eigenfaces were recombined randomly, then one would be able to generate a face-like image not necessarily resembling any of those in the original set. This approach offers a means of producing novel faces in a holistic fashion, and provides a practical solution that is well fitted to the problem of face recognition (Hancock & Frowd, 2002; Turk & Pentland, 1991).

It was thought a good idea to introduce the eigenface approach due to the inefficiency of the linear search, as used by featural systems (which make up the majority of existing facial composite systems). With featural systems, the need for the recognition of isolated features is compounded with the need for further recall of both features and feature positions to produce a final composite (Caldwell & Johnston, 1991). Operating on only one face, which is current practice of building composites with the majority of existing facial composite systems, means that a witness must continuously describe changes necessary to make the composite better. Hancock and Frowd (2002) believe that this "recall" process is a naturally hard task for anyone to do. Recognising a face, on the other hand, tends to be both easy and accurate. Even if a face is recalled, it is hard to communicate the memory (Ellis, 1984; Hancock & Frowd, 2002). A natural alternative then, according to Hancock and Frowd (2002), is to present groups of faces and allow a witness to select a few of these based on their similarity to the target face. Selecting in this way is rather like picking criminals from a mugshot album, which avoids the task of having to describe the target face.

The method used to create a composite with E-Face (i.e. getting witnesses to select, from a set of faces, a face that looks most similar to the target during each generation, as opposed to having them recall the face verbally or scrutinise isolated features) therefore capitalises on the need for face recognition and assists witnesses in moving into "recognition mode" (Tredoux, da Costa, Nunez & Rosenthal, 1999).

In a series of three experiments conducted in 1999, Tredoux et al. (1999) tested the first version of the E-Face system. In experiment 1, 15 participants used E-Face to reconstruct three faces whilst having their photographs in full view during reconstruction. 267 judges evaluated the resulting composites through matching and rating tasks. For the matching task, judges were required to select which face most resembled a reconstruction out of a 7-person photo-lineup. For the rating task, judges were required to rate each of the seven faces in the lineup according to its degree of similarity to the reconstruction, using a 9-point scale (1 = "not at all like him"; 9 = "looks a lot like him"). An overall identification rate of 51% was found for the matching task, and the average rating score for the three target faces was 72% (6.5 out of 9).

In experiment 2, 15 different participants used E-Face to reconstruct two faces (one familiar and one unfamiliar) from memory, after being exposed to them for 15 seconds each. 227 judges evaluated the resulting composites in the same way as in experiment 1. The identification rate for the familiar face was 58%, and 17% for the unfamiliar face. The average rating score was 72% (6.5 out of 9) for the familiar face and 47% (4.23 out of 9) for the unfamiliar face. In experiment 3, six participants used E-Face to reconstruct two unfamiliar faces from memory. 230 judges evaluated the resulting composites in the same way as in experiments 1 and 2, except that each judge evaluated only one reconstruction. An overall identification rate of 43% was found for both faces (Tredoux et al., 1999).

Tredoux et al.'s (1999) identification rates, for their three experiments evaluating the E-Face system, compare favourably to those recorded in other studies in which faces were reconstructed from memory (e.g. Christie and Ellis, 1981 with 23%; Davies, 1983 with 60% and 19%; Ellis et al., 1975 with 13%; Green and Geiselman, 1989 with 33%, 46% and 19%). Following this study, the first version of the E-Face system was revised, firstly to include a much larger high quality database. Secondly, the "ghost-like" composites (resulting from PBIL's scaling and centring solution), which did not allow for variation in shape, were eliminated. Thirdly, the system was revised to introduce colour, as the first version only used greyscale images.

Since Tredoux et al. (1999) only used the white database of the E-Face system in their evaluation; the other databases of the system needed to be evaluated. In 2000, as part of my Psychology Honours research project, I conducted a study that evaluated the black and coloured databases. In the first reconstruction condition of my study, ten participants used the E-Face system to make "in view" reconstructions of two faces (one black; one coloured). In the second reconstruction condition, ten different participants used the E-Face system to make "from memory" reconstructions of the same two faces, following an exposure to a two-minute video-clip of the same target persons committing a crime. Fifteen independent judges rated the resulting composites for similarity to their respective target and to three foils, using a 7-point scale.

On average, the "in view" reconstructions received a 46% rating score (3.22 out of 7) whereas the "from-memory" reconstructions received 42% (2.93 out of 7). An average rating score of 50% (3.55 out of 7) was found for the coloured target and 37% (2.60 out of 7) for the black target. In order to get an idea of the number of hits and misses made by the judges, a thresholding was conducted using the judges' ratings for each reconstruction compared to its respective target as well as to the other three foils. The results of this thresholding indicated that the average identification rate for the black target was 40% and a measly 11% for the coloured target (Prag, 2000).

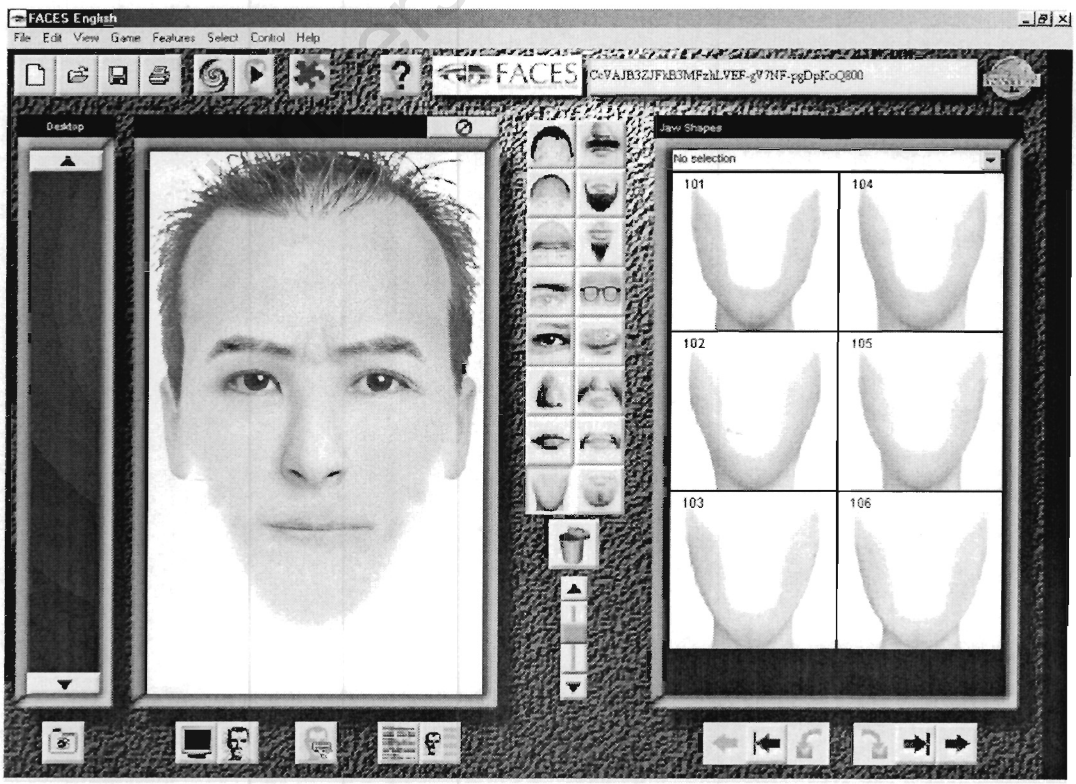
In another empirical study conducted in 2000, in which Tredoux collaborated with Nunez, 36 participants made E-Face reconstructions of one of three faces (black, white or coloured face) from memory or "in-view". Average identification rates were found to be 38% for the black database, 52% for the coloured database, and 76% for the white database (Tredoux, Nunez, Oxtoby, Rosenthal, da Costa, Prag & Pooley, 2002).

On the basis of the results of the afore-mentioned E-Face studies and verbal feedback from participants, it became clear that the decision to map face images to a common shape was a restrictive one. E-Face was therefore revised to introduce shape as a searchable and manipulable aspect. With this new version of E-Face, when a selection is made from the array of faces which users are presented with, both the texture and shape coefficients describing that face are read, and used to conduct parallel searches of texture and shape space (Tredoux et al., 2002).

To test the impact of the incorporation of shape into the latest version of E-Face, another empirical study was conducted. The study made an evaluation of the system under optimal ("in view", expert operator) and less optimal ("from memory", witness operator) conditions. The same procedure and measures were used as in the previous study mentioned above. Foils were similar-looking faces in a database of 300. The enhanced version of E-Face yielded an identification rate of 50% for novice/witness operators and an impressive 86% for expert operators (Tredoux et al., 2002).

4.2. Faces

Figure 2: The Faces system



The Faces system has been designed and created by a Canadian company called InterQuest Inc. The third version of the Faces software has been evaluated in this study. It contains a database of close to 4,000 specially coded facial features. The feature database comprises black-and-white facial features of 1,500 people, who volunteered to participate in photo sessions in an effort to develop this widely used facial composite system. By clicking these features with a mouse, adult faces of any sex or race between the ages of 17 to 60 can be created in a relatively brief period of time. All of the selected features are automatically blended together, making the resulting picture resemble a good-quality black-and-white photograph (Interquest Ltd., 1998).

As can be seen in Figure 2 above, the Faces operator has 16 feature libraries to examine when building the facial composite. Those feature libraries include the following: hair, head shapes, forehead lines, eyebrows, eyes, noses, lips, jaw shapes, moustaches, beards, goatees, glasses, eye lines, smile lines, mouth lines and chin lines. The "selection area" displays, in sequential order, miniature images of a selected feature library. Therefore, if the operator selects the "eye" library, for instance, pages of narrow eyes, deep-set eyes, etc. will be displayed. One can browse through these pages using the navigation buttons at the bottom of the "selection area" (Interquest Ltd., 1998).

The operator can also use the "first/last Page" buttons to go directly to the first or last page of the individual facial features. In addition, Faces has a handy "previous/next selected feature" button, which enables the operator to browse through all previously selected features. As an alternative to browsing through each and every page of feature sub-categories, one can use the drop-down "sub-categories selection menu", which displays the names of selected feature sub-categories in words, e.g. narrow nose, pointed nose, round nose, etc. As part of the "control area", one can use the "delete selected feature" button to remove selected features. Using the position slider, features can be moved up and down. Selected features can also be moved further apart, closer together, and reduced/enlarged in size (Interquest Ltd., 1998).

Once the operator is satisfied that the composite created best represents the target face, it can be saved onto the system. Once the image is saved, the Faces system automatically assigns an "InterCode" (a personalised, alphanumeric ID code), which can be used in future to recall the composite on the system. This "InterCode" also allows Faces operators to share facial composites with each other in a speedily manner. Printing the composite is simply done by clicking the "print" icon (Interquest Ltd., 1998).

Faces designers claim that they have developed a user-friendly system that creates images of high quality (Interquest Ltd., 1998). Unfortunately, no known empirical studies have been conducted to test the effectiveness of the Faces system.

Chapter 5: Interview Techniques to Aid Facial Composite Production

Several interview techniques have been developed thus far to assist the eyewitness or victim of a crime to recall specific details relating to the criminal event and/or the perpetrator. These include the "guided memory interview", the "cognitive interview", and the "enhanced cognitive interview". Various law enforcement agencies and forensic artists have developed their own unique interview techniques for this purpose too. However, a standardised interview technique has not yet been developed for the specific purpose of facial composite production. The "cognitive interview" procedure developed for this specific purpose is discussed here.

5.1. Guided Memory Interview

In 1981, Malpass and Devine invented a technique called the "guided memory interview". This interview takes the witnesses step by step through the build-up to the criminal event as well as through the proceedings of the criminal event. The technique reminds witnesses of their immediate reactions to the criminal event, their thoughts and feelings of the offender, the crime scene and its surroundings, and the offender's general appearance (Malpass & Devine, 1981).

An investigation into the effectiveness of this technique turned out to be quite fruitful. Five months after witnessing a staged vandalism, Malpass and Devine (1981) tracked down all the witnesses and asked them to pick out the vandal from a 5-person target-present lineup. Recognition accuracy was greater for witnesses who had undergone the guided memory interview (60%) than for control witnesses (40%). Krafka and Penrod (1985), who based their context reinstatement manipulation upon the technique developed by Malpass and Devine (1981), found that accurate identification was at 60% (32.7% greater) when context reinstatement was done two hours after witnesses were exposed to the perpetrator and at 50% (19.2% greater) when context reinstatement was done 24 hours after.

Deffenbacher (1988) points out that one drawback of Malpass and Devine's technique is that it depends on the interviewer being well aware of the version of events relating to the crime scene, which is unlikely to happen in real life.

5.2. Interview Techniques adopted by Forensic Artists

For Boylan, a forensic sketch artist who has worked on investigations such as the Unabomber case and the Polly Klaas murder, the most important element to look for from an eyewitness or victim is that he or she has experienced trauma during the criminal incident. Boylan (2000) explains that trauma is what embeds images into memory. She adds that emotional involvement is an underlying prerequisite for a good witness, and hence the production of an accurate sketch. During her interviews with eyewitnesses or victims of crime, she carefully attempts to move beyond the conscious memory of the event to reach the subconscious level, which is where the accurate image of the offender normally resides, according to Boylan (2000). Central to her technique is patient questioning. She engages witnesses in conversation, mixing questions about the offender with discussion of unrelated topics that do not provoke any anxiety. She also avoids asking obvious, unhelpful questions such as, "Did he have a long, thin nose or a short, wide one?" Boylan (2000) asserts that people do not remember faces as collections of parts; therefore her questioning is predominantly holistic. Although her technique is time-consuming, the accuracy of her composites perhaps confirms the validity of her method. It should be pointed out that Boylan's interviewing technique does not rise above the level of anecdote since empirical tests have not been conducted.

Taylor, a freelance portrait artist who has worked for roughly two decades at the Texas Department of Public Safety in Austin and taught at several law enforcement academies including the FBI academy, uses a similar approach to Boylan. Taylor (2000), however, engages in more of an information-gathering approach prior to the interview. Taylor (2000) finds out from the police whether her witness was a victim or merely an observer. If her witness was a victim, Taylor (2000) finds out the nature and degree of violence and psychological trauma that he/she endured. In addition, prior to the interview, Taylor (2000) finds out whether the offender was of the same race as the witness (which Taylor says can make a difference in how well the witness discerns certain features) and how apprehensive the witness was in making an identification. All this information directs her as to how she should conduct the interview with the witness. Taylor (2000) uses a cognitive style of interviewing which allows the witness to describe the event in his/her own words. Then with carefully crafted questions, she gains additional information.

Taylor (2000) has come up with a "composite-specific interview", which she says has evolved over time. It represents a compilation of thoughts, methods and experience of various sketch artists around the world, and comprises the following six stages: -

i) *Pre-interview:* Before the interview, the artist should have an understanding of situational factors like duration of view, point of view and distance, lighting conditions and time since the criminal event. The artist should also be aware of the crime and general information about the victim, especially physical and psychological conditions. The artist should not be exposed to any photographs of possible suspects prior to conducting the interview.

ii) *Introduction/rapport-building stage:* The artist should warmly greet the witness. Any fears, anxieties and potential psychological blockages should be dispelled. The artist should make the witness aware that a break is welcome at the witness's request. The artist should then explain the composite drawing process.

iii) *Initial drawing stage:* The artist should do an initial "run-through" to gain the facial description, and pay particular attentions to what aspects of the description are most emphasised by the witness. The artist should concentrate on the proportional arrangement of the features and establish the basic facial character. The artist should show the witness the initial drawing and gather first impressions.

iv) *Fine-tuning drawing stage:* The artist should use multiple-choice question to fine-tune the initial drawing. Thereafter the artist should selectively introduce photographic references to the witness. The artist should not introduce too many images to avoid deterioration of the witness's mental image of the perpetrator.

v) *Finishing touches:* Once the fine-tuning is complete, the individual components of the face should be reviewed from top to bottom, reinstating the context of the event if appropriate.

vi) *Concluding the interview:* The artist should conclude the interview by informing the witness that he/she should feel free to contact the artist if he/she recalls additional information. Witness anxieties regarding aspects of the investigation, release of the composite image to the media, etc. should be addressed (Taylor, 2000).

5.3. South African Police Service (SAPS) Interview

Law enforcement agencies around the world have developed their own methods of interviewing witnesses or victims of crime, prior to getting them to produce a facial composite of the perpetrator. According to Van Dijk (personal communication, October 1, 2001), a composite operator for the SAPS since 1998, the following forms part of the SAPS Interview: -

i) The investigating officer first asks the witness to provide a description of the entire criminal event.

ii) Thereafter the investigating officer asks the witness to provide a description of the perpetrator. The investigating officer prompts the witness by a series of verbal descriptors about the perpetrator, i.e. race, age, height, build, scars, tattoos, hair, eyes, nose, mouth, chin, teeth, facial hair, complexion, jewellery, hats and spectacles.

iii) The investigating officer gives the witness the option of not describing those characteristics or features about the perpetrator that they were not able to recall.

iv) Once this information is ascertained, the witnesses proceed to assist the investigating officer (composite system operator) to produce a facial composite of the perpetrator.

5.4. The Original and Advanced Cognitive Interview

Fisher, Geiselman and Raymond (1987a) discovered that the standard interview protocol typically used by police during an investigation consists of a sequence of short, closed-ended questions. The interview procedure adopted by the SAPS is no exception.

In an effort to elicit more complete and more accurate information from witnesses about the events of the crime and of the perpetrator at the description phase of a police interview, Geiselman and Fisher, in 1984, developed a set of retrieval techniques designed to improve spoken, verbal recall. They called the procedure the "Cognitive Interview" (CI). In its original form, the cognitive interview focuses on guiding witnesses through four general memory-jogging techniques derived from some basic, well-established psychological principles of memory: -

1) *Reinstate the context:* Think about physical surroundings and personal emotional reactions that existed at the time of critical past events;
2) *Report everything:* Report everything that comes to mind about those events, no matter how fragmentary or seemingly inconsequential;
3) *Recall the events in different orders:* Recount events in a variety of chronological sequences (i.e. beginning to end, reverse order, forward or backward from highly memorable points); and
4) *Change perspectives:* Adopt different perspectives while recalling events, such as having a crime victim describe the offenders from his or her own point of view and from that of a bystander at the scene (Geiselman, Fisher, Firstenberg, Hutton, Sullivan, Avetissian & Prosk, 1984).

According to Bower (1997), actual crime victims and witnesses often experience more anxiety, display poorer communication skills, and confront more confusion about their roles in an interview. In an attempt to address these issues, Fisher, Geiselman, Raymond, Jurkevich and Warhaftig (1987) made revisions to the original CI to develop the "Enhanced Cognitive Interview" (ECI). As part of the ECI, the interviewer begins by building rapport and encouraging the witness to take an active role in recalling information instead of merely responding to a list of questions. The witness first describes the criminal event in his/her own words, without being interrupted. The interviewer then probes further with specific techniques like the "change perspectives" technique, for example. The ECI is based on four core principles: -

1) *Memory-event similarity:* Attempt to have the witness mentally recreate the environment surrounding the incident;
2) *Focused retrieval:* Help the witness to focus by refraining from asking too many short-answer, undirected or irrelevant questions that tend to break concentration;
3) *Extensive retrieval:* Encourage witnesses not to terminate the memory retrieval after the first attempt, for example, by continuing chronologically from beginning to end; and
4) *Witness-compatible questioning:* Interviewers should place themselves in the witness's frame of mind and ask questions compatible with the situation in which the witness found himself or herself (Bower, 1997; Fisher & Geiselman, 1992; Fisher, McCauley & Geiselman, 1994; Köhnken, 1995; Memon, Vrij & Bull, 2003; Milne & Bull, 1999; Osterburg & Ward, 1997).

Most researchers have found the CI and ECI to be very effective at increasing the amount of correct information recalled by a witness, without increasing the relative level of confabulation. To illustrate this, Geiselman, Fisher, MacKinnon and Holland (1986) found that the CI elicited 35% more correct information than the standard police interview. In another study, Fisher, Geiselman and Amader (1989) trained seven police detectives to use the CI technique and found that these trained detectives elicited 47% more correct information after than before their training, and 63% more information than did the nine untrained detectives. George and Clifford (1996) noticed similar results after police officers (from Hertfordshire Constabulary in the UK) received CI training. In a series of studies conducted at the University of California using police officers as interviewers and students, non-students and children as witnesses to realistic crimes, the CI was found to elicit significantly more correct information with no apparent increase in confabulations (Memon & Bull, 1991).

Bekerian and Dennett (1993) conducted a meta-analysis of 27 studies. This revealed that, on average, 30% more information was accurately reported and that relatively little false information was reported. In a meta-analysis of 42 studies (55 individual comparisons involving nearly 2500 interviewees) conducted by Köhnken, Milne, Memon and Bull (1999), it was found that overall the CI substantially increased the number of correct details by roughly 41% but also led to an increase in errors or confabulated details. Geiselman (1999), after reviewing a collection of CI research papers, concluded that the CI can be effective in eliciting significantly more information from interviewees, typically without affecting error rates.

Despite overwhelming evidence mentioned above as to the usefulness of both the CI and ECI, several authors have criticised the comparison in many of these studies with a "standard police interview", as it has been found that there is little standardisation of normal interview techniques. According to Taylor (2000), the "standard police interview" is based on a question-and-answer approach with frequent interruptions, seeking primarily the "who, what, when, where, why and how". Since police practice varies in different parts of the world, it comes as no surprise that some "standard interview techniques" are better than others. Hence it is difficult to make definitive claims as to the exact advantage of the CI (Ainsworth, 1998). Memon and Higham (1999) reiterate that the standard interview is not a useful control group against which to evaluate particulars of the CI as there is too much variability in its administration. Memon, Cronin, Eaves and Bull (1996) provide further criticism by purporting that some of the improved performance shown by CI witnesses relates to motivation, i.e. the more motivated the interviewer, the greater the performance of the witness.

Ainsworth (1998) asserts that as with any other technique, the CI does have its limitations. Firstly, whilst the CI may be a useful technique in getting co-operative and helpful witnesses to recall more information, it may not work with more reluctant or obstructive witnesses. Secondly, the CI is time-consuming and may not be viable for trivial crimes or in situations where police officers are under pressure and severely pressed for time. Thirdly, since the CI needs to be adapted to the specific needs of the interview (Fisher & MacCauley, 1995), it requires greater flexibility and concentration from the interviewer, which may not come naturally to many police officers.

Despite these limitations, the bulk of research on the CI and ECI has demonstrated its effectiveness in helping individuals recall information in the description phase of a police investigation. One could thus infer that if used during the composite phase of investigation, the CI would elicit more accurate, detailed information and result in more accurate composites (Koehn et al., 1999).

The efficacy of the CI in assisting with the construction of facial composites has however shown to yield mixed results. In a study conducted by Koehn et al. (1999), in which participants were required to produce faces using the Mac-a-Mug Pro composite system, an advantage of the CI over the standard police interview could not be found. The fact that the composites in the Koehn et al. (1999) study produced recognition rates all near floor level may serve to explain this result. Quite the opposite was found in a study conducted by Luu and Geiselman (1993). They used the FIS composite system (which permits the examination of individual facial features either in the context of other facial features or in isolation) and found improved composite production when using the CI. Koehn et al. (1999) suggest that these conflicting results can be explained by the fact that the CI is a holistically based interview and thus the production of better quality composites will be facilitated by a holistically based composite system, like the FIS system. However Davies and Milne (1985) found that a "guided memory interview" and "context revival" procedure enhanced facial composites even when a feature based composite system (Photo-Fit) was used. Koehn et al. (1999) concede that they cannot find an explanation for this, which decreases the validity of their suspicion.

5.5. Modifications made to the Original CI to suit Facial Composite Production

Koehn et al. (1999) claim that since the CI has been used traditionally with verbal tasks (i.e. describing information about a criminal event), that it would be inappropriate to use it for the visual task of facial composite production without making minor revisions to it. They argue that since Schooler and Engstler-Schooler (1990) found that verbalisation about a stimulus can impair later recognition (in a series of six experiments), the first way that the CI should be revised is to promote more pictorial processing and minimise verbal processing. This will allow for the "verbal overshadowing" effect to be overcome. Further support for this effect comes from three experiments conducted by Dodson, Johnson and Schooler (1997) who found that face recognition suffers when a verbal description of a previously seen face is generated. Laughery and Fowler (1980) would agree with the above arguments as they purport that a possibility for the limited usefulness of facial composite systems relates to the suggestions that a key limiting factor is the witness's ability to verbally describe a face. A logical solution would therefore be to remove or at least minimise the verbal component of the task (Laughery and Fowler, 1980).

However a review by Meissner, Brigham and Kelley (2001) revealed that the verbal overshadowing effects are quite small, and are specific only to certain experimental manipulations. In Meissner and Brigham's (2001) meta-analysis of the verbal overshadowing effect in face identification, their analysis of the pattern of means showed that overshadowing effects were more likely to occur when the identification task immediately followed the description task, and when participants were given an elaborative (as opposed to a free recall) instruction during the description task. Finger and Pezdek (1999) found that verbal overshadowing was reduced after a delay had occurred between verbal description and face identification, which led them to believe that the verbal description does not overwrite the visual memory of the face but instead makes it less accessible at the time of face identification. The outcomes of the above studies raise doubt as to the relevance of the first change made to the original cognitive interview by Koehn et al. (1999).

Koehn et al. (1999) advise that a second way the CI should be revised is to encourage the interviewee to search for any trait judgements and/or labels they may have inadvertently made about the target's face at the earlier encoding. This second modification is substantiated by evidence that such judgements and labels assigned at encoding are helpful at retrieval for recognising faces (Chance & Goldstein, 1976; McKelvie, 1976; Mueller, Carlomusto & Goldstein, 1978 in Koehn et al, 1999).

Interestingly, recent work by Frowd, McQuiston-Surret, Ness and Hancock (2004) seems to validate the modifications made to the original CI to suit facial composite production mentioned above. In 2 experiments, Frowd et al. (2004) examined the role of the interview used with witnesses and the subsequent starting point for composite production. Eighty "mock" witnesses were asked to construct a composite using the UK PROfit system under varying interview and starting conditions. Initial data indicate that a process involving a novel "holistic" interview, requiring witnesses to make a number of personality judgments about their target face, may be preferable to the traditional approach involving the original CI.

5.6. CI Procedure developed for this Study

Since a standardised procedure has not yet been developed regarding how to conduct the CI for the specific task of facial composite production, the implementation of the CI procedure for experiment 2 of this study has been guided by Koehn et al.'s modifications made to the original CI, by Malpass and Devine's (1981) "guided memory interview", by suggestions offered by Ron Fisher (personal communication, June 3, 2002) and Chris Meissner (personal communication, May 31, 2002), as well as by interview techniques adopted by highly regarded forensic artists, Boylan and Taylor. The following CI procedure was the outcome: -

i) In order to help the witness concentrate, the interviewer should ask the witness to close his/her eyes or stare at some blank field in the interview room.

ii) The interviewer should then ask the witness to play the entire criminal event in their heads - not to verbalise it, but merely think about it.

iii) Whilst the witness does this, the interviewer should ask the witness to think of the moment when he/she had the best view of the perpetrator.

iv) Once he/she has retrieved this best view, the interviewer should ask him/her to think about his/her thoughts and feelings at that particular moment.

v) Keeping this best view in mind, the interviewer should ask the witness to search for trait judgments, labels, first impressions that he/she associated with the perpetrator.

vi) The interviewer should then ask the witness to focus intensely on the time when he/she had the best view of the perpetrator's face, so as to acquire a mental image of it.

vii) The interviewer should ask the witness to "refresh" that mental image of the perpetrator's face several times during the composite production process in order to minimise or prevent disruption of the mental image (caused by exposure to the facial composite system).

Chapter 6: Experiment 1

Prior to the conduct of this experiment, the E-Face system comprised three databases, i.e. (1) white, (2) black and (3) coloured. The coloured database comprised coloured and Indian faces. Within the broad aim of evaluating the effectiveness of the E-Face composite system, a key aim of this experiment was to discover whether the E-Face system would be improved if the Indian faces were removed from the coloured database, and used to create a new Indian database. It was hypothesised that better quality composites of coloured faces would be produced using a purely coloured database, and that better quality composites of Indian faces would be produced using a purely Indian database.

To test this hypothesis, three new E-Face databases were created, i.e. (1) coloured database comprising only coloured faces, (2) Indian database comprising only Indian faces, and (3) Indian/coloured database comprising equal numbers of Indian and coloured faces. Each of these databases comprised coloured and/or Indian males between the ages of 18 and 21.

6.1. Database Building Process

The first step in developing these databases was to collect photographs of 18- to 21-year-old male faces. Secondary schools, tertiary institutions (i.e. universities and colleges), and shopping malls were used to source the faces that would make up the respective databases. After being granted permission from the respective authorities at each of these locations, a digital camera and external flash was set up on a tripod. All in all, 657 photographs were collected; 332 coloured and 325 Indian faces. One "front-on" photograph was taken of each participant's face. Participants were guided by the photographer with respect to the way they should orient their faces so that they were in consistent geometrical relation to the film plane. The photographer also instructed them not to display any facial expressions (e.g. smile, frown, etc.) whilst the photograph was being taken, to keep their mouths closed, and to look directly into the camera lens. The 657 participants involved were offered a packet of crisps in exchange for photographs of their faces.

The 657 photographs collected were transferred from the digital camera memory card to computer. Each of these digital images were carefully examined to ensure that they were of good quality with respect to lighting, exposure and orientation, otherwise the eigenface analysis would produce unusable images. This sorting process resulted in the elimination of 32 coloured faces and 25 Indian faces, leaving a remaining 300 coloured faces and 300 "Indian faces" to use to create the respective databases. These 600 digital images were cropped with the aid of PhotoImpact (a "picture-editing" software programme) to remove the background surrounding each participant's face, leaving behind a mugshot image of each face.

In preparation for building the coloured and Indian databases, the 300 coloured faces were separated from the 300 Indian faces in preparation for the next phase of the database building process. In preparation for building the Indian/coloured database, 150 Indian faces were randomly selected from the 300 Indian faces, and 150 coloured faces were randomly selected from the 300 coloured faces.

Once this separation process was complete, the images were "landmarked" and then warped in an effort to map all the images to an average shape. The faces need to be normalised before they can be treated as sufficiently "similar" to process as data points in PCA. Each image was "landmarked" using a specially developed DOS-based "landmarking" programme, which writes Cartesian co-ordinates in terms of particular screen dimensions. Each facial image was loaded individually into the "landmarking" programme, and a face template was used to carefully and precisely place 81 points (which make up the face template), on or around the various features of the face. Nine points of the face template go around each of the eyes; eight points around each of the eyebrows; six around the nose; nine around the mouth; seven around each of the ears; ten around the outer face; and eight around the hair. Each of these points is used as a vertex for a set of triangular tessellations, and the faces are then bi-linearly mapped into the average face shape (Tredoux et al., 2002).

Once the "landmarking" phase was complete, the images were warped using a computer-based programme (written in Microsoft Visual C++) called "FaceWarp", specially developed by Oxtoby in 2000. This very flexible piece of software has the ability to warp any image annotated with the "landmarking" programme to any appropriately defined shape (usually the average of an ensemble of annotated faces). "FaceWarp" was therefore used to generate an average face for the 300 faces that made up each of the three databases (see Figure 3).

Figure 3: Average faces generated by "FaceWarp" – coloured, Indian and Indian/coloured (from left to right)

Each image was individually loaded into "FaceWarp" and then warped to the shape of its database average face, as can be seen in the Figure 4 below. Each warp was examined to ensure that they were of acceptable quality. When certain warps were not found to be of acceptable quality (e.g. severely deformed, asymmetrical), their original images were "landmarked" a second time. This involved shifting the Cartesian points and placing them more precisely on and around the various facial features. Once this was done, those images were warped a second time. This process of "landmarking" and warping continued until each warp was found to be of acceptable quality (see Figure 5 for warped images of acceptable quality).

Figure 4: "FaceWarp"

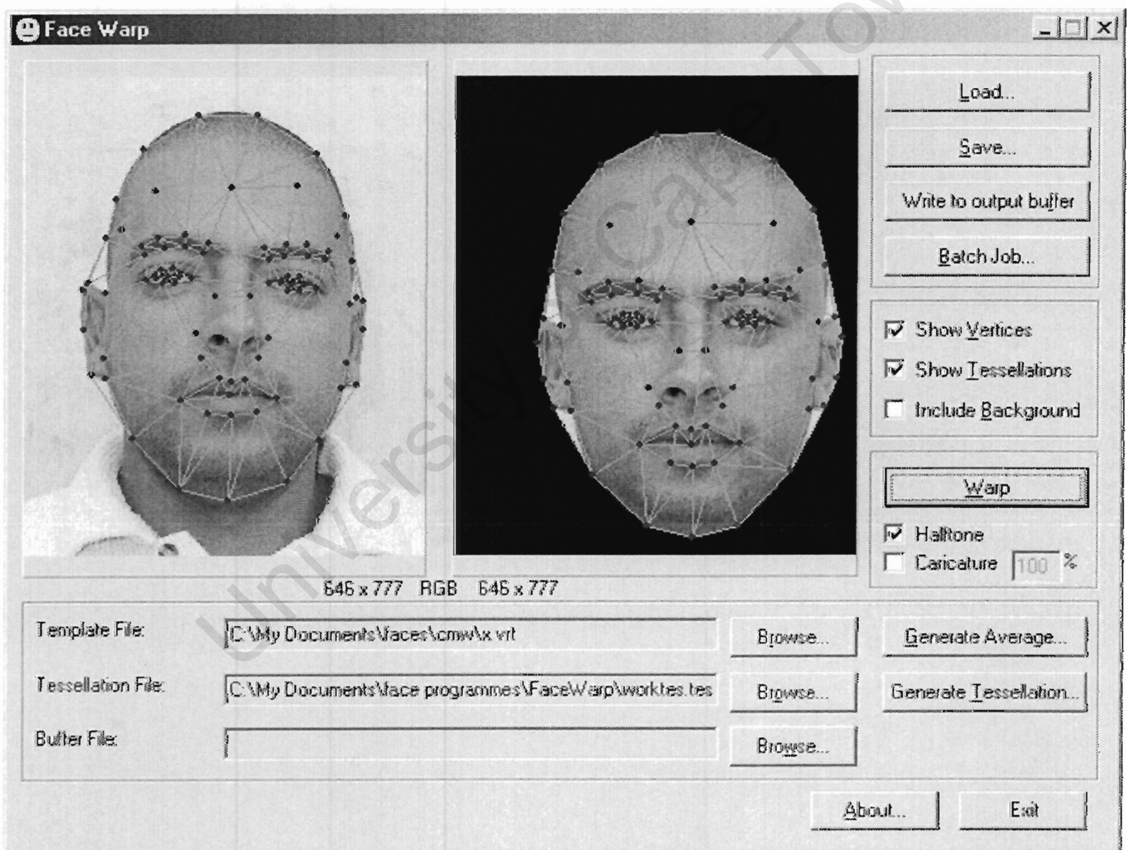
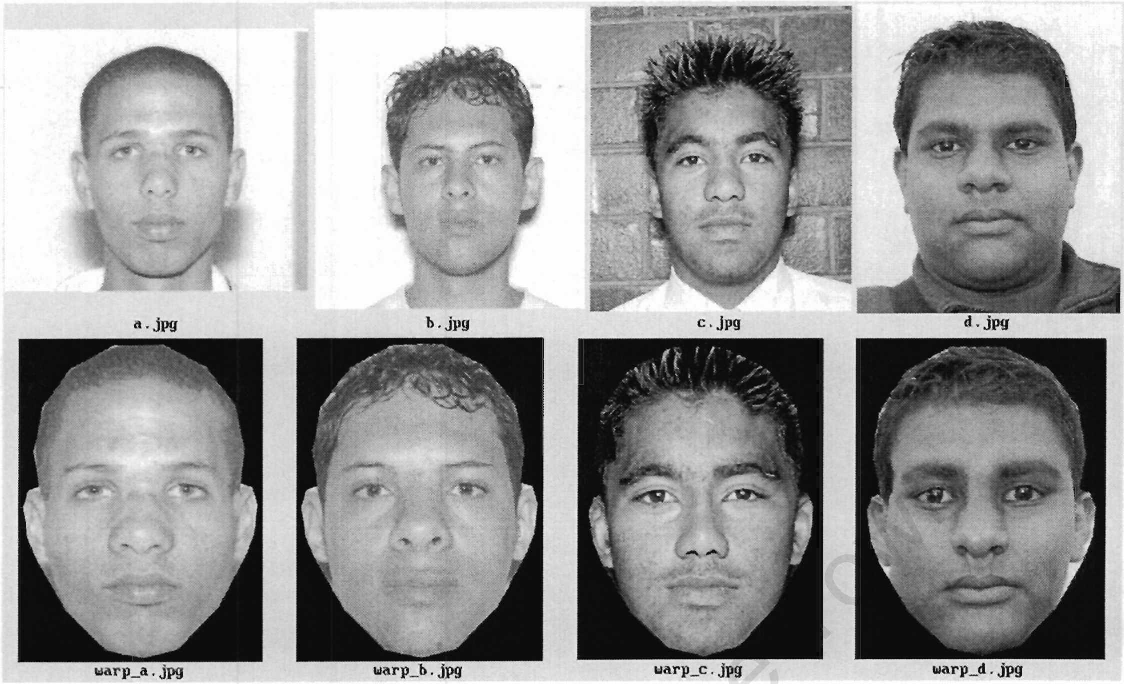


Figure 5: Examples of original images (top) and warped images (bottom)



Once the warping process was complete, Principal Component Analysis (PCA) was conducted on the normalised faces, to derive "eigenfaces". This was done separately for each of the three databases. Specially developed software utilities were developed for this purpose, run in "MS-DOS Prompt" mode. First, the average face for the database being built was calculated. This average face was then subtracted from the ensemble of faces in the database, which is a characteristic step in PCA. Thereafter the eigenvectors of a face image matrix was found and written to a text file. A matrix multiplication programme, capable of rapidly multiplying extremely large matrices, was used to create the eigenfaces required for E-Face. An average face of all the standardised faces was then generated, "landmarked" and saved.

To add shape to the facial composite system, PCA was conducted on the "face-shape" data, derived from the "landmarking" stage when the faces were annotated. The shape and texture eigenmodels were then combined to build an "appearance model", as described by Cootes and Taylor (1999). Thereafter, the "landmark scale converter" was used to reduce the size of the eigenfaces so that they fit properly into the display boxes of the E-Face system.

The process involved in building the coloured, Indian and Indian/coloured databases has been outlined above. With the creation of these three databases of the E-Face, the experiment could proceed.

6.2. Method

The experiment consisted of two phases, i.e. (1) a reconstruction phase and (2) an evaluation phase. For the reconstruction phase of this experiment, two coloured faces and two Indian faces were used as targets for reconstruction. Participants were required to make "in view" reconstructions of one coloured face and one Indian face, using one of three databases each time they reconstructed a face. The three databases that were used by participants in their reconstruction attempts were the ones developed for the purpose of this experiment, i.e. (1) the coloured database, (2) the Indian database and (3) the Indian/coloured database. To clarify, each participant reconstructed two faces with photographs of the target faces in full view during the composite production process, thereby assessing the fundamental ability of the E-Face system and three of its databases under ideal conditions for novice users. Careful counter-balancing took place in order to prevent confounding the potential learning curve with the ability to reconstruct actual faces.

Independent judges evaluated the quality of the resulting composites through a rating task. The two best reconstructions of each target face, determined by the judges' ratings of the reconstructions, were then further evaluated by another group of independent judges using a mugshot test procedure.

Participants

A total of 64 members of the Cape Town public participated in this experiment. As they were walking in the city centre, they were approached and asked to contribute to an experiment. Each participant was paid for agreeing to participate. Their level of education ranged between high school and university education. They were all familiar enough with computers to perform their impending tasks. Twenty-four members of the Cape Town public (13 males: 11 females) participated in the reconstruction phase. The average age of the participants was 28 years old (ranging from 19 to 43 years of age). With regard to the race distribution of participants in the reconstruction phase, it was ensured that the three major South African races or population groups were equally represented to allow for the investigation of the "other-race effect". Hence 8 black, 8 white and 8 coloured (4 coloured; 4 Indian) participants were selected.

Twenty members of the Cape Town public (11 males; 9 females) acted as "judges" in the first part of the evaluation phase, and 20 (8 males; 12 females) participated in the second part of the evaluation phase. Different participants were used for each part of the evaluation phase to avoid the risk of contamination. As in the reconstruction phase of the experiment, it was ensured in the evaluation phase that there were equivalent numbers of white, black and coloured participants.

Those who participated in the reconstruction phase were not allowed to participate in the evaluation phase. Also, those who participated in the first part of the evaluation phase were not allowed to participate in the second part of the evaluation phase. Prior to participating, all participants were informed that a facial composite system was being tested.

Materials

Three databases specially developed for the purpose of this experiment were used, i.e. (1) coloured database, (2) Indian database, and (3) Indian/coloured database.

The four faces used as target faces during the reconstruction phase, and the face used as the trial face were of different-looking 18-year-old male students who volunteered to have their faces photographed and used for the purpose of this experiment. Targets had to be between the ages of 18 and 21 because the E-Face system does not yet have the ability to reconstruct facial composites outside of this age range. In addition, the E-Face system does not yet have the ability to generate female composites, thus only male targets could be used.

Two of the four targets were coloured; the other two were Indian. The student representing the trial face was Indian. The four targets and the student representing the trial face were thoroughly informed about the nature of the research. It was explained that by allowing the experimenter to take digital images of their faces from different angles, the E-Face system could be evaluated under the "in view" condition.

A digital camera was used to take photographs of the four target faces and the trial face from three different angles, i.e. (1) front-on, (2) side profile and (3) three-quarter profile (see Figures 6 to 10 below). These digital images were laser-printed in colour onto five high-quality, white A4 sheets of paper. On each page, there were three images of one target face, each image (9cm x 11cm in dimension) taken from the three different angles. This was done so that participants in the reconstruction phase could get a good idea of what the faces looked like whilst reconstructing them.

Figure 6: Trial face

Figure 7: First coloured target (cm1)

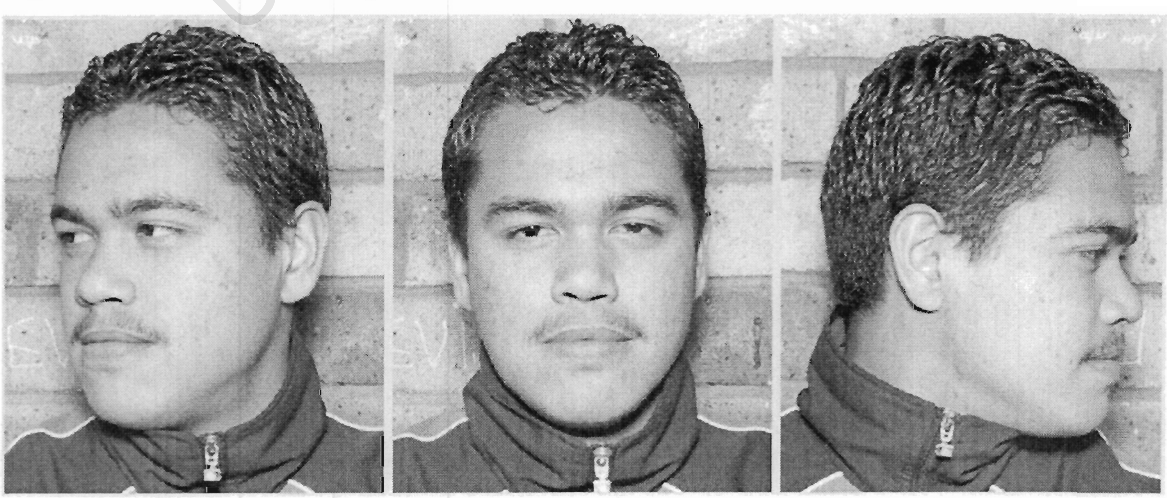


Figure 8: Second coloured target (cm2)

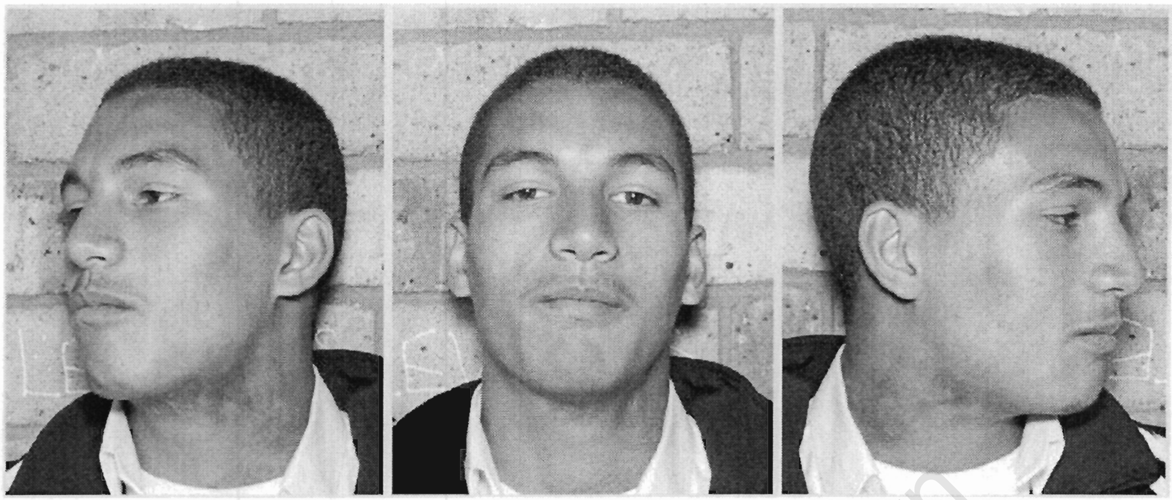


Figure 9: First Indian target (im1)

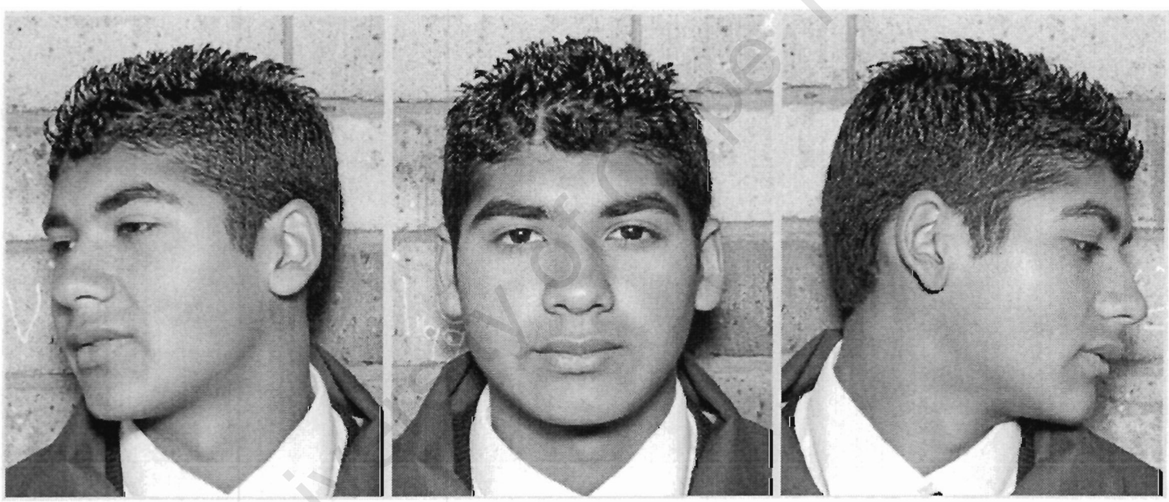
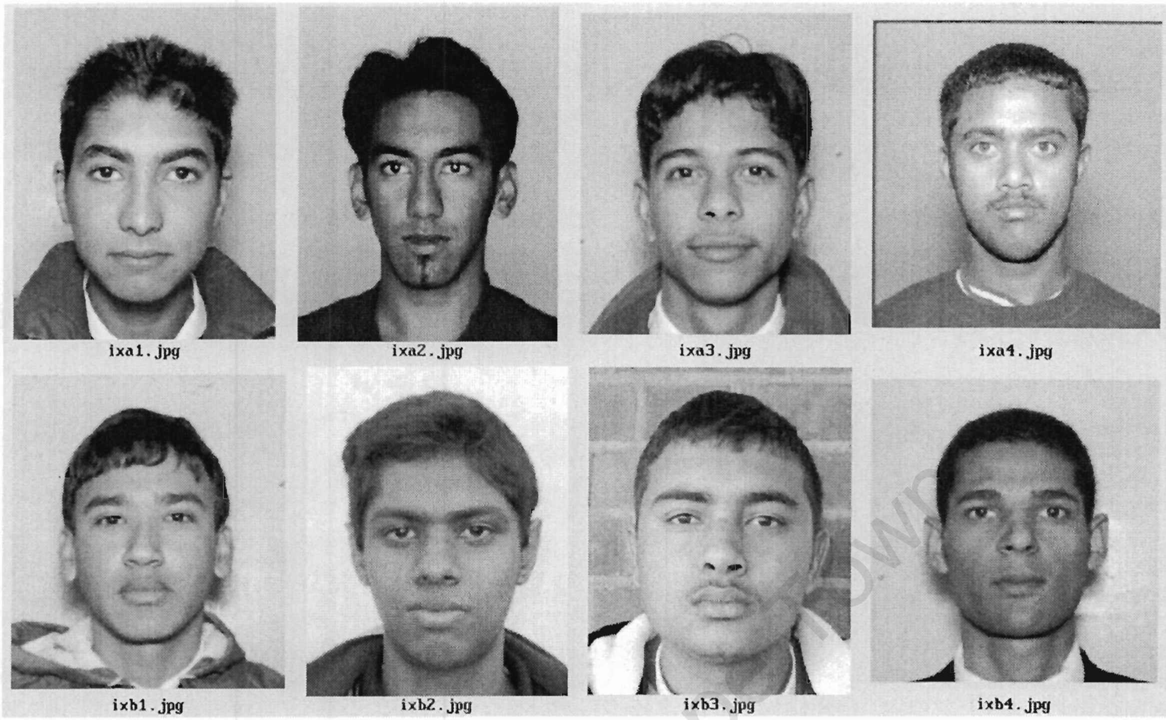


Figure 10: Second Indian target (im2)

For the rating task, a computer programme for rating faces called "Face Test" was used, which was specially developed by Nunez in 2000. The four target faces, the 48 reconstructions made by the participants, and the 16 distracter faces, were all loaded into this specially developed programme in preparation for the rating task. As shown in Figures 11 and 12, the 16 distracters comprised eight coloured faces (randomly drawn from the coloured database) and eight Indian faces (randomly drawn from the Indian database). To clarify, there were four distracter faces to each of the four target faces.

Figure 11: Eight coloured distracters

Figure 12: Eight Indian distracters



For the mugshot test procedure, the two best rated reconstructions of each of the four target faces were identified (see Figures 13 and 14). These eight images (11cm x 15cm in dimension) were individually laser-printed in colour onto eight high-quality, white A4 sheets of paper. The four target faces and 156 distracter faces (randomly and equally selected from the Indian and coloured databases) were used to compile four mugshot albums. Each mugshot album included the target and 39 distracter mugshots. For each mugshot album, the mugshots were laser-printed in colour onto four high-quality, white A4 sheets of paper, with 10 mugshots on each sheet (see Figure 15). Each mugshot was 5cm x 8cm in dimension. The four mugshot albums were randomly arranged in two different orders, therefore a total of eight mugshots albums were produced in preparation for the mugshot test procedure.

Figure 13: Highest rated composites (bottom) of each target face (top)

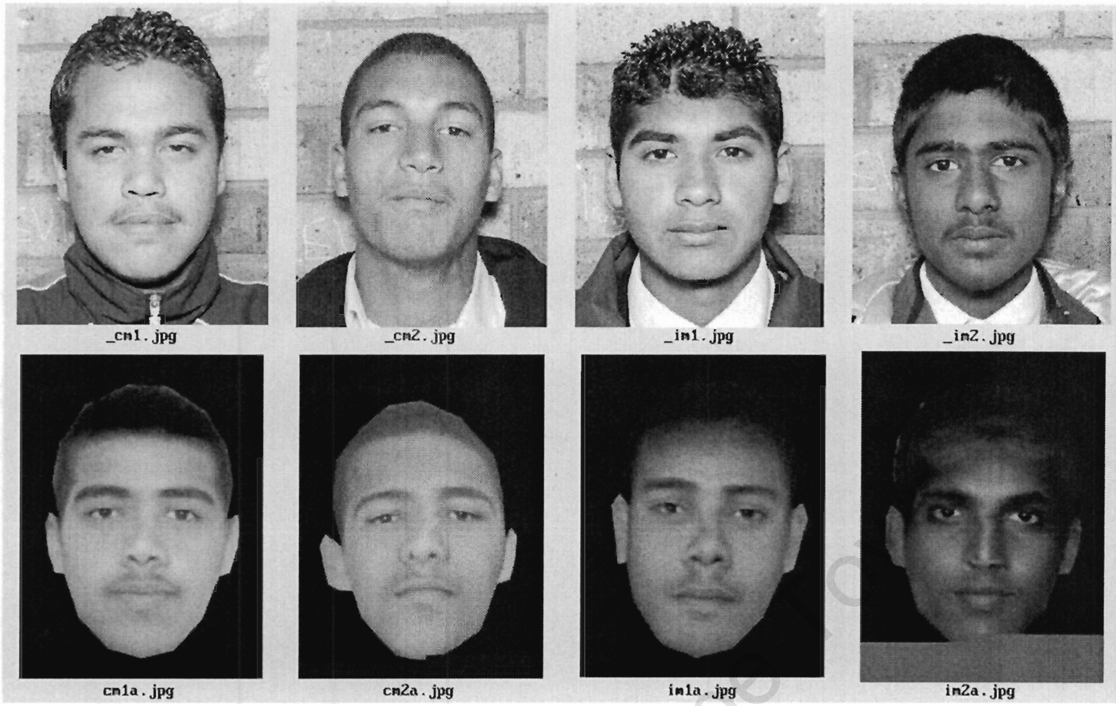


Figure 14: Second highest rated composites (bottom) of each target face (top)

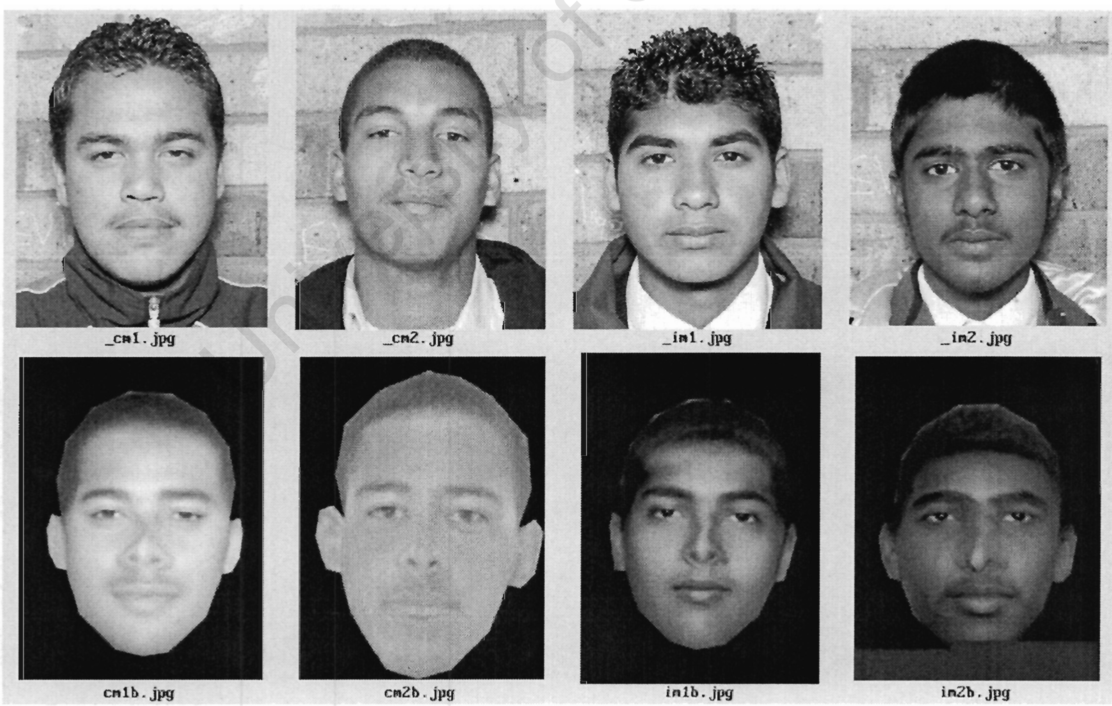
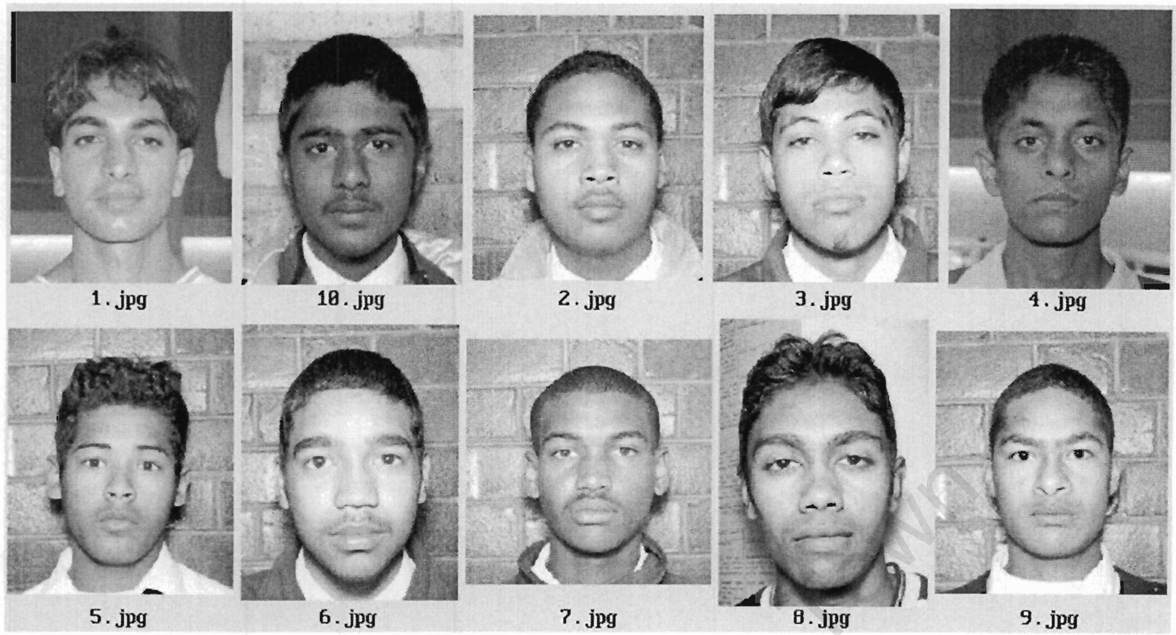


Figure 15: First sheet of mugshot album containing im2 (second, top)



Reconstruction Phase

The participants were told that their task was to use E-Face to produce a reconstruction of two of the four targets. Participants were notified that, throughout the composite production process, they would have three photographs of each of the target faces they were required to reconstruct, in full view. Each of these photographs was taken from three different angles, i.e. (1) front-on, (2) side profile and (3) three-quarter profile.

Participants performed their task individually. Initially the experimenter briefly outlined the nature and involvement of their impending tasks. A demonstration of the composite system was made and structured guidelines were provided. Participants were advised not to look at a single feature or features of the eigenface, but to rather look at the eigenface as a whole when deciding which one to select. Thereafter each participant was taken through a trial run, where the experimenter guided the participant through the process of reconstructing the trial face. Participants were given an opportunity to ask any questions at this point. Once participants felt confident that they understood how the system operated, they were given the first of two faces to reconstruct in view using one of three databases (see figures 16 to 18). The databases to be used by participants when reconstructing each face, was randomly assigned.

Figure 16: E-Face system (coloured database)

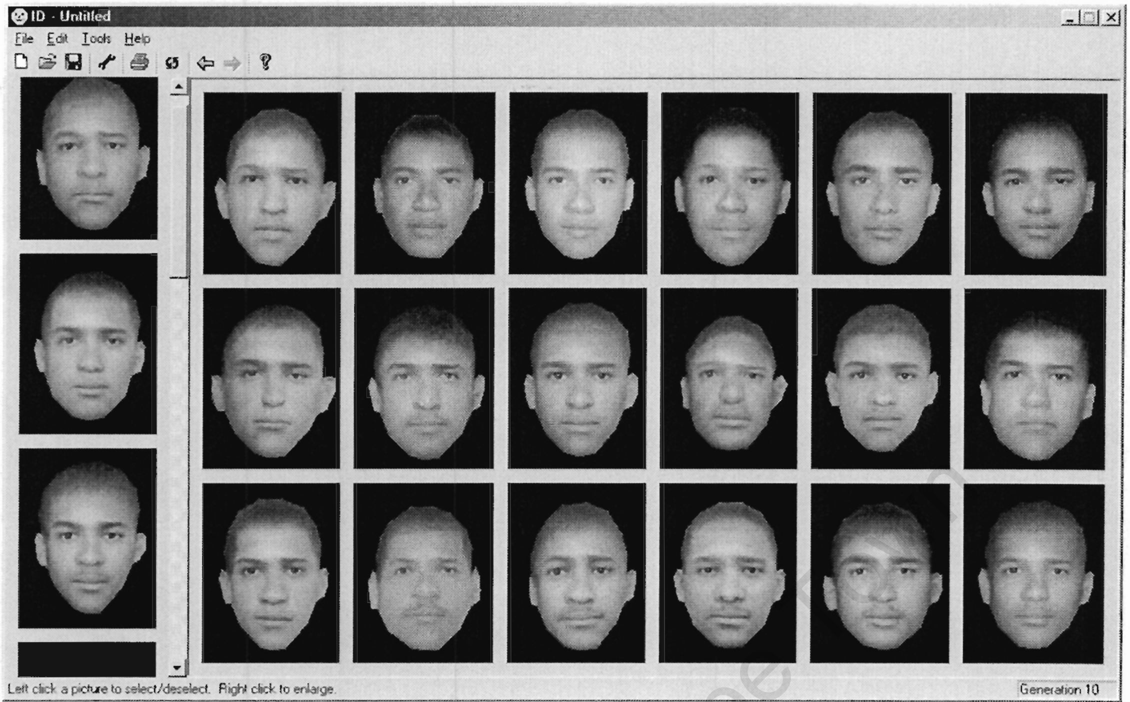


Figure 17: E-Face system (Indian database)

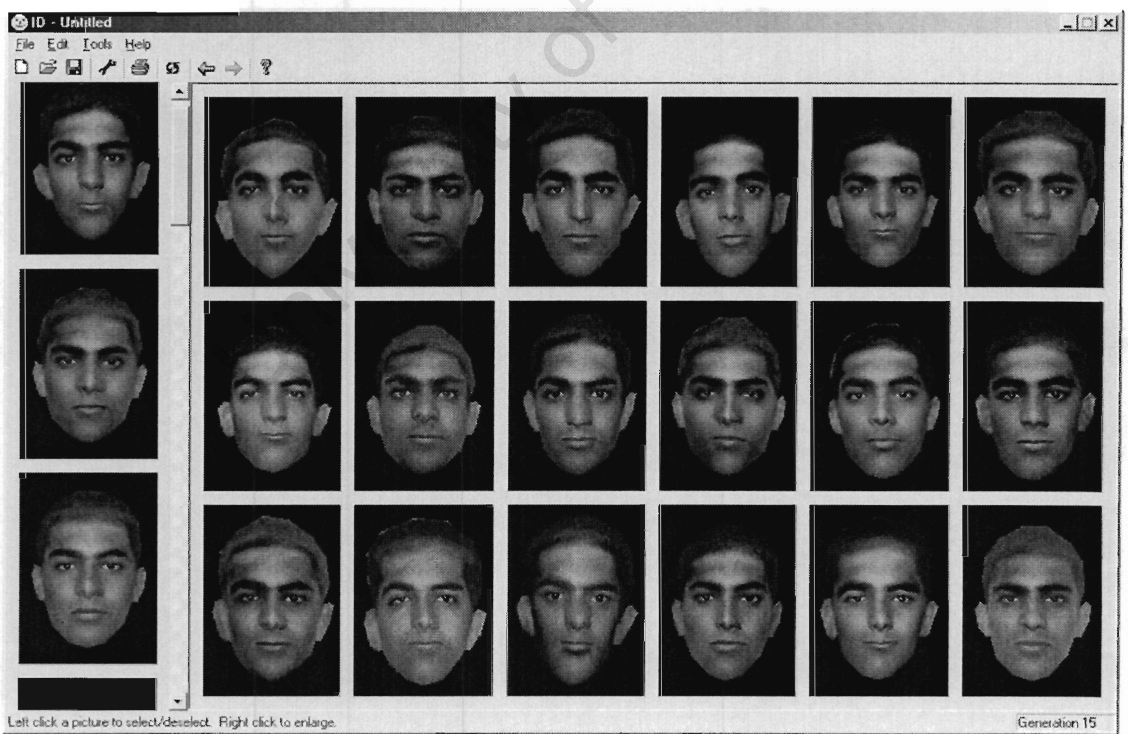
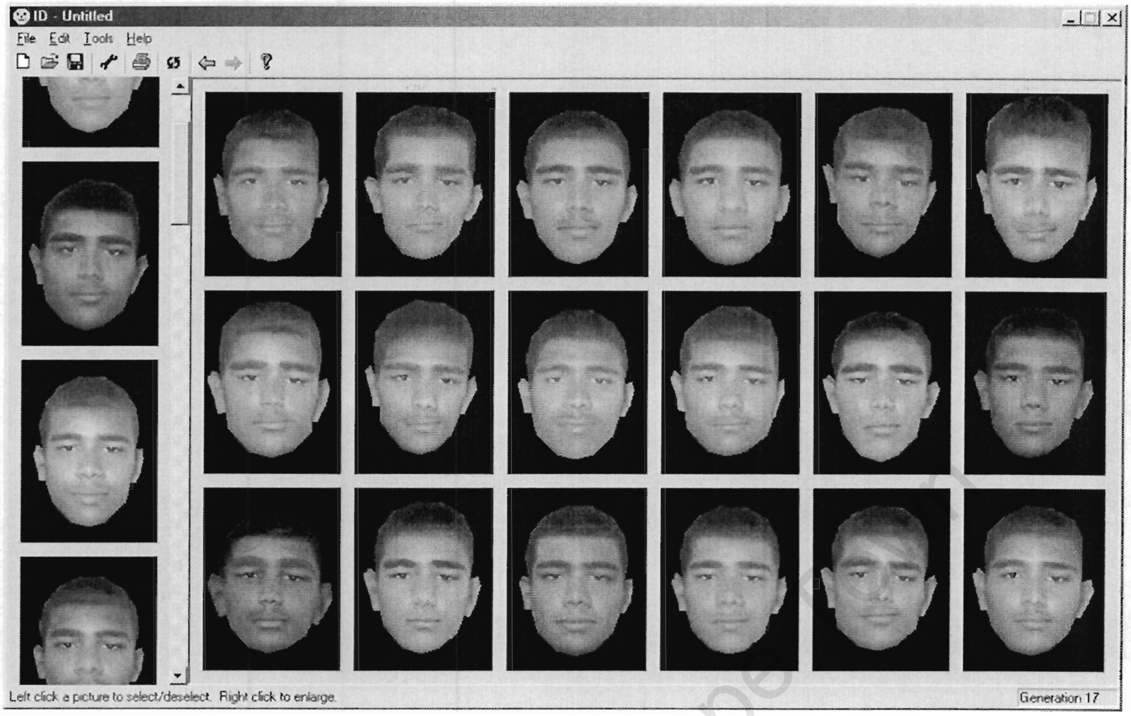


Figure 18: E-Face system (Indian/coloured database)



Participants were notified that there were no time constraints for this task, and that they were allowed to make as many changes as they wished until they were satisfied with their reconstructions. Participants were allowed to ask for assistance during the composite production process only if they were having trouble using the composite system. In cases where assistance was required, the experimenter ensured that participants' composite production process was not influenced in any way.

Participants were allowed to stop once they believed that the composite they had produced represented a "good enough" likeness to the target face. They then made their final choice by referring to the history of all eigenfaces selected from the first to the last generation. Their choice from this array was the final product of the reconstruction process. Upon completion of the reconstruction of the first face, participants were asked, on a 7-point Likert-type scale, how accurate they thought their reconstructions were to the target face (1 for "highly inaccurate"; 7 for "highly accurate"). The reconstruction was then saved as a digital image onto a computer. The time taken for each participant to complete each reconstruction, and the number of generations made to arrive at the finally chosen eigenface, was recorded.

Participants then proceeded to reconstruct a second face using one of the three databases, constantly referring to the photographs of the target face during the entire procedure. Upon completion of the reconstruction of the second face, participants were again asked, on a 7-point Likert-type scale, how accurate they thought their reconstructions were to the target face. The reconstruction was saved, and the time taken to complete the reconstruction and the number of generations made to arrive at the finally chosen eigenface, was recorded.

At the end of the entire reconstruction procedure, participants were taken through a debriefing session were asked to express their thoughts and feelings about the task in general, about which face they found more difficult to reconstruct, about the databases used for the reconstruction of the two faces, and about how the composite system could be improved.

To summarise, 48 "in view" reconstructions were made during the reconstruction phase. Twelve reconstructions were made of each of the four target faces. Of the 12 reconstructions made of each of the four faces, four were made using the coloured database; four were made using the "Indian database; and four were made using the Indian/coloured database. Therefore 16 reconstructions were made using each of the three databases.

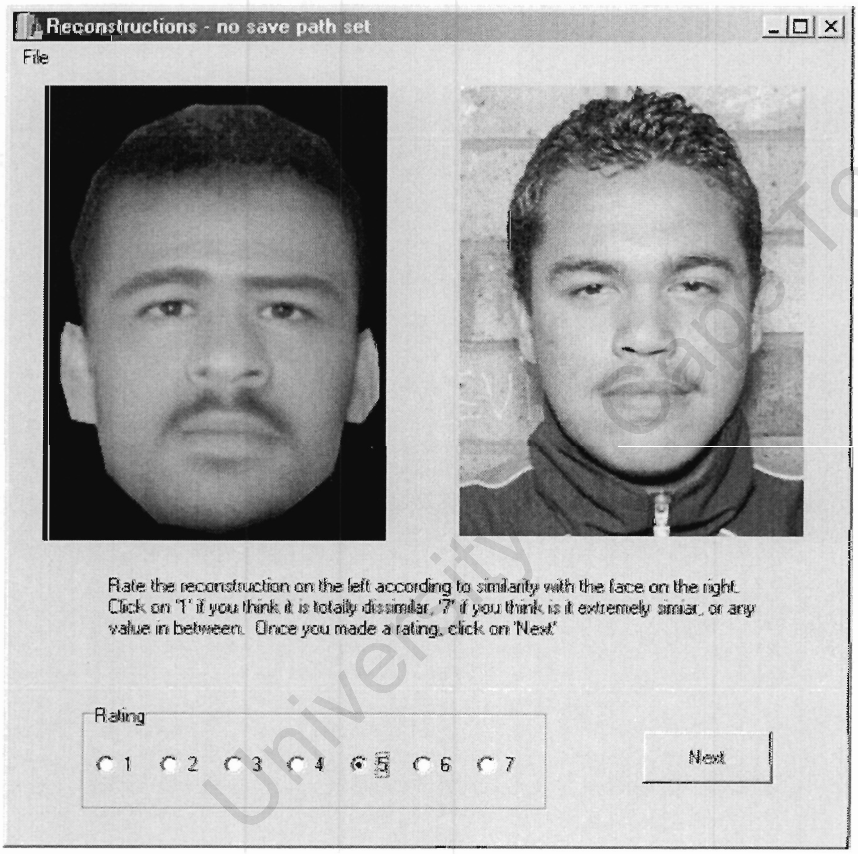
Evaluation Phase

In a separate phase of the experiment, the quality and utility of the 48 composites produced by participants in the reconstruction phase was assessed. Forty evaluators individually participated in two evaluation tasks, 20 evaluators per task. These two tasks were (1) a rating task and (2) a mugshot test procedure. Each task was self-paced and individually performed. Prior to performing their tasks, evaluators were informed about the nature of the research and about how their participation contributed to achieving the aims and objectives of the experiment.

<u>Task 1: Rating task</u>

For the rating task, evaluators used a specially developed computer programme for rating faces called "Face Test". They were seated in front of a computer screen and were presented with two facial images next to each other, each image 5cm x 7cm in dimension. Beneath these two facial images, were the following instructions, "Rate the reconstruction on the left according to similarity with the face on the right. Click on "1" if you think it is totally dissimilar, "7" if you think it is extremely similar, or any value in between. Once you made a rating, click on 'Next'". Beneath these instructions were numbers 1 to 7 and the "next" button (see Figure 19).

Figure 19: "Face Test" (specially developed programme for rating faces)



The experimenter made it known to the evaluators that the image on the left would always be an E-Face facial composite produced by participants in the reconstruction phase, while the face on the right may or may not represent the corresponding target face, i.e. the face that participants in the reconstruction phase were trying to reconstruct. When deciding on their ratings, the evaluators were instructed to look at each face as a whole, rather than focus on specific features. Each evaluator rated each of the 48 reconstructions against its corresponding target face and four distracters, thus a total of 240 ratings were made. Each evaluator rated a different random order of reconstructions.

Task 2: Mugshot test procedure

The two best reconstructions of each of the four target faces, determined by evaluator ratings given to each of the reconstructions, were further evaluated using a mugshot test procedure. Thus eight of the original 48 facial composites were further evaluated. Four mugshot albums were created, each comprising one of the four targets and 39 distracter faces. These four albums were arranged in two different orders, therefore eight mugshot albums were prepared for this second evaluation phase.

Evaluators were handed one (of eight) facial composite together with one (of eight) mugshot album. Their task was to examine the mugshot album, and then compile a list of all those mugshots (each of which were numbered) that they believed were most likely matches to the facial composite being evaluated. Upon completion of the first mugshot test, evaluators were handed the second (of eight) facial composite together with the second (of eight) mugshot album. The same procedure ensued until all eight composites were evaluated. A different random order of composites and corresponding mugshot albums were presented to each evaluator.

By allowing evaluators to set the criterion for the initial set of possibles and have them decide how many mugshots they think represent the composite, one is able to measure precision as well as accuracy. A precise and accurate composite will be one where there is a small chosen field and the target is within that field.

6.3. Results

Reconstruction Phase

On average, the participants took 16.17 minutes to reconstruct each composite; recorded times ranged from 7 to 29 minutes. Participants took an average of 11.46 generations to arrive at their finally chosen composite; recorded generations ranged from 4 to 45. Upon completion of their reconstructions, participants were asked on a 7-point Likert-type scale how accurate they thought their reconstructions were to the target face, 1 being "highly inaccurate" and 7 being "highly accurate". All reconstructions made by participants received an average satisfaction rating of 4.77 out of 7, which equates to 68%. Table 1 below details mean reconstruction times, mean number of E-Face generations, and mean operator satisfaction ratings. (Coloured Male has been abbreviated to "cm"; Indian Male to "im"; and Indian Coloured to "ic").

Table 1: Mean reconstruction times (in minutes); mean number of generations made during E-Face reconstructions; and mean operator satisfaction ratings (out of 7)

			Reconstruction Times				Number of Generations				Satisfaction Ratings			
T/Fce	D/Bse	N	Mean	Min	Max	Std.Dev.	Mean	Min	Max	Std. Dev.	Mean	Min	Max	Std.Dev.
		48	16.17	7	29	5.85	11.46	4	45	6.69	4.77	2	7	1.29
cm1		12	14.25	7	24	4.77	11.08	5	19	4.89	5.17	3	7	1.19
cm2		12	15.75	8	26	5.94	10.33	5	18	4.01	4.92	3	7	1.08
im1		12	18.17	10	29	7.09	13.67	5	45	10.75	4.5	2	7	1.31
im2		12	16.5	8	24	5.42	10.75	4	22	5.31	4.5	2	7	1.57
	cm	16	15.75	7	26	5.64	11	5	21	5.1	4.88	2	7	1.31
	im	16	16.25	8	29	6.66	12.81	4	45	9.66	4.44	2	7	1.31
	ic	16	16.5	9	27	5.56	10.56	5	22	4.23	5	3	7	1.26
cm1	cm	4	15.5	7	24	7.85	10.75	6	19	5.74	6	6	6	0
cm1	im	4	12.75	11	15	1.71	13.25	5	19	6.24	5.5	4	7	1.29
cm1	ic	4	14.5	11	19	3.7	9.25	7	12	2.22	4	3	5	0.82
cm2	cm	4	16.25	8	26	7.68	12	6	14	4	4.5	3	6	1.29
cm2	im	4	18.25	13	26	5.56	11.75	7	18	4.57	4.5	4	5	0.58
cm2	ic	4	12.75	9	19	4.35	7.25	5	9	1.71	5.75	5	7	0.96
im1	cm	4	14.25	10	20	4.65	10	5	21	7.57	4.5	4	5	0.58
im1	im	4	23.25	13	29	7.32	19.5	10	45	17.02	3.75	2	5	1.5
im1	ic	4	17	10	27	7.26	11.5	9	15	2.65	5.25	4	7	1.5
im2	cm	4	17	15	22	3.37	11.25	8	18	4.57	4.5	2	7	2.08
im2	im	4	10.75	8	13	2.63	6.75	4	10	2.5	4	2	5	1.41
im2	ic	4	21.75	18	24	2.87	14.25	7	22	6.18	5	3	6	1.41

Verbal feedback

It was unanimously agreed amongst participants in the reconstruction phase that the E-Face system was user-friendly, and that no extensive computer expertise was necessary to operate it. Participants found it useful to have a history of each face selected in each generation to refer to, as this assisted them in making their final selection. The ability of the system to go backwards and forwards during the reconstruction procedure was also commended by participants, especially in situations when they accidentally made a selection, or when they were not satisfied with an eigenface selected in a particular generation.

Many participants however found the task of producing E-Face composites more challenging and confusing as they progressed through the reconstruction procedure. This was because the eigenfaces, which they were required to select from, became more and more similar as they proceeded from one generation to the next. Also, by being exposed to so many eigenfaces during the task, some participants found themselves drifting away from their mental images of the target face.

In terms of how the E-Face system could be improved, many participants suggested that there be a facility to manipulate and refine the finally chosen face. By importing the finally chosen face into a picture editing or freehand drawing programme, for example, one could make facial refinements (like stretching, narrowing, and shifting facial features, adjusting face shape, etc.) and add specific detail (like scars, ageing lines, etc.). Participants also suggested that it would be easier to make a final selection if the history of faces selected in each generation could be viewed at a single glance, instead of having to scroll down a narrow panel of faces (as it is currently set up). Also, to improve the E-Face system, participants would like to see the quality of certain eigenfaces, which they were required to choose from, be improved so that they look less distorted and more realistic.

Evaluation Phase

For the evaluation phase, results are separately reported for the two evaluation tasks of the experiment, i.e. (1) the rating task and (2) the mugshot test procedure.

Task 1: Rating task

For the rating task, evaluators used a 7-point Likert-type scale to rate the 48 reconstructions (i.e. facial composites produced during the reconstruction phase) against its corresponding target face and four distracters. Ratings were made according to similarity, 1 being "extremely dissimilar" and 7 being "extremely similar". Table 2 below shows the similarity rating means for the reconstructions.

Table 2: Rating means (out of 7)

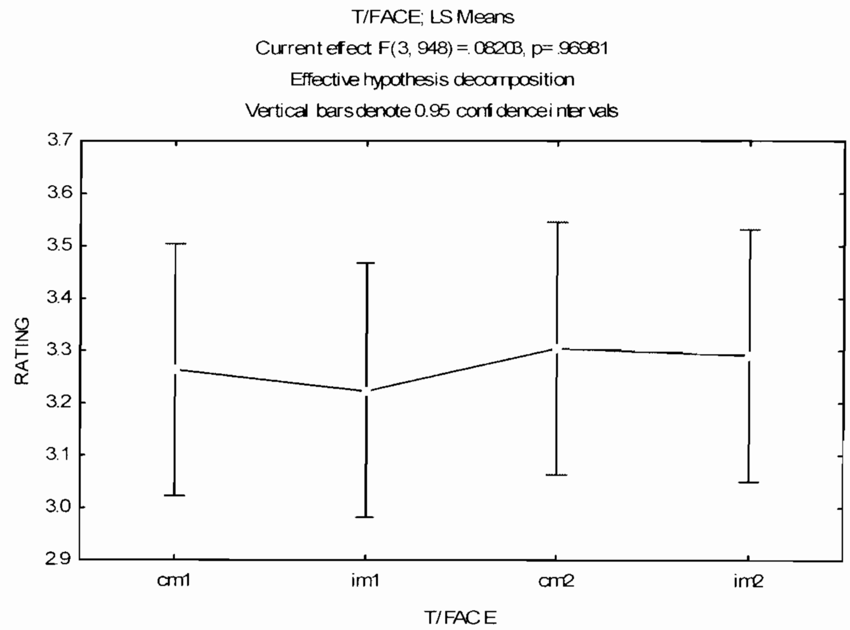
T/Face	D/Base	Valid N	Mean	Minimum	Maximum	Std. Dev.
		960	3.27	1	7	1.90
cm1		240	3.26	1	7	1.77
cm2		240	3.30	1	7	1.96
im1		240	3.23	1	7	1.94
im2		240	3.29	1	7	1.93
	cm	320	3.31	1	7	1.95
	im	320	3.37	1	7	1.84
	ic	320	3.14	1	7	1.91
cm1	cm	80	3.38	1	7	1.85
cm1	im	80	3.23	1	7	1.74
cm1	ic	80	3.19	1	7	1.72
cm2	cm	80	3.51	1	7	2.04
cm2	im	80	3.19	1	7	1.84
cm2	ic	80	3.21	1	7	2.01
im1	cm	80	3.19	1	7	1.94
im1	im	80	3.38	1	7	1.84
im1	ic	80	3.11	1	7	2.06
im2	cm	80	3.15	1	7	1.98
im2	im	80	3.68	1	7	1.93
im2	ic	80	3.05	1	7	1.85

Table 3: Two-way factorial ANOVA rating results

	SS	Df	MS	F	*p*
Intercept	10270.42	1	10270.42	2834.517	0.000000*
T/Face	0.89	3	0.30	0.082	0.969810
D/Base	8.70	2	4.35	1.201	0.301399
T/Face*D/Base	19.06	6	3.18	0.877	0.511159
Error	3434.92	948	3.62		

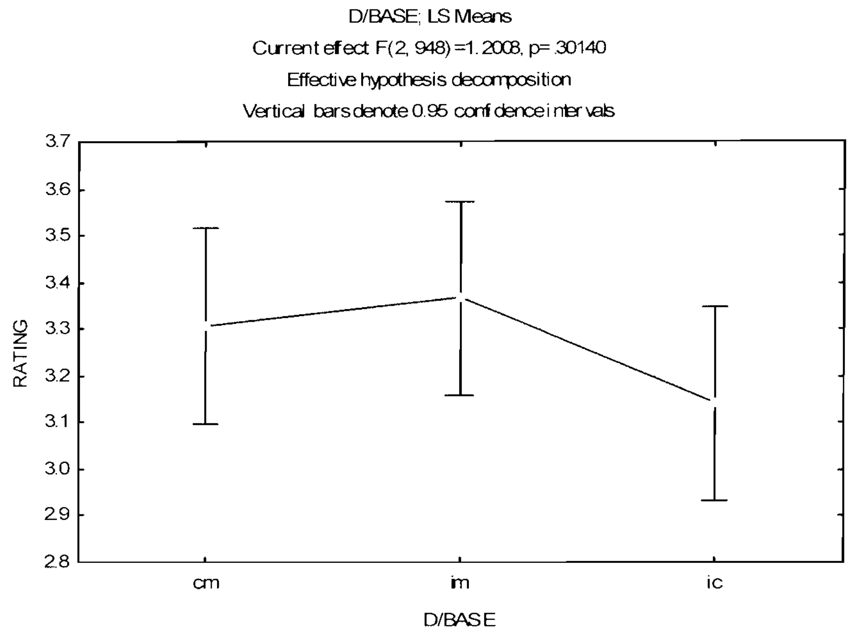
All reconstructions made by participants received an average rating score of 3.27 out of 7, which equates to 47%. As can be seen in Table 3 above, the two-way factorial ANOVA results show that the main effect for "Target Face" was not significant, $F(3, 948) = 0.082$, $p > 0.05$. "Cm2" (M = 3.30) performed the best while "im1" (M = 3.23) performed the worst, however the difference between their rating scores is not significant. Figure 20 below presents the four target rating means graphically.

Figure 20: "Target" rating means (out of 7)



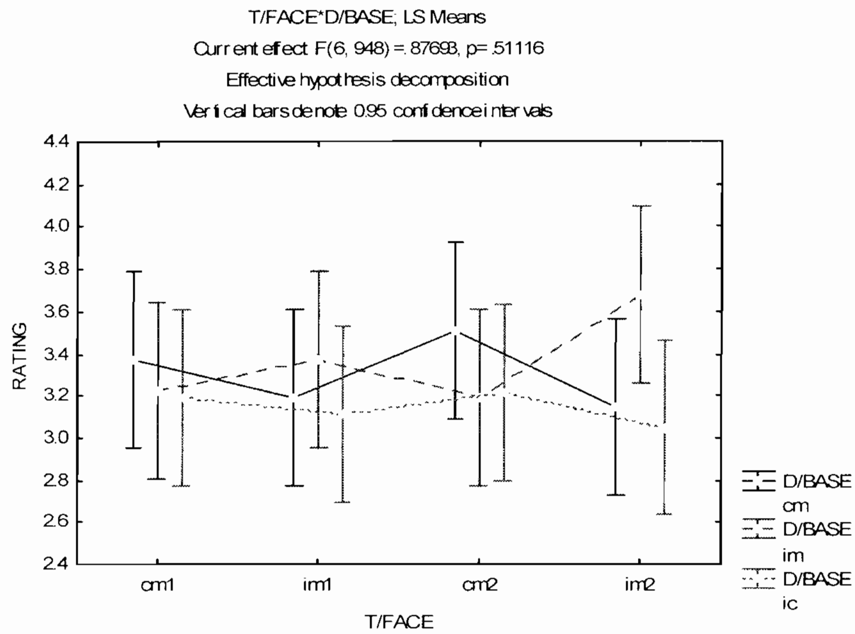
The main effect for "Database" was not significant, F (2, 948) = 1.201, $p > 0.05$. The "im" database (M = 3.37) performed the best of the three databases. This was shortly followed by the "cm" database (M = 3.31), and then the "ic" database (M = 3.14). Figure 21 below presents the four database rating means graphically.

Figure 21: "Database" rating means (out of 7)



The interaction between "Target Face" and "Database" was not significant, F (6, 948) = 0.877, $p > 0.05$. For "cm1", the highest evaluation rating was obtained when the "cm" database was used (M = 3.38), and the lowest when the "ic" database was used (M = 3.19). For "cm2", the highest evaluation rating occurred when the "cm" database was used (M = 3.51), and the lowest when the "im" database was used (M = 3.19). For "im1", the highest evaluation rating was obtained when the "im" database was used (M = 3.38), and the lowest when the "ic" database was used (M = 3.11). For "im2", the highest evaluation rating was obtained when the "im" database was used (M = 3.68), and the lowest when the "ic" database was used (M = 3.05). Figure 22 below presents the "Target x Database" rating means graphically.

Figure 22: Rating means for "Target x Database" (out of 7)



Evaluation Identification

As part of the rating task, evaluators were required to make similarity ratings of the 48 reconstructions against the corresponding target face as well as against four distracter faces in order to obtain identification rates made by evaluators. By setting up the rating task in this way, one could argue that evaluators are not only making similarity ratings between the composites and their corresponding target faces, but that they are also attempting to identify the target from a 5-person target-present lineup. To clarify, a thresholding was conducted using the evaluators' rating scores for each reconstruction compared to its respective target, as well as to the four distracters. If the target received a rating higher than or equal to the highest rating of the rest of the four distracters, it would be classified as a "hit". This allows one to infer that the target would have been selected from a 5-person lineup by the evaluator. Conversely, if the target received a rating lower than the rest of the four distracters, it would be classified as a "miss". Hence, one could infer that the target would not have been selected from a 5-person lineup by the evaluator.

The results of the thresholding revealed that all reconstructions made by participants received an identification rate of 51% (494 hits out of 960). "Cm1" performed the best with an identification rate of 55% (132 hits out of 240); "im2" the worst with 45% (109 hits out of 240). Table 4 details the mean identification rates.

Table 4: Mean evaluation identification rates

T/Face	D/Base	Valid N	Hits	Identification Rate %
		960	494	51.46
cm1		240	132	55.00
cm2		240	131	54.58
im1		240	122	50.83
im2		240	109	45.42
	cm	320	177	55.31
	im	320	163	50.94
	ic	320	154	48.13
cm1	cm	80	42	52.50
cm1	im	80	44	55.00
cm1	ic	80	46	57.50
cm2	cm	80	51	63.75
cm2	im	80	41	51.25
cm2	ic	80	39	48.75
im1	cm	80	44	55.00
im1	im	80	39	48.75
im1	ic	80	39	48.75
im2	cm	80	40	50.00
im2	im	80	39	48.75
im2	ic	80	30	37.50

To determine whether or not evaluators had given the target the highest rating purely by chance, a cumulative binomial test was conducted (see Appendix B). When evaluators made 4 or more hits out of 12 from a 5-person lineup ($p < 0.010406$), it would be unreasonable for them to have given the target face the highest rating purely by guessing. In contrast, if evaluators made less than 4 hits out of 12 from a 5-person lineup, it could have been sheer luck that they had given the target the highest rating. Two of the 20 evaluators managed hits purely by chance for "cm1"; three for "cm2"; six for "im1"; and five for "im2".

Measurement of the "other-race effect"

The evaluation identification results were used to measure the "other-race effect", which was explored separately for the reconstruction and evaluation phases. In the reconstruction phase, it was ensured that the three major South African races or population groups were equally represented to allow for the investigation of the "other-race effect". Hence 8 black, 8 white and 8 coloured (4 coloured; 4 Indian) participants were selected.

The black participants were best at reconstructing "cm1" (59% identification rate), while the white participants were worst (50% identification rate). The white participants were best at reconstructing "cm2" (59% identification rate), while the black participants were worst (43% identification rate). The black participants were best at reconstructing "im1" (53% identification rate), while the white participants were worst (48% identification rate). The black participants were best at reconstructing "im2" (58% identification rate), while the white participants were worst (38% identification rate). (See Appendix C for the mean "other-race" reconstruction identification rates outlined above).

As in the reconstruction phase of the experiment, it was ensured in the evaluation phase that there were equivalent numbers of white, black and coloured participants. The coloured evaluators were best at recognising "cm1" (62% identification rate), while the black evaluators were worst (49% identification rate). The white evaluators were best at recognising "cm2" (60% identification rate), while the black evaluators were worst (51% identification rate). The white evaluators were best at recognising "im1" (68% identification rate) and "im2" (51% identification rate), while the coloured evaluators were worst at recognising "im1" (37% identification rate) and "im2" (42% identification rate). (See Appendix D for the mean "other-race" evaluation identification rates outlined above).

These results do not provide clarity as to which of the three races is better at reconstructing coloured and Indian faces. However, even if they did provide this clarity, generalisation could not be drawn from them not only because of the small sample size used, but also because the above "other-race" comparisons are "untested".

Task 2: Mugshot Test Procedure

The best reconstructions of each of the four target faces (hereafter referred to as "cm1A", "cm2A", "im1A" and "im2A") and the second best reconstructions of each of the four target faces (hereafter referred to as "cm1B", "cm2B", "im1B" and "im2B") were selected for further evaluation in a mugshot test procedure. These were determined by examining the evaluators' rating scores given to each of the reconstructions. The best rating scores for "cm1", "cm2", "im1" and "im2" were 3.95, 4.20, 3.85 and 3.95 (out of 7) respectively; second best rating scores were 3.55, 3.95, 3.65 and 3.90 (out of 7) respectively (see Appendix E). These eight reconstructions can be viewed in Figures 13 and 14.

As part of the mugshot test procedure, evaluators were handed one (of eight) reconstruction together with one (of eight) 40-person target-present mugshot album. Their task was to examine the mugshot album, and then compile a list of all those mugshots (each of which were numbered) that they believed were most likely matches to the facial composite being evaluated. Upon completion of the first mugshot test, evaluators were handed the second (of eight) facial composite together with the second (of eight) mugshot album. The same procedure ensued until all eight composites were evaluated. A different random order of composites and corresponding mugshot albums were presented to each evaluator. By allowing evaluators to set the criterion for the initial set of possibles and have them decide how many mugshots they think represent the composite, one is able to measure precision as well as accuracy. A precise and accurate composite will be one where there is a small chosen field and the target is within that field.

The average number of mugshots selected from the 40-person mugshot albums was 5.34 (ranging from 1 to 27). During the entire mugshot procedure, a total of 47 hits out of 160 were made, amounting to an identification rate of 29%. As can be seen in Table 5, "cm2B" performed the best (50% identification rate), followed by "cm2A" (45% identification rate), and then "im2B" (35% identification rate). "Im2A" performed the worst, with an identification rate of only 10%.

Table 5: Average number of mugshots selected from each mugshot album including identification rates for each composite

T/Face	Valid N	Total Selected	Valid N	Hits	Identification Rate %
cm1A	40	4.15	20	3	15
cm1B	40	3.95	20	6	30
cm2A	40	5.6	20	9	45
cm2B	40	6.05	20	10	50
im1A	40	5.5	20	2	10
im1B	40	6.05	20	5	25
im2A	40	5.75	20	5	25
im2B	40	5.7	20	7	35

Table 6: Hypergeometric distribution results for each composite

Target	Hypogeometric distribution *p*
cm1A	0.20012
cm1B	0.18952
cm2A	0.03429*
cm2B	0.00017*
im1A	0.25230
im1B	0.10486
im2A	0.09251
im2B	0.01177*

The results of the mugshot test procedure are not good, especially considering the fact that the two highest rated reconstructions of each of the four target faces were evaluated during this procedure. However, as can be seen in Table 6 above, hypogeometric distribution results revealed significant probability values for the three composites with the highest identification rates (i.e. "cm2A", "cm2B" and "im2B"). Therefore one could argue that, out of the eight composites evaluated during the mugshot test procedure, there were three accurate composites, at least as defined as being selected above chance level in a large-choice task.

6.4. Discussion

As can be seen from the results of Experiment 1, it takes a relatively short period of time to produce a facial composite using the E-Face system (roughly 16 minutes on average) which is approximately five minutes quicker than that found in my previous investigation into the effectiveness of the E-Face system (Prag, 2000).

The E-Face system seems to produce reasonably satisfactory composites as participants in this experiment gave them an average satisfaction rating of 68%. Participants commended the E-Face system for being easy to operate and having useful features, e.g. the ability to go backwards and forwards through generations during the reconstruction procedure, and the ability to refer to a history of previously chosen eigenfaces. They did however complain about the eigenfaces becoming too similar too quickly when progressing from one generation to the next, which probably explains the low number of average generations that participants made before making their final choice as to which eigenface best represent the target face. It took an average of approximately 11 generations of the E-Face system to arrive at the finally chosen composite, which is almost identical to the average number of generations found Prag's (2000) study. This probably accounts for the relatively short period of time used on average to produce E-Face composites in this experiment and in Prag's (2000) study.

All E-Face composites produced by participants in the reconstruction phase were given an average rating score of 47% by independent evaluators. This score is 3% greater than that found in Prag's (2000) study, in which an evaluation was made of the black and coloured E-Face databases. However, this rating score does not compare too favourably against the 72% found in Tredoux et al.'s (1999) first experiment of their evaluation of the first version of E-Face, in which "in view" reconstructions were made of three white male targets.

The thresholding results, which allow one to infer whether or not the target would have been selected from a 5-person lineup by the evaluator, revealed that all composites produced by participants in the reconstruction phase received an overall identification rate of 51%. This result is considerably better than the 26% found in Prag's (2000) study, and is more than double the naming rate found for Frowd's (2001) "in view" EvoFIT reconstructions and Brace et al.'s (2000) "in view" E-Fit reconstructions of famous faces. The thresholding results of this experiment also compare very favourably against the low 10% naming rate found when "in view" E-Fit reconstructions were made of famous faces in an experiment conducted by Davies and Oldman (1999). Interestingly, the same identification rate of 51% for this experiment was found in Tredoux et al.'s (1999) first experiment, where judges were required to select out of a 7-person photo lineup, which face most resembled the "in view" E-Face reconstructions of three white male targets.

The results of the second evaluation task for this experiment, the mugshot test procedure, revealed an overall identification rate of 29%, which is in line with the majority of the above-mentioned studies. This is not an especially good identification rate, especially considering that the two highest rated reconstructions of each of the four targets were evaluated during this procedure. However, further statistical analyses indicate that out of the eight composites evaluated during this procedure, there were three accurate composites (defined as being selected above chance level in a large-choice task).

The identification rates found in this experiment does not, however, compete with the 83% matching accuracy found when novice "in view" E-Fit reconstructions were made of familiar faces (Davies et al., 2000), and the 81% matching accuracy found when novice "in view" Whatsisface reconstructions were made (Gillenson and Chandrasekeran, 1975). Despite this, it can be seen that the E-Face system, in general, compares quite well against previously conducted studies where novice operators have used various facial composite systems (including the E-Face system) to produce facial composites under "in view" conditions.

No significant differences were found between the rating scores and identification rates of the four targets, therefore no conclusions can be made as to which of the four target faces were the easiest or most difficult to reconstruct. However, the fact that different mean scores were obtained each of the four target faces implies that certain faces are more difficult to reconstruct than others. In order to minimise and/or eliminate this problem, the faces that make up the E-Face databases need to be expanded so that they are more representative of all types of faces.

Even if this problem is resolved, as Ellis et al. (1975) quite rightly points out, people differ quite markedly in their ability to reconstruct faces. The variation in the accuracy of the composites produced indicates that certain participants have the ability to reconstruct faces better than others using this composite system, which suggests a clear witness effect. Some achieved reasonable likenesses to the original faces while others made quite poor reconstructions.

It is interesting to note that firstly, for the two coloured targets, the highest evaluation rating was obtained when the coloured database was used. Secondly, for the two Indian targets, the highest evaluation rating was obtained when the Indian database was used. Thirdly, for all targets except "cm2", the lowest evaluation rating was obtained when the Indian/coloured database was used. Therefore it would appear that better quality composites of coloured faces are produced using a purely coloured database, and better quality composites of Indian faces are produced using a purely Indian database. While the results of the rating task seem to support the hypothesis of the experiment, no such inference can be drawn since the interaction between "Target Face" and "Database" was found to be not significant in this experiment.

The thresholding results appear to disprove the hypothesis as they revealed that for all targets except "cm2", the highest identification rate was made when a database that did not correspond to the race of target face, was used. Based on these findings, a strong argument cannot be made for removing the Indian faces from the coloured database, and creating a new Indian database.

It is interesting to see how various methods of evaluation reveal different results, and even contradictory results, as in the particular case above. This can create some confusion as to which evaluation method to draw conclusions from; however, it would seem advisable to incorporate at least two evaluation tasks per experiment to provide some measure of comparison between findings. It has been said that a possible explanation for the poor performance of facial composite systems during laboratory testing is that the way in which the composite pictures were judged during experiments was inadequate and insensitive, and essentially unfair to the systems being investigated (Barber, 1988).

It has been consistently found in several eyewitness testimony studies that faces from another race are more difficult to recall and recognise than those from the person's own race. This phenomenon is often referred to as the "other-race effect". In a review of a number of previous studies on the other-race effect, Bothwell, Brigham and Malpass (1989) concluded that it was a significant factor in many eyewitness identification cases, and can account for between 10% and 12% of the variance in performance in face recognition studies. Similar conclusions were made in another meta-analysis carried through by Shapiro and Penrod (1986). Interestingly however, Lindsay and Wells (1983) found that recognition rates were higher for cross-race than for same-race identification when target-absent and target-present lineups were used. Despite these findings, there is general consensus amongst eyewitness identification researchers that people recognise faces of their own race more accurately than faces of other races.

Chance and Goldstein (1996) provide two explanations for the other-race effect. Firstly, they suggest that ethnic attitudes might impair or enhance people's recognition of faces of other races. Secondly, the amount of experience or familiarity people have with other racial groups has an impact on their face perception and memory processes. Also, as a result of stereotyping, people tend to assign more similar characteristics to members of other groups than they do to members of their own (Ainsworth, 1998).

An investigation into the other-race effect for this experiment, which was conducted separately for the reconstruction and evaluation phases, revealed that the bulk of the findings go against the general consensus that people recognise faces of their own race more accurately than faces of other races. The findings are instead more in line with those of Lindsay and Wells (1983), who found that recognition rates were higher for cross-race than for same-race identification.

According to Frowd (2001), it is of little forensic relevance to get participants to make "in view" reconstructions of faces (particularly of familiar faces) because in a real case scenario, composites are created of unfamiliar faces from memory. Composites produced with the target in view constitute a non-ecologically valid operational procedure (Frowd, 2001); therefore caution must be applied in making generalisations from this study. Nonetheless, the value of producing facial composites under "in view" conditions should not be underestimated, as these ideal conditions serve to test the optimality of facial composite systems.

Chapter 7: Experiment 2

The idea behind this experiment was to see how well the E-Face system performs against a facial composite system currently used by law enforcement agencies around the world. The Faces system was selected as the comparative facial composite system to the E-Face system. Various conditions and multiple comparisons were implemented in this experiment to allow for a thorough evaluation of these two systems. To test the optimality of both systems, expert users produced "in view" reconstructions of target faces. These reconstructions were evaluated against novice "in view" reconstructions. To test the two systems under more realistic and forensically important conditions, "from memory" reconstructions were made by novice users after being exposed to the targets in a live staged event. Prior to making these "from memory" reconstructions, participants either underwent the SAPS Interview or the Cognitive Interview (CI). This allowed for a comparison to be made between a standard police interview and an interview technique devised for the specific task of facial composite production.

7.1. Method

The experiment comprised two phases, (1) a reconstruction phase and (2) an evaluation phase. Within the reconstruction phase, there were four different conditions, two "from memory" and two "in view" conditions. In the two "from memory" conditions, participants reconstructed a target face between one and fifteen day(s) following a one-minute exposure to one of four targets during a live staged event. For the first "from memory" condition, participants underwent the SAPS Interview prior to reconstructing the target face; for the second "from memory" condition, participants underwent the CI prior to reconstructing the target face. The two "from memory" conditions were designed to evaluate the practical ability of the two systems under reasonably realistic conditions, so as to apply it within forensic contexts.

In the first "in view" condition, different participants reconstructed one of the four targets with photographs of the target face in full view during the composite production process, thereby assessing the fundamental ability of the composite systems under ideal conditions for novice users. In the second "in view" condition, two experts used the composite system that they had expertise in to make reconstructions of all four targets with photographs of the targets in full view during reconstruction, thereby assessing the two composite systems under absolute optimal conditions. Participants used one of two composite systems during all conditions of the reconstruction phase.

For the evaluation phase, independent judges meticulously evaluated the quality and utility of the resulting composites via three evaluation tasks, i.e. (1) the rating task, (2) the sorting task, and (3) the ranking task.

Participants

A total of 132 "cognitive psychology" students from the University of Cape Town (UCT) participated in this experiment. Participation in the experiment was a requirement for students taking the "cognitive psychology" course. Seventy-two students (15 males; 57 females) participated in the reconstruction phase. The average

age of the students was 21 years old (ranging from 19 to 26 years of age). Seventy-one percent of the students who participated in this phase belonged to the "white" population group.

Additional participants in the reconstruction phase were the two expert facial composite operators, those being Van Dijk and Tredoux. Van Dijk, a composite operator for the SAPS since 1998, is an expert in the Faces system. Tredoux has special expertise of the E-Face system since he has been extensively involved in developing and refining the system.

Sixty students (13 males; 47 females) participated in the evaluation phase, their average age being 20 years old (ranging from 19 to 24 year of age). Seventy-four percent of the evaluators belonged to the "white" population group. With regards to the race distribution of participants in both phases of the experiment, white people were deliberately selected as the majority as all the targets were white. This was done in the hope that floor effects would be minimised.

Materials

Two facial composite systems were evaluated during this experiment, i.e. (1) the E-Face system, more specifically the white database of this system (see Figure 1), and (2) the Faces system, more specifically the third version of this system (see Figure 2).

The four target faces were different-looking white males, aged between 18 and 21 years, and selected from a tertiary institution in Cape Town, called the "Cape Technikon". There were several reasons for targets being selected from this specific target group. Firstly, either white or black targets could be used since the feature libraries of the Face system (version 3) are not thoroughly representative of coloured and Indian faces. Secondly, only male targets could be used since the E-Face system does not yet have the ability to generate female composites. Thirdly, targets could only be between the ages of 18 and 21 because the E-Face system does not yet have the ability to reconstruct facial composites outside of this age range. Lastly, the reason behind selecting targets from "Cape Technikon" was to decrease the probability that their faces would be familiar to UCT students, who participated in the experiment.

The four targets were thoroughly informed about the nature of the research. They were told that by allowing the experimenter to take digital images of their faces from different angles, it would allow for the two facial composite systems to be evaluated under the "in view" condition. They were also made aware that their participation in the live staged events would form a crucial role in an evaluation of two facial composite systems under "from memory" conditions. The targets were paid R50.00 each.

As can be seen in Figures 23 to 26, a digital camera was used to take photographs of the Targets 1 to 4 from three different angles, i.e. (1) side profile, (2) three-quarter profile, and (3) front-on. These digital images were laser-printed in colour onto four high-quality, white A4 sheets of paper. Each sheet of paper had three photographs of one target face on it, each photograph (9cm x 8cm) taken from the three afore-mentioned angles. This was done so that participants in the "in-view" reconstruction condition had a clear idea of what the target faces looked like whilst reconstructing it.

Figure 23: Target 1

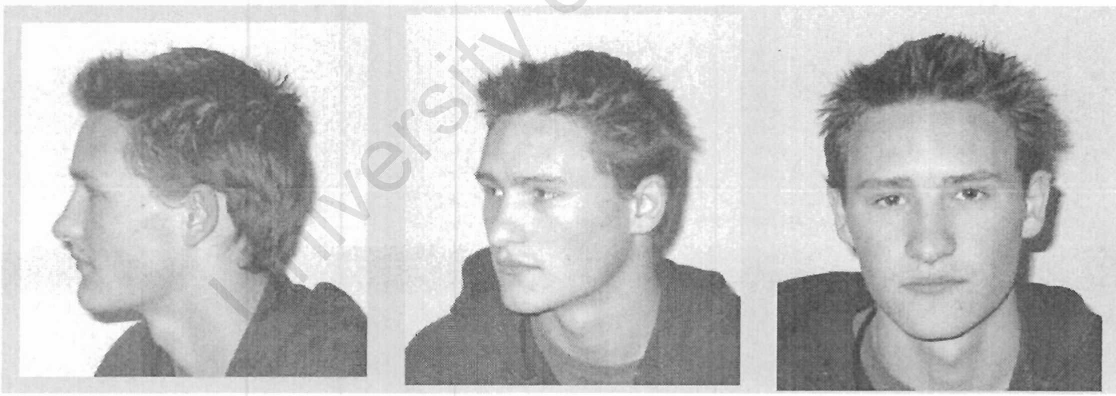


Figure 24: Target 2

Figure 25: Target 3

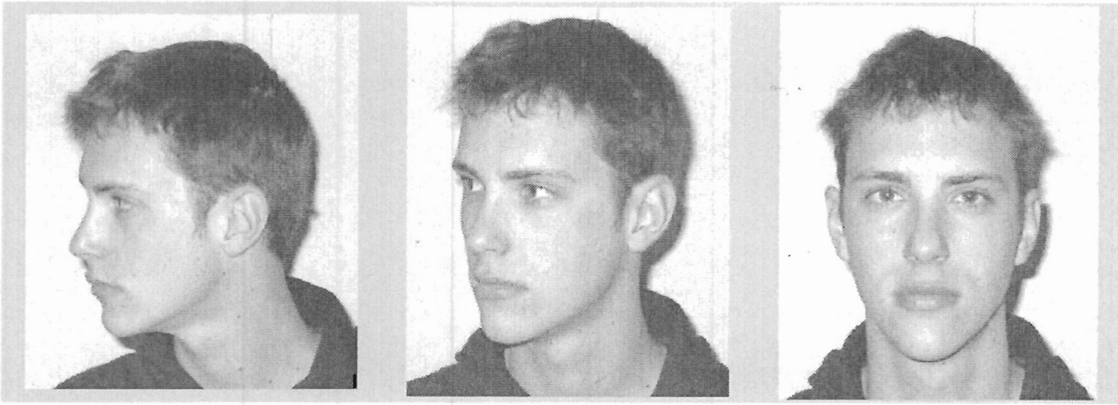
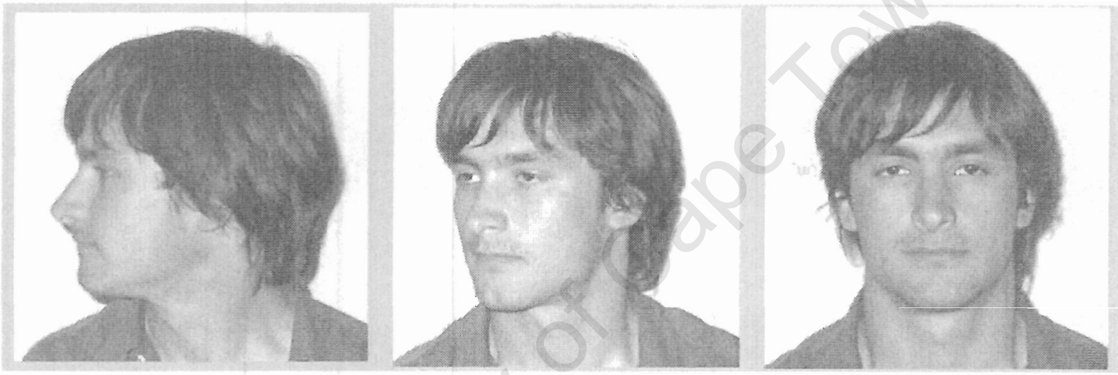


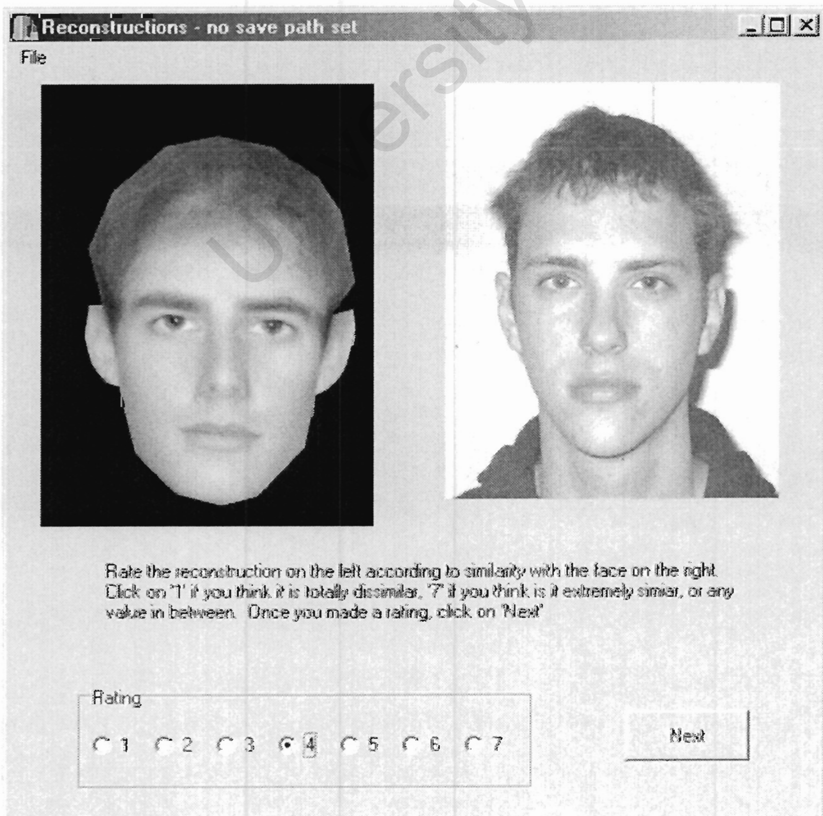
Figure 26: Target 4



As in Experiment 1, "Face Test" was used for the rating task (first task of the evaluation phase). The four target faces and 80 reconstructions made by the participants (including experts) were loaded into this specially developed programme in preparation for the rating task. Figure 27 shows a Faces reconstruction being rated in "Face Test", while Figure 28 shows an E-Face reconstruction being rated.

Figure 27: A Faces reconstruction being rated in "Face Test"

Figure 28: An E-Face reconstruction being rated in "Face Test"



For the sorting and ranking tasks (second and third tasks of the evaluation phase), the front-on digital photographs of the four targets used during the "in view" condition, were laser-printed in colour onto four high-quality, white A4 sheets of paper. Each photograph was 9cm x 8cm in dimension. The 80 reconstructions made by participants (including experts) in all conditions of the reconstruction phase were also laser-printed onto high-quality, white A4 sheets of paper. The E-Face reconstructions were printed in colour; the Faces reconstructions in black-and-white, since the Faces system is a black-and-white system. All printed reconstructions were cut out to produce 80 "composite cards" (9cm x 8cm) in preparation for the sorting and ranking tasks.

Reconstruction Phase

The 72 participants in the reconstruction phase were required to produce one facial composite using either the Faces system or the E-Face system. They were split evenly into three condition groups, thus 24 participants comprised each of the three condition groups, which were (1) "from memory" reconstructions after undergoing the SAPS Interview, (2) "from memory" reconstructions after undergoing the CI, and (3) "in view" reconstructions. Participants in the two "from memory" reconstruction conditions were required to make a reconstruction of the target face after being exposed to a live staged event, where the target was exposed to participants for roughly sixty seconds. Four live events involving the exact same scenario (of the target making an announcement to students in a classroom) were staged using the one of the four targets each time.

<u>Live Staged Events</u>

Forty-eight cognitive psychology students were required to sign up to one of four tutorial sessions, co-ordinated and run by the experimenter. Each of the four tutorial sessions comprised 12 students, and took place over the course of two days. These tutorial sessions were set up to stage the live events. Prior to attending the tutorial sessions, students were told that they would participate in a research experiment, however they were not told anything about the nature of the research.

At the beginning of each tutorial session, students were made to believe that the research experiment was about intelligence. They were given a pen-and-paper MENSA intelligence test (see Appendix A), which they were told to complete in 15 minutes. This distractor task was used to disguise the tasks that students would partake in during the actual experiment. About 10 minutes into the tutorial session, whilst students were in the process of completing their intelligence tests, one of the four targets entered the classroom. He politely interrupted the tutor (i.e. the experimenter), and asked to make an urgent announcement to students in the classroom, to which the tutor consented. Here follows the dialogue that took place between the target, tutor and students in the classroom during each of the four live staged events: -

Target: Hi everyone. I have lost my cell-phone. I think that it is somewhere in this building. It is very important that I get the cell-phone back as soon as possible, because I have stored many important numbers on it. It has even got my bank details stored on it.

Tutor: What kind of cell-phone is it?

(The target provides a brief description of the cellular phone.)

Tutor: Have any of you students seen the cell-phone?

(The students respond by saying that they had not seen the cell-phone.)

Tutor: The cell-phone is not perhaps in this classroom, is it?

(The students look around the classroom to see if they could find it.)

Target: I do not think that it will be in this classroom because I came to the psychology building this morning just to meet one of the lecturers, not to attend a class. If anyone of you finds the cell-phone, please give it to Sheila, the secretary of the psychology department. I am offering a reward of R100 to the person who finds it. Thank you for your time. Goodbye.

(The target leaves the classroom.)

Each of the four targets rehearsed the event with the tutor prior to the actual staging of it. The targets were instructed to stand at the front of classroom when making their announcements so that each student in the classroom could get a good look at them. The targets were urged to make eye contact with each student in the classroom to allow for sufficient interaction between the target and the students. They were also encouraged to make the event seem as realistic and plausible as possible. Each of the four live staged events lasted approximately sixty seconds. After the live staged event, the students continued completing their intelligence tests for roughly five minutes more, after which the tutor collected their test papers. At the end of the tutorial session, the students were instructed to sign up for a follow-up session. They were told that this follow-up session would comprise a one-to-one interview (between each student and the tutor) regarding their MENSA intelligence tests.

The follow-up sessions took place in a quiet interview room to avoid distractions and to promote focused concentration. They occurred between one and 15 day(s) after students were exposed to one of the four targets in the live staged events. The 12 students in each of the four tutorial sessions were equally divided into two groups according to two conditions of the reconstruction phase, i.e. (1) "from memory" reconstructions after undergoing the SAPS Interview, and (2) "from memory" reconstructions after undergoing the CI. Therefore 24 students participated in each of these two conditions, which were administered during the follow-up sessions.

The SAPS Interview condition

Once seated in the interview room, the tutor revealed to each of the 24 students individually that the research experiment was not at all about intelligence testing (as they were made to believe), but that it was in fact related to the evaluation of two facial composite technologies under varying conditions. It was further revealed that the tutor was actually the experimenter, and that the intelligence testing acted as a mere disguise. They were told that the actual reason for setting up the tutorial sessions was to stage live events, which would allow for brief target exposure to students in the classroom.

Once the truth was revealed to the students, they were thoroughly informed about the nature of the research. They were then told that they were about to undergo a standard interview used by the SAPS intended to help witnesses or victims of a crime recall the face of their perpetrator. Following the SAPS Interview, they were informed that they would be instructed to use one of two facial composite systems to produce a reconstruction of the target that they were exposed to during the live staged event. Following this briefing of their impending tasks, the experimenter gave a full demonstration of how to operate the composite system that they were to use, based on structured guidelines stipulated in the respective training manuals of both composite systems.

Participants who used the (featural) Faces system were told that the order in which they selected the various facial features was up to them to decide, however they were advised to first select the feature that they remembered most distinctly when building the facial composite, and then move on to the next salient feature, and so on. Participants who used the (configural) E-Face system were advised not to look at a single feature (or features) of the eigenfaces they were presented with in each generation, but to rather look at each eigenface as a whole when deciding which one to select.

Each participant was given an opportunity to ask any questions at this point. Once they felt confident that they understood how the facial composite system operated, they were taken through the SAPS interview: -

v) The experimenter first asked each participant to provide a description of the entire live staged event, from the moment that the target entered the classroom until he left the classroom.

vi) Thereafter the experimenter asked each participant to provide a description of the target. The experimenter prompted each participant by a series of verbal descriptors about the target, i.e. race, age, height, build, scars, tattoos, hair, eyes, nose, mouth, chin, teeth, facial hair, complexion, jewellery, hats and spectacles.

vii) The experimenter gave each participant the option of not describing those characteristics or features about the target that they were not able to recall.

Once the participants underwent the SAPS Interview, they proceeded in their efforts to reconstruct the target's face using the composite system demonstrated to them. They were notified that there were no time constraints for the task, and that they were allowed to make as many changes as they wished until satisfied with their reconstruction. Participants were allowed to ask the experimenter for assistance during the composite production process only if they were having trouble operating the composite system. In cases where assistance was required, the experimenter ensured that he did not influence participants' composite production process in any way.

Upon completion of their reconstructions, each participant was asked to rate the accuracy of their reconstructions to the target face using a 7-point Likert-type scale, 1 being "highly inaccurate" and 7 being "highly accurate". All reconstructions were saved as digital images onto a computer. The time taken for each participant to complete their reconstructions, and the number of generations made to arrive at the finally chosen eigenface (for those participants who used the E-Face system), was recorded.

Participants were then taken through a debriefing session where asked to express their thoughts and feelings about the task in general, about the usefulness of the SAPS Interview, about the strengths and weaknesses of the composites system they used, and about how the composite system they used could be improved. Before dismissing each student from the interview room, the experimenter pleaded with them not to provide any information whatsoever to their fellow students about the follow-up session or about the actual nature of the research experiment. It was stressed to them that without their co-operation, the results of the experiment would probably be distorted.

To summarise, 24 reconstructions were made during the SAPS Interview condition, i.e. six reconstructions of each of the four targets, half of which were made using the E-Face system, and the other half using the Faces system.

The CI condition

The procedure and instructions administered to the 24 participants in the CI condition were the same as those administered during the SAPS Interview condition, except that the participants underwent a CI instead of an SAPS Interview. Once the truth was revealed to the students about the intelligence testing and the live staged event, they were thoroughly informed about the nature of the research. They were then told that they were about to undergo a CI intended to help witnesses or victims of a crime recall the face of their perpetrator. Following the CI, they were informed that they would be instructed to use one of two facial composite systems to produce a reconstruction of the target that they were exposed to during the live staged event. Following this briefing of their impending tasks, the experimenter gave a full demonstration of how to operate the composite system that they were to use, based on structured guidelines stipulated in the respective training manuals of both composite systems.

Each participant was given an opportunity to ask any questions at this point. Once they felt confident that they understood how the facial composite system operated, they were taken through the CI, a slightly altered version of the original CI so that it could be used for the specific task of facial composite production: -

viii) In order to help the participant concentrate, the experimenter asked the participant to close his/her eyes or stare at some blank field in the interview room. No eye contact was to be made between the experimenter and the participant during the CI.

ix) The experimenter then asked the participant to play the entire live staged event in his/her head, from the moment that the target entered the classroom until he left the classroom - not to verbalise it, but merely think about it.

x) Whilst the participant did this, the experimenter asked the participant to think of the moment when he/she had the best view of the target.

xi) Once he/she has retrieved this best view, the experimenter asked him/her to think about his/her thoughts and feelings at that particular moment.

xii) Keeping this best view in mind, the experimenter asked the participant to search for trait judgments, labels, first impressions that he/she associated with the target.

xiii) The experimenter then asked the participant to focus intensely on the time when he/she had the best view of the target's face, so as to acquire a mental image of it.

Once the participants underwent the CI, participants proceeded in their efforts to reconstruct the target's face using the composite system demonstrated to them. The experimenter asked the participants to "refresh" that mental image of the target's face during the composite production process, more specifically after every other facial feature was selected from the feature library (of the Faces system) and/or after every other eigenface (of the E-Face system) was selected from the first generation to the last. This was done in the hope that it would minimise or prevent disruption of that mental image of the target face, caused by exposure to the feature library (of the Faces system) and/or the eigenfaces (of the E-Face system).

To summarise, 24 reconstructions were made during the CI condition, i.e. six reconstructions of each of the four targets, half of which were made using the E-Face system, and the other half using the Faces system.

<u>The Novice In-View (IV) condition</u>

The 24 students in this condition performed their tasks in a computer laboratory, four students at a time. These participants were not previously exposed to any of the targets, as opposed to those in the two "from memory" conditions. In this condition, the participants were told that their task was to use one of two facial composite systems to produce a reconstruction of one of the four targets. Participants were notified that, throughout the composite production process, they would have three photographs of the target person they were required to reconstruct in full view. Each of these photographs was taken from three different angles, i.e. (1) front-on, (2) side profile and (3) three-quarter profile (see Figures 22 to 25).

Following this briefing of their impending tasks, the experimenter gave a full demonstration of how to operate the composite system that they were to use, based on structured guidelines stipulated in the respective training manuals of both composite systems. Participants who used the (featural) Faces system were told that the order in which they selected the various facial features was up to them to decide. Participants who used the (configural) E-Face system were advised not to look at a single feature (or features) of the eigenfaces they were presented with in each generation, but to rather look at each eigenface as a whole when deciding which one to select.

Each participant was given an opportunity to ask any questions at this point. Once they felt confident that they understood how the facial composite system operated, they proceeded to reconstruct the target face, constantly referring to the photographs of the target face during the entire reconstruction procedure. As in conditions 1 and 2, participants were notified that there were no time constraints for the task, and that they were allowed to make as many changes as they wished until satisfied with their reconstruction. Participants were allowed to ask the experimenter for assistance during the composite production process only if they were having trouble operating the composite system. In cases where assistance was required, the experimenter ensured that he did not influence participants' composite production process in any way.

As in the two "from memory" conditions, upon completion of their reconstructions, each participant was asked to rate the accuracy of reconstructions to the target face using a 7-point Likert-type scale, 1 being "highly inaccurate" and 7 being "highly accurate". All reconstructions were saved as digital images onto a computer. The time taken for each participant to complete their reconstructions, and the number of generations made to arrive at the finally chosen eigenface (for those participants who used the E-Face system), was recorded. Participants were then taken through a debriefing session where asked to express their thoughts and feelings about the task in general, about the strengths and weaknesses of the composites system they used, and about how the composite system they used could be improved.

To summarise, 24 reconstructions were made during the Novice IV condition, i.e. six reconstructions of each of the four targets, half of which were made using the E-Face system, and the other half using the Faces system.

The Expert In-View (IV) condition

For the Expert IV condition, two expert facial composite operators were asked to make reconstructions of all of the four targets using the facial composite system they had expertise in. One operator was an expert in the Faces system, while the other an expert in the E-Face system. As in the Novice IV condition, the expert operators were notified that they would have three photographs of each of the targets in full view during their reconstruction procedures. Each of these photographs was taken from three different angles, i.e. (1) front-on, (2) side profile and (3) three-quarter profile (see Figures 22 to 25). Upon completion, the two expert operators were asked to rate the accuracy of their reconstructions to the target face using a 7-point Likert-type scale, 1 being "highly inaccurate" and 7 being "highly accurate". All expert reconstructions were saved as digital images onto a computer.

To summarise, 8 reconstructions were made during the Expert IV condition, i.e. two reconstructions of each of the four targets, half of which were made using the E-Face system, and the other half using the Faces system.

Evaluation Phase

In a separate phase of the experiment, the quality and utility of the 80 composites produced by participants (including experts) in the reconstruction phase was thoroughly assessed. Sixty evaluators individually participated in three evaluation tasks, 20 evaluators per task. Different evaluators were used for each task to avoid the risk of contamination. These three tasks were (1) a rating task, (2) a sorting task and (3) a ranking task. Each task was self-paced. Prior to performing their tasks, evaluators were informed about the nature of the research and about how their participation contributed to achieving the aims and objectives of the experiment.

Task 1: Rating task

For the rating task, evaluators used a specially developed computer programme for rating faces called "Face Test". They were seated in front of a computer screen and were presented with two facial images next to each other, each image 5cm x 7cm in dimension. Beneath these two facial images, were the following instructions, "Rate the reconstruction on the left according to similarity with the face on the right. Click on "1" if you think it is totally dissimilar, "7" if you think it is extremely similar, or any value in between. Once you made a rating, click on 'Next'". Beneath these instructions were numbers 1 to 7 and the "next" button (see Figures 26 and 27).

The experimenter made it known to the evaluators that the image on the left would always be a facial composite produced by participants in the reconstruction phase using one of two composite systems, while the face on the right would always represent the corresponding target face, i.e. the face that participants in the reconstruction phase were trying to reconstruct. When deciding on their ratings, the evaluators were told to look at each face as a whole, rather than focus on specific features.

Each evaluator rated each of the 80 reconstructions against its corresponding target face, thus a total of 80 ratings were made. Each evaluator rated a different random order of reconstructions.

Task 2: Sorting task

For the sorting task, evaluators were provided with a pack of 80 composite cards (high-quality 9cm x 8cm prints of the 80 facial composites produced in the reconstruction phase) and a front-on, 9cm x 8cm, colour photograph of each of the four targets. Evaluators were told that 20 composite cards represented each of the four target faces. Their task was to sort the 80 composite cards according to what they perceived to be the matching target face. Evaluators were encouraged to guess should they experience any uncertainty during the sorting task. The pack of 80 composite cards was shuffled prior to each evaluator performing their sorting task, thus each evaluator sorted a different random order of composite cards.

After each evaluator completed their sorting task, the experimenter made a record (using the numbers that appeared on the back of each composite card) of which facial composites were sorted to each target face. Each facial composite received a score of either 0 or 1 depending on whether or not an evaluator correctly assigned the composite card to the matching target face. To clarify, a composite that was correctly assigned to the matching target face by the evaluator received a score of 1, whereas a composite that was not assigned to the matching target face received a score of 0.

This forced-choice sorting procedure is quickly administered and has been shown to correlate significantly with more direct measures of identification accuracy such as the ability to identify a face in a large mug album on the basis of a single composite (Christie et al., 1981). It is also the most ecologically valid and realistic of the three evaluation tasks.

<u>Task 3: Ranking task</u>

The pack of 80 composite cards and the front-on photographs of each of the four targets used in the sorting task, were used in the ranking task. Prior to evaluators beginning their ranking task, the experimenter sorted the 80 composite cards according to their matching targets. In other words, evaluators were presented with four stacks of composite cards laid out on a table, each stack comprising 20 composite cards. Above each of the four stacks of composite cards, were the matching target faces.

Evaluators were required to scrutinise each composite card relative to the photograph of the target and then rank order each of the four stacks of composite cards from most similar-looking to least similar-looking (to the matching target faces). The four stacks of 20 composite cards were shuffled prior to each evaluator performing their ranking task, therefore each evaluator ranked a different random order of composite cards. After each evaluator completed their ranking task, the experimenter made a record of their rankings, using the numbers that appeared on the back of each composite card. Each facial composite received a score between 1 and 20 depending on where they were ranked by evaluators, e.g. a composite ranked fifth out of 20 by an evaluator would be given a score of 5.

7.2. Results

Results for conditions 1, 2 and 3 (SAPS Interview, CI and Novice IV) were compared against each other, to make a comparison between "in view" composites and "from memory" composites. The Expert IV composites were deliberately excluded from this comparison since novice operators produced the "from memory" composites. Results for conditions 3 and 4 (Novice IV and Expert IV) were compared against each other to see how "in view" composites produced by expert operators compared against those produced by novice operators. Results for the reconstruction phase and the evaluation phase are reported separately for both comparisons.

Reconstruction Phase (SAPS Interview v CI v Novice IV)

For the first three experimental conditions, participants took an average of 25.68 minutes to reconstruct each composite; recorded times ranged from 8 to 55 minutes. With the E-Face system, participants took an average of 34.86 generations to arrive at their finally chosen composite; recorded generations ranged from 15 to 60. Upon completion of the reconstructions, participants were asked on a 7-point Likert-type scale how accurate they thought their reconstructions were to the target face, 1 being "highly inaccurate" and 7 being "highly accurate". All reconstructions made by participants received an average satisfaction rating of 3.67 out of 7, which equates to 52%. Table 7 below details mean reconstruction times and mean operator satisfaction ratings, while Table 8 shows the mean number of E-Face generations used to arrive at the finally chosen composite.

Table 7: Mean reconstruction times (in minutes); and mean operator satisfaction ratings (out of 7)

				Reconstruction Times				Satisfaction Ratings			
T/Face	Reconstructor	Condition	Valid N	Mean	Min	Max	Std. Dev.	Mean	Min	Max	Std. Dev.
			72	25.68	8	55	11.57	3.67	1	6	1.17
Target 1			18	24.94	8	55	12.85	4.11	2	6	1.08
Target 2			18	24.78	8	49	12.89	3.56	2	6	1.29
Target 3			18	23.83	8	44	9.26	3.17	1	6	1.29
Target 4			18	29.17	12	52	11.14	3.83	2	5	0.86
	Faces		36	26.69	8	53	12.06	3.56	1	6	1.23
	E-Face		36	24.67	8	55	11.14	3.78	2	6	1.12
		SAPS	24	21.75	8	53	11.73	3.63	2	6	1.17
		CI	24	22.83	8	55	11.88	3.42	1	6	1.35
		Novice IV	24	32.46	18	52	7.87	3.96	2	6	0.95

Table 8: Mean number of generations made during E-Face reconstructions

T/Face	Condition	Valid N	Mean	Minimum	Maximum	Std. Dev.
		36	34.86	15	60	12.74
Target 1		9	38.00	27	60	11.17
Target 2		9	30.89	15	60	14.68
Target 3		9	34.78	25	55	10.35
Target 4		9	35.78	15	60	15.25
	SAPS	12	34.67	25	57	10.05
	CI	12	41.50	25	60	14.51
	Novice IV	12	28.42	15	43	10.55

Verbal feedback

The verbal feedback from participants in the reconstruction phase is reported separately for the E-face system, the Faces system, and the two "from memory" conditions.

E-Face system

As in Experiment 1, it was unanimously agreed amongst participants in the reconstruction phase that the E-Face system was user-friendly, and that no extensive computer expertise was necessary to operate it. Participants found it useful to have a history of each face selected in each generation to refer to, as this assisted them in making their final selection. The ability of the system to go backwards and forwards during the reconstruction procedure was also commended by participants, especially in situations when they accidentally made a selection, or when they were not satisfied with an eigenface selected in a particular generation. Participants were also impressed with the quality of the outcome, and the speed at which a facial composite could be created with this system.

As in Experiment 1, many participants however found the task of producing E-Face composites more challenging and confusing as they progressed through the reconstruction procedure. This was because the eigenfaces, which they were required to select from, became more and more similar as they proceeded from one generation to the next. Also, by being exposed to so many eigenfaces during the task, some participants found themselves drifting away from their mental images of the target face. There were also a couple of complaints about the eigenfaces becoming too similar too quickly.

In terms of how the E-Face system could be improved, many participants suggested that there be a facility to manipulate and refine the finally chosen face. By importing the finally chosen face into a picture editing or freehand drawing programme, for example, one could make facial refinements (like stretching, narrowing, and shifting facial features, adjusting face shape, etc.) and add specific detail (like scars, ageing lines, etc.). Participants also suggested that it would easier to make a final selection if the history of faces selected in each generation could be viewed at a single glance, instead of having to scroll down a narrow panel of faces (as it is currently set up). Also, to improve the E-Face system, participants would like to see the quality of certain eigenfaces, which they were required to choose from, be improved so that they look less distorted and more realistic.

Other suggested improvements were an increase in the thumbnail size of each eigenface; a corresponding side profile to each eigenface; more variation in hairstyle, eye and hair colour; and more variation in eigenfaces so that they do not become too similar too quickly. Some participants, who remembered the target looking worried during the live staged event, would like to see the E-Face system have the ability to add facial expressions to the finally chosen facial composites.

Faces system

Participants had several positive comments to make about the Faces system. It was unanimously agreed amongst participants in the reconstruction phase that the Faces system was user-friendly, and that no extensive computer expertise was necessary to operate it. Participants commended the system on the detail that could be added to the facial composite, i.e. facial hair, eye lines, smile lines, mouth lines and chin lines, glasses and other accessories. They also appreciated the manipulations that could be made to selected features of the facial composite e.g. moved further apart, closer together, deleted, reduced / enlarged in size, and moved up and down with the aid of the position slider. Participants found it useful that the Faces system kept a record of all selected facial features (by highlighting the selected feature codes in red once selected) as this allowed them to go back and choose the most appropriate features to make up the facial composite. Participants were also impressed with the quality of the outcome, and the speed at which a facial composite could be created with this system.

However, participants using the Faces system experienced some difficulties too. Some participants commented that there were too many options to choose from in the feature libraries, which made the task challenging and confusing. Others complained that the feature library was not representative of all facial features, as they could not locate certain facial features. Quite a number of participants commented that the selection of mouths were very feminine. This made it difficult to choose appropriate mouths for facial composites of the male targets.

In terms of how the Faces system could be improved, participants suggested that colour be added to the current black-and-white system so that eye and hair colour can be distinguished, for example. Other suggested improvements included increasing the spread of masculine lips and hair types, having the facility to shorten and lengthen a particular hair type, and the facility to add facial expressions to the facial composites.

SAPS Interview Condition

Participants who underwent the SAPS Interview generally thought that it assisted them in recalling the target face, especially when prompted by the interviewer through the series of verbal descriptors of the target. Many participants found that they had a limited vocabulary for verbal descriptions, which resulted in them providing a quite brief description or no description at all of a particular verbal descriptor. There were a few participants who could not recall any information at all about certain verbal descriptors.

CI condition

Participants who underwent the CI thought that it was a good technique to aid recall of the target face. The CI generally helped participants to focus and thus gain a clearer mental image of the target face. The constant request by the interviewer to get participants to "refresh" the best view of the target face was found to be effective in getting the majority of participants to maintain their mental images of the target face throughout the composite production process. There were a few participants who did not find the CI particularly helpful in recalling the target face. For these participants, the CI helped to recall specific details during the live staged event, like the voice and body language of the target, for example.

In general, participants found that it required extensive amounts of concentration and focus to carry through the reconstruction task. Participants who had experienced a long delay between target exposure and target reconstruction found this task quite demanding. Also, the fact that the targets were exposed to the participants for a relatively short period of time (approximately one minute) did not make the task any easier. Nevertheless they were intrigued by both facial composite systems. Overall, participants were impressed with the way the live staged event was executed, commenting that it was well planned, well acted and completely unsuspicious.

Evaluation Phase (SAPS Interview v CI v Novice IV)

For the evaluation phase, the results of the first three experimental conditions are separately reported for the three evaluation tasks of the experiment, i.e. (1) the rating task, (2) sorting task and (3) ranking task.

<u>Task 1: Rating Task</u>

For the rating task, evaluators were asked to use a 7-point Likert-type scale to rate the composites produced by participants in the reconstruction phase against the target faces they were trying to reconstruct. Ratings were made according to similarity, 1 being "extremely dissimilar" and 7 being "extremely similar".

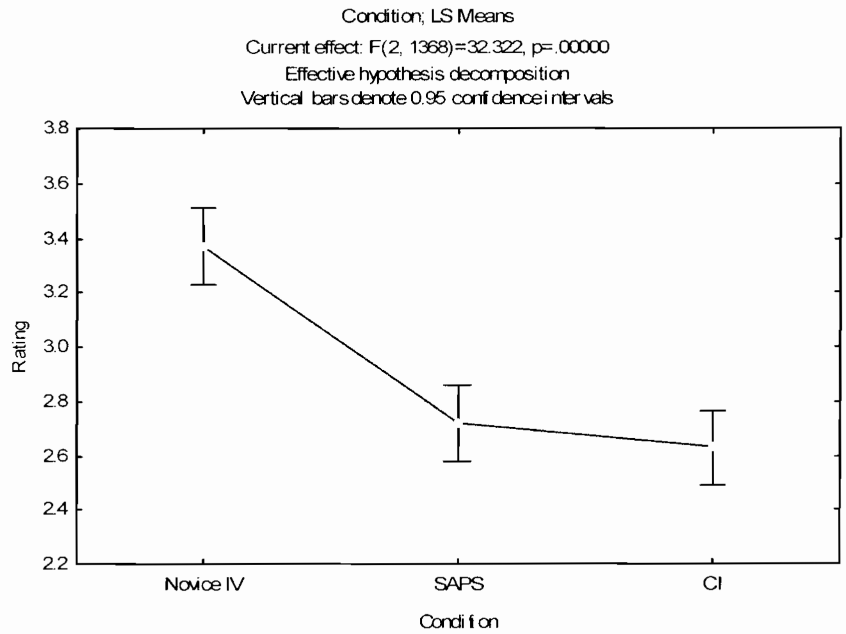
Table 9: Four-way factorial ANOVA rating results

Source	*SS*	*df*	*MS*	*F*	*p*
Intercept	12151.03	1	12151.03	4963.234	0.000000*
(1)Reconstructor	1.06	1	1.06	0.431	0.511395
(2)Condition	158.26	2	79.13	32.322	0.000000*
(3)Target	28.02	3	9.34	3.816	0.009733*
(4)C.Number	1.06	2	0.53	0.216	0.805418
Reconstructor*Condition	107.34	2	53.67	21.922	0.000000*
Reconstructor*Target	37.52	3	12.51	5.109	0.001617*
Condition*Target	44.67	6	7.45	3.041	0.005860*
Reconstructor*C.Number	20.50	2	10.25	4.188	0.015378*
Condition*C.Number	29.04	4	7.26	2.965	0.018747*
Target*C.Number	29.49	6	4.92	2.008	0.061736
Reconstuctor*Condition*Target	55.01	6	9.17	3.745	0.001060*
1*2*4	18.00	4	4.50	1.838	0.119149
Reconstructor*Target*C.Number	33.39	6	5.57	2.273	0.034555*
Condition*Target*C.Number	92.04	12	7.67	3.133	0.000205*
1*2*3*4	73.41	12	6.12	2.499	0.003025*
Error	3349.15	1368	2.45		

All reconstructions made by participants received an average rating score of 2.90 out of 7, which equates to 41%. (See Appendix F for the similarity rating means for the reconstructions made under the first three experimental conditions). As can be seen in Table 9 above, the four-way factorial ANOVA results show four main effects. Although the main effect for "Composite Number" is an important one, it does not aid the interpretation of the key experimental effects, thus it has been dropped in the results analysis. The main effect for "Reconstructor" was not significant, F (1, 1368) = 0.43, $p > 0.05$. Faces composites (M = 2.93) performed marginally better than E-Face composites (M = 2.88).

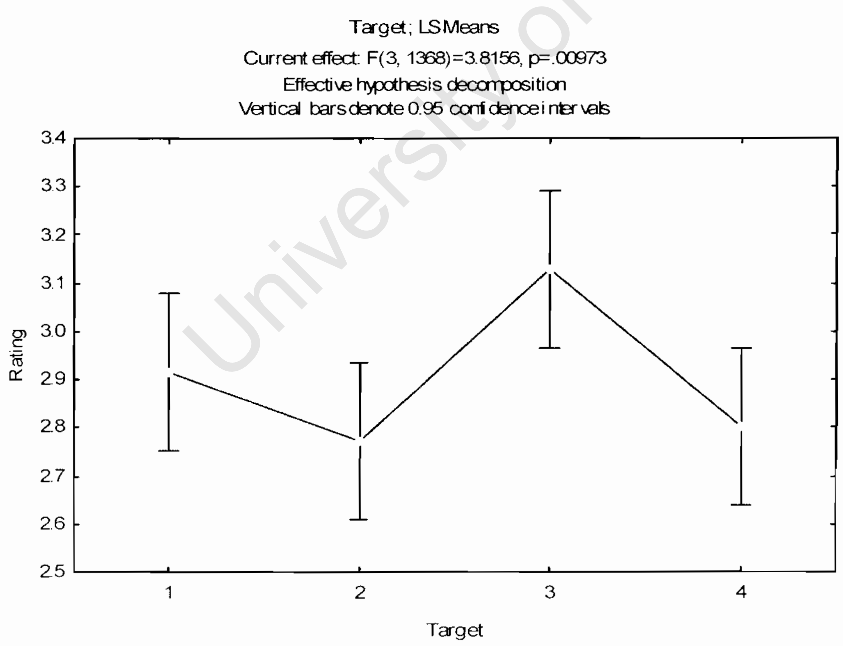
The main effect for "Condition" was significant, F (2, 1368) = 32.32, $p < 0.01$. Post-hoc analysis using Tukey HSD comparison tests revealed that Novice IV composites (M = 3.37) performed significantly better than SAPS composites (M = 2.72) and CI composites (M = 2.63). SAPS composites were found to be slightly better than CI composites: however there was no significant difference in their mean rating scores. Figure 29 below presents the three condition rating means graphically.

Figure 29: "Condition" rating means (out of 7)



The main effect for Target was significant, F (3, 1368) = 3.82, $p < 0.01$. Post-hoc analysis using Tukey HSD comparison tests revealed that Target 3 composites (M = 3.13) received the highest ratings on average, significantly higher than Target 2 composites (M = 2.77) and Target 4 composites (M = 2.80). Target 1 composites (M = 2.92) came in second. Figure 30 below presents the four target rating means graphically.

Figure 30: "Target" rating means (out of 7)

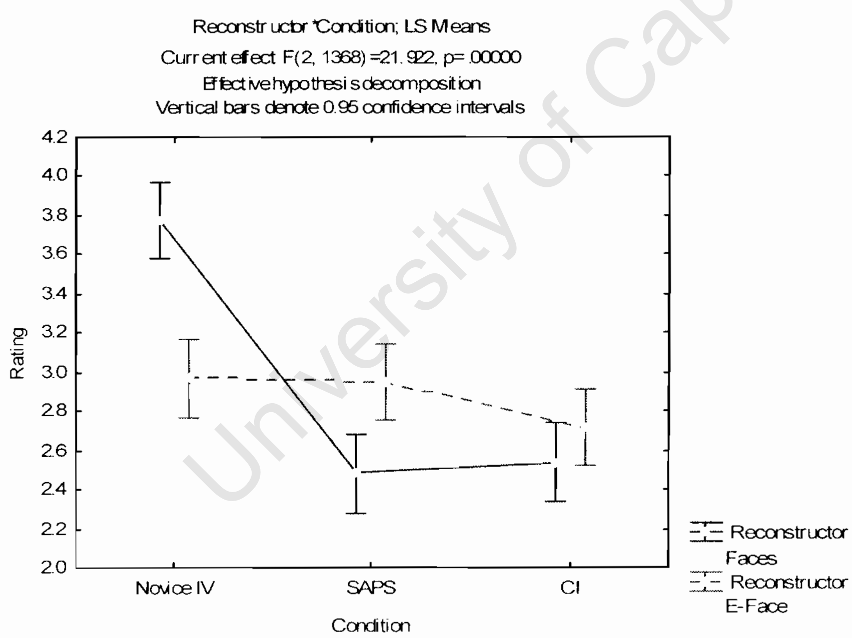


The interaction between "Reconstructor" and "Condition" was significant, $F(2, 1368) = 21.92$, $p < 0.01$. As can be seen in Table 10 below, the simple effects analysis reveals that there was a significant interaction between "Condition" and the Faces system, $F(2, 1368) = 52.29$, $p < 0.01$, such that Novice IV composites (M = 3.78) performed significantly better than SAPS composites (M = 2.48) and CI composites (M = 2.54) for this reconstructor. There was also a significant interaction between "Reconstructor" and the SAPS condition, $F(1, 1368) = 10.67$, $p < 0.01$, such that Faces composites (M = 2.48) received considerably lower mean ratings than E-Face composites (M = 2.95) in this condition. Figure 31 below presents the "Reconstructor x Condition" rating means graphically.

Table 10: "Reconstructor (R) x Condition (C)" interaction

Source	*SS*	*df*	*MS*	*F*	*P*
R at Novice IV	78.408	1	78.408	32.003	0.000000*
R at SAPS	26.133	1	26.133	10.667	0.001797*
R at CI	3.852	1	3.852	1.572	0.225169
C at E-Face	9.378	2	4.689	1.914	0.161600
C at Faces	256.219	2	128.110	52.290	0.000000*
Error	3349.15	1368	2.45		

Figure 31: Rating means for "R x C" (out of 7)

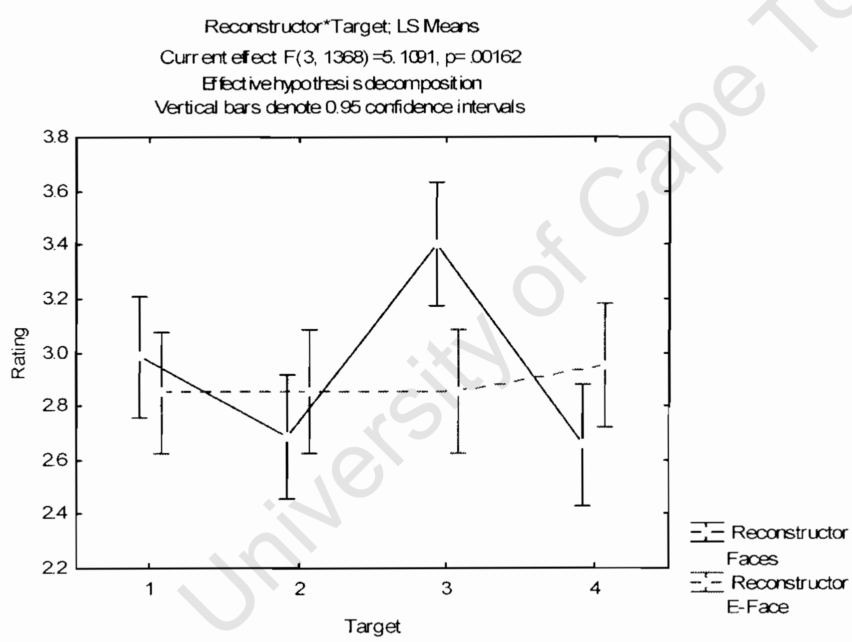


The interaction between "Reconstructor" and "Target" was significant, $F(3, 1368) = 5.11$, $p < 0.01$. As can be seen in Table 11 below, the simple effects analysis reveals that there was a significant interaction between "Reconstructor" and Target 3, $F(1, 1368) = 10.89$, $p < 0.01$, such that Faces composites ($M = 3.40$) performed significantly better than E-Face composites ($M = 2.86$) of Target 3. There was also a significant interaction between "Target" and the Faces system, $F(3, 1368) = 8.75$, $p < 0.01$, such that Faces composites of Target 3 ($M = 3.40$) performed significantly better than those of Target 2 ($M = 2.69$) and Target 4 ($M = 2.66$). Figure 32 below presents the "Reconstructor x Target" rating means graphically.

Table 11: "Reconstructor (R) x Target (T)" interaction

Source	SS	df	MS	F	P
R at T1	1.600	1	1.600	0.653	0.428057
R at T2	2.500	1	2.500	1.020	0.341078
R at T3	26.678	1	26.678	10.889	0.002374*
R at T4	7.803	1	7.803	3.185	0.111460
T at E-Face	1.256	3	0.419	0.171	0.921817
T at Faces	64.293	3	21.431	8.747	0.000106*
Error	3349.15	1368	2.45		

Figure 32: Rating means for "R x T" (out of 7)



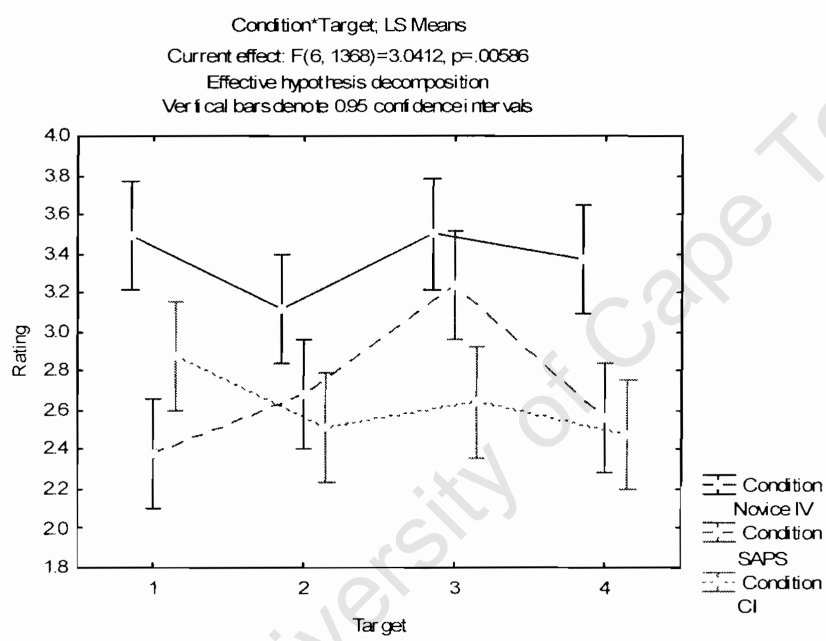
The interaction between "Condition" and "Target" was significant, $F(6, 1368) = 3.04$, $p < 0.01$. As can be seen in Table 12 below, the simple effects analysis reveals that there was a significant interaction between "Condition" and all four targets. There was also a significant interaction between "Target" and the SAPS condition, $F(3, 1368) =$

6.74, $p < 0.01$, such that Target 3 composites (M = 3.24) received considerably higher ratings than Target 1 composites (M = 2.38) and Target 4 composites (M = 2.56) in this condition. Figure 33 below presents the "Condition x Target" rating means graphically.

Table 12: "C x T" interaction

Source	*SS*	*Df*	*MS*	*F*	*p*
C at T1	74.017	2	37.008	15.105	0.000000*
C at T2	23.022	2	11.511	4.698	0.014824*
C at T3	46.539	2	23.269	9.498	0.000297*
C at T4	59.356	2	29.678	12.113	0.000052*
T at Novice IV	11.508	3	3.836	1.566	0.260614
T at SAPS	49.550	3	16.517	6.741	0.000329*
T at CI	11.640	3	3.880	1.584	0.216817
Error	3349.15	1368	2.45		

Figure 33: Rating means for "C x T" (out of 7)

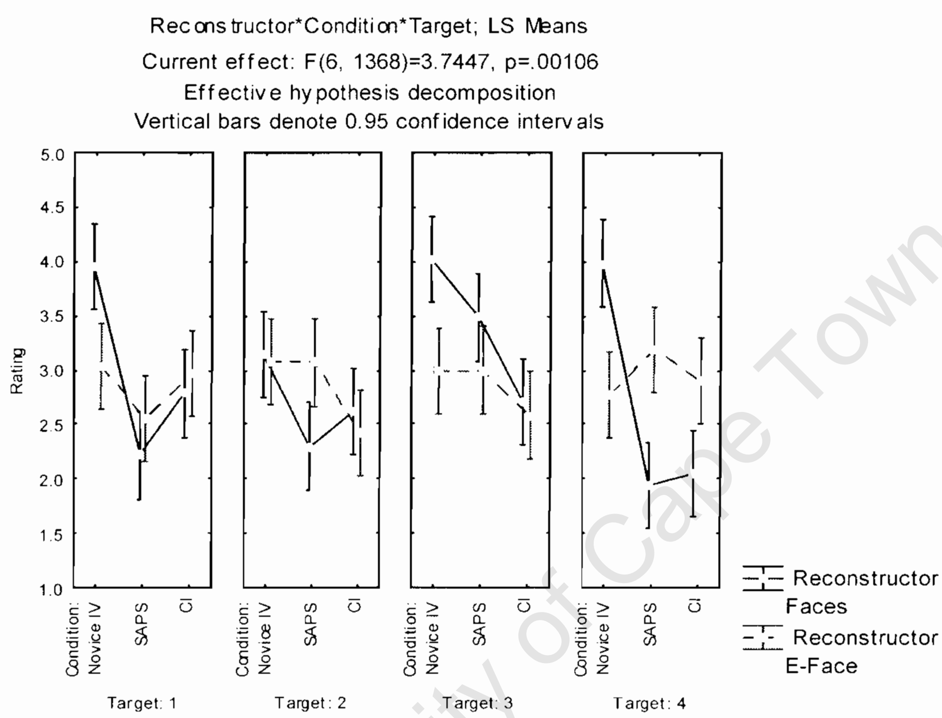


The three-way interaction between "Reconstructor", "Condition" and "Target" was significant, F (6, 1368) = 3.75, $p < 0.01$. As can be seen in Table 13 below, the simple interaction effects analysis reveals that there was a highly significant interaction between "Reconstructor" and "Condition" for Target 1 composites, F (2, 1368) = 5.70, $p < 0.01$; for Target 2 composites, F (2, 1368) = 3.36, $p < 0.05$; and for Target 4 composites. F (2, 1368) = 21.46, $p < 0.01$. Figure 34 below presents the "RC x T" rating means graphically.

Table 13: "RC x T" interaction

Source	*SS*	*df*	*MS*	*F*	*P*
RC at T1	27.950	2	13.975	5.704	0.002430*
RC at T2	16.467	2	8.233	3.361	0.047047*
RC at T3	12.772	2	6.386	2.607	0.096421
RC at T4	105.156	2	52.578	21.460	0.000000*
Error	3349.15	1368	2.45		

Figure 34: Rating means for "RC x T" (out of 7)



The best-rated composites for Targets 1, 3 and 4 were created using the Faces system, under the Novice IV condition. They received evaluation ratings of 4.30, 4.65 and 4.95 (out of 7) respectively. The composite that achieved the best rating for Target 2, with an evaluation rating of 3.75 out of 7, was created using the E-Face system, under the SAPS Interview condition (see Appendix G and Figure 35).

Figure 35: Best-rated composites (bottom) of each target face (top)

Task 2: Sorting Task

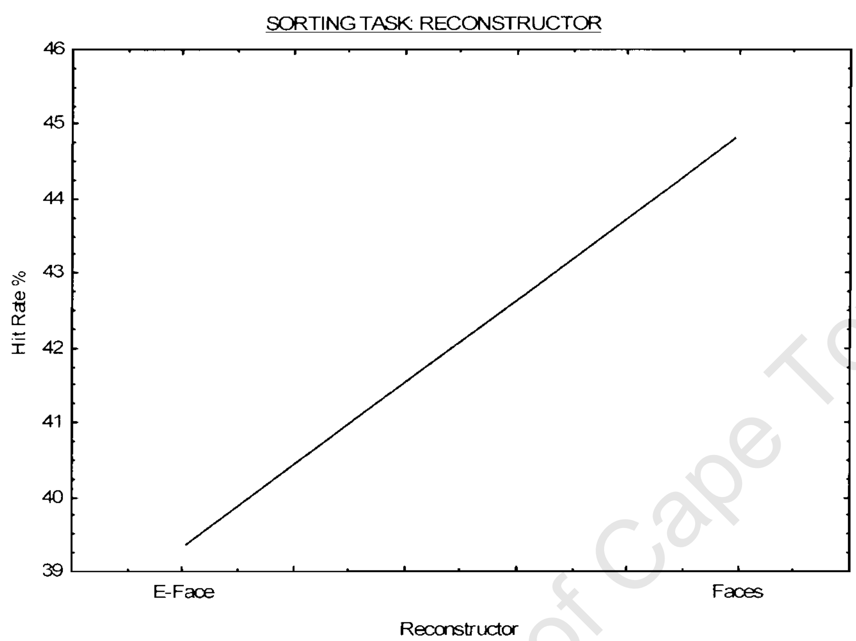
For the sorting task, evaluators were to sort the facial composites produced by participants in the reconstruction phase according to what they perceived to be the matching target face. Each of the 80 facial composites sorted by the evaluators received a score of either 0 or 1, depending on whether or not they were correctly assigned to the matching target face by the evaluators.

Table 14: Sorting logistic regression results

	χ^2	*df*	*P*	χ^2 difference from full model	*p*
Reconstructor	4.5614	1	0.03271*	39.91338	0.000000*
Condition	38.847	1	0.00000*	5.62741	0.059997
Target	0.91191	1	0.33962	43.56285	0.000000*
Reconstructor x Condition	43.535	2	0.00000*	0.940103	0.332259
Reconstructor x Target	5.4762	2	0.06471	38.99857	0.000000*
Condition x Target	39.784	2	0.00000*	4.690382	0.030339*
Reconstructor x Condition x Target	44.475	3	0.00000*	-	-

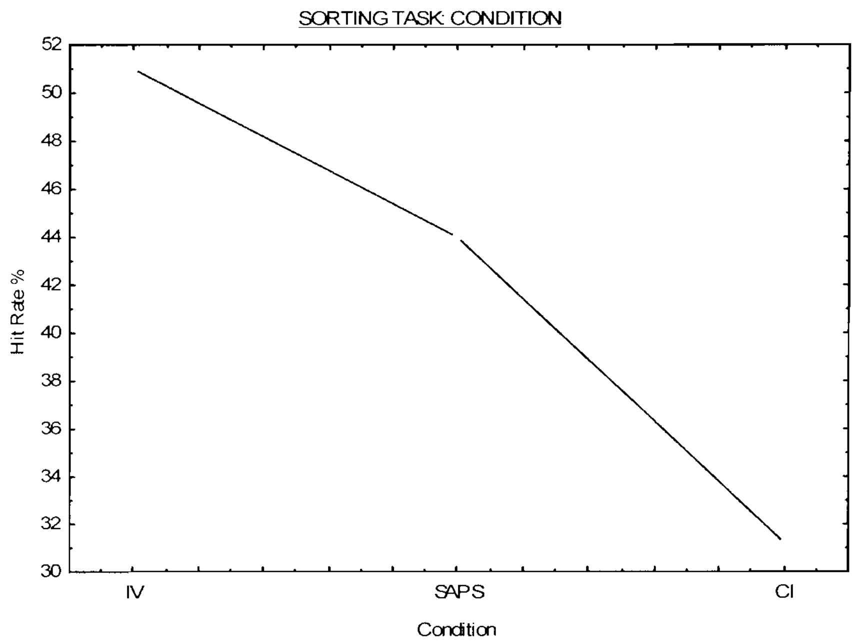
All reconstructions made by participants received an average hit rate of 42% (606 hits out of 1440). Faces composites were found to be slightly better than E-Face composites, receiving a hit rate of 45% (323 hits out of 720) compared to the E-Face hit rate of 39% (283 hits out of 720). (See Appendix H for the sorting hit rates for the reconstructions made under the first three experimental conditions). A logistic regression was used as a modelling tool for the sorting data. The "Composite Number" variable was again dropped, as it does not aid the modelling or interpretation of the key experimental effects. The logistic regression (see Table 14 above) produced a χ^2 of 4.561 on 1 *df* for the "Reconstructor" factor, which is significant at 0.033. This suggests that the type of reconstructor used (i.e. Faces and/or E-Face) made a significant contribution to the prediction of the sorting outcome. However, since these comparisons are "untested", such an inference cannot be made. Figure 36 below presents the reconstructor sorting hit rates graphically.

Figure 36: "Reconstructor" sorting hit rates



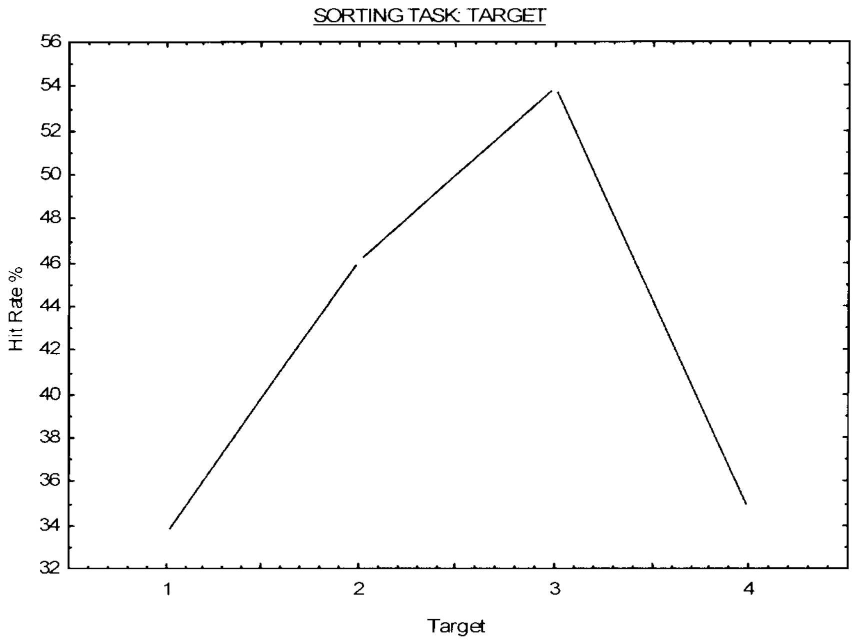
The sorting results revealed that Novice IV composites (hit rate = 51%) were better than SAPS composites (hit rate = 44%); which in turn were better than CI composites (hit rate = 31%). The logistic regression produced a significant χ^2 of 38.847 on 1 *df* for the "Condition" factor. This suggests that the condition that participants were put through (i.e. Novice IV, SAPS Interview and/or CI) made a significant contribution to the prediction of the sorting outcome. However, since these comparisons are "untested", such an inference cannot be made. Figure 37 below presents the condition sorting hit rates graphically.

Figure 37: "Condition" sorting hit rates



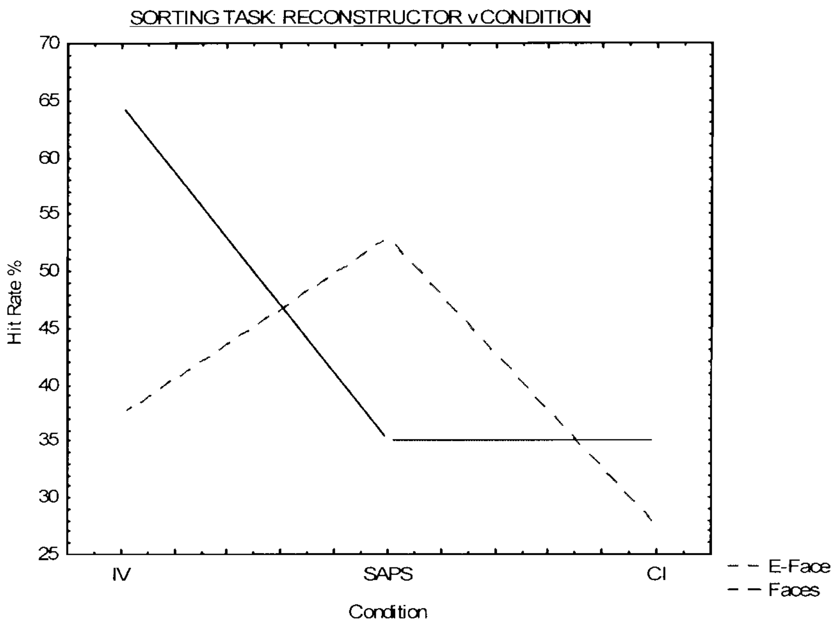
The sorting results revealed that Target 3 was easiest to reconstruct, followed by Targets 2, 4 and then 1. Their hit rates were 54%, 46%, 35% and 34% respectively. The logistic regression produced a χ^2 of 0.912 on 1 *df* for the "Target" factor, which is not significant at 0.340. This suggests that the target reconstructed by participants (i.e. Targets 1, 2, 3 and/or 4) did not make a significant contribution to the prediction of the sorting outcome. However, since these comparisons are "untested", such an inference cannot be made. Figure 38 below presents the target sorting hit rates graphically.

Figure 38: "Target" sorting hit rates



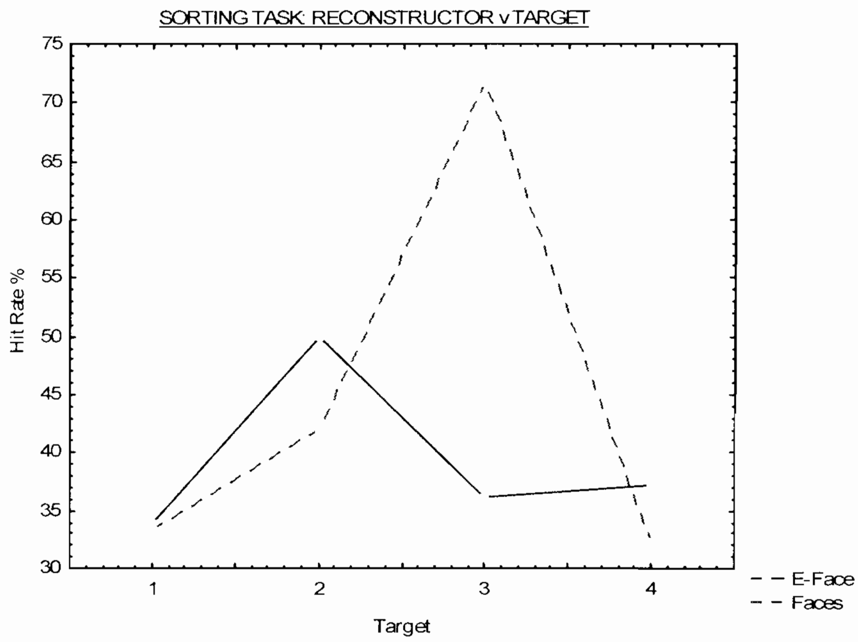
For the Faces system, the sorting task results show that Novice IV composites (hit rate = 65%) were considerably better than those produced from memory. The CI and SAPS composites were found to be the same in quality, both receiving a hit rate of 35%. For the E-Face system, the sorting task results show that SAPS composites (hit rate = 53%) were better than Novice IV composites (hit rate = 38%), which in turn were better than CI composites (hit rate = 28%). The logistic regression results reveal that for the "Reconstructor x Condition" interaction, the change in χ^2 from the full model (i.e. the model including all three factors) is 0.940, which is not significant at 0.332. This suggests that the model which best fits the sorting data is the "Reconstructor x Condition" model. However, since these comparisons are "untested", such an inference cannot be made. Figure 39 below presents the "Reconstructor x Condition" sorting hit rates graphically.

Figure 39: "R x C" sorting hit rates



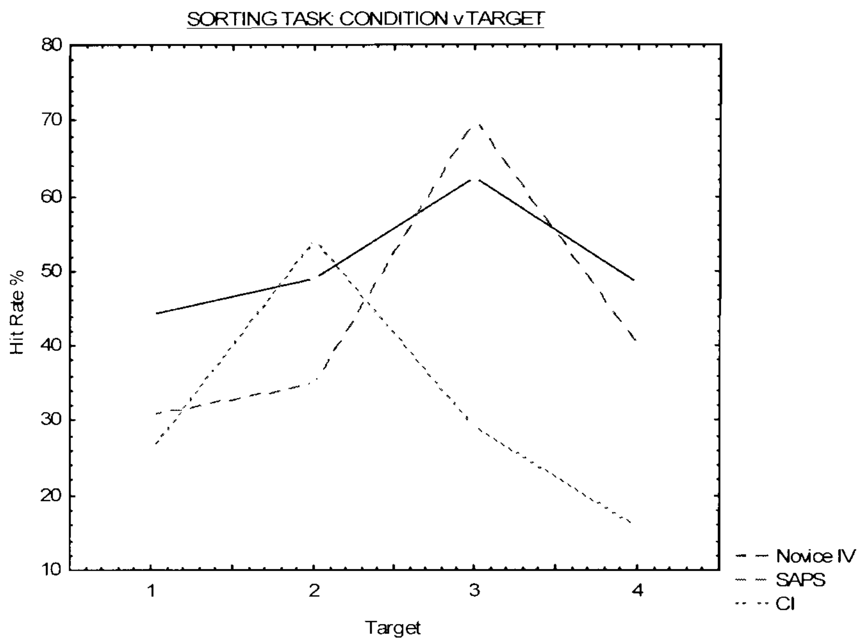
For the Faces system, the sorting task results show that Target 3 performed remarkably well against the other targets, achieving a hit rate of 72%. For the E-Face system, the sorting task results show that Target 2 performed the best with a hit rate of 50%. The logistic regression results reveal that for the "Reconstructor x Target" interaction, the change in χ^2 from the full model (i.e. the model including all three factors) is 38.999, which is significant. This suggests that the "Reconstructor x Target" model is not the best fit for the sorting data. However, since these comparisons are "untested", such an inference cannot be made. Figure 40 below presents the "Reconstructor x Target" sorting hit rates graphically.

Figure 40: "R x T" sorting hit rates



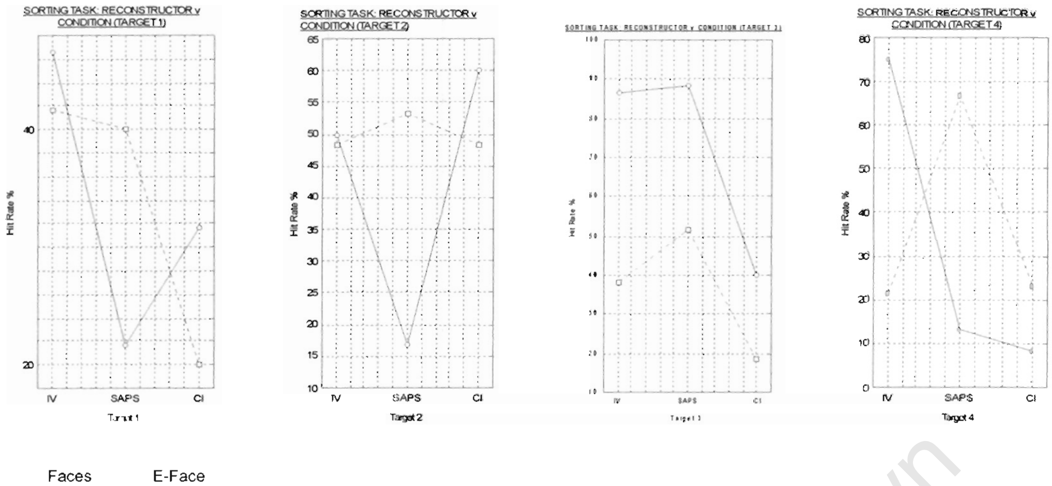
For the SAPS Interview condition, the sorting task results show that Target 3 performed remarkably well against the other targets, achieving a hit rate of 70%. For the CI condition, the sorting task results show that Target 2 performed the best with a hit rate of 54%. For the Novice IV condition, the sorting task results show that Target 3 performed the best with a hit rate of 63%. The logistic regression results reveal that for the "Condition x Target" interaction, the change in χ^2 from the full model (i.e. the model including all three factors) is 4.690, which is significant at 0.030. This suggests that the "Condition x Target" model is not the best fit for the sorting data. However, since these comparisons are "untested", such an inference cannot be made. Figure 41 below presents the "Condition x Target" sorting hit rates graphically.

Figure 41: "C x T" sorting hit rates



For all targets reconstructed under the Novice IV condition, the Faces system performed better than the E-Face system. For Target 1, the Novice IV condition resulted in the best sorting hit rates for both reconstructors, i.e. 47% for the Faces system and 42% for the E-Face system. For Target 3, the SAPS condition resulted in the best sorting hit rates for both reconstructors, i.e. 88% for the Faces system and 52% for the E-Face system. The sorting logistic regression results indicate that the "Reconstructor x Condition x Target" model (i.e. the full model) is not the best fit for the sorting data. However, since these comparisons are "untested", such an inference cannot be made. Figure 42 below presents the "Reconstructor x Condition x Target" sorting hit rates graphically.

Figure 42: "R x C x T" sorting hit rates



The best-sorted composites for Targets 1 and 2 were created using the E-Face system, while the best-sorted composites for Targets 3 and 4 were created using the Faces system. The best-sorted composites for Targets 1 and 4 were created under the Novice IV condition, while the best-sorted composites for Targets 2 and 3 were made under the SAPS condition. Hit rates for the best-sorted composites for Targets 1, 2, 3 and 4 were 55%, 100%, 100% and 85% respectively (see Appendix I and Figure 43).

Figure 43: Best-sorted composites (bottom) of each target face (top)

Task 3: Ranking Task

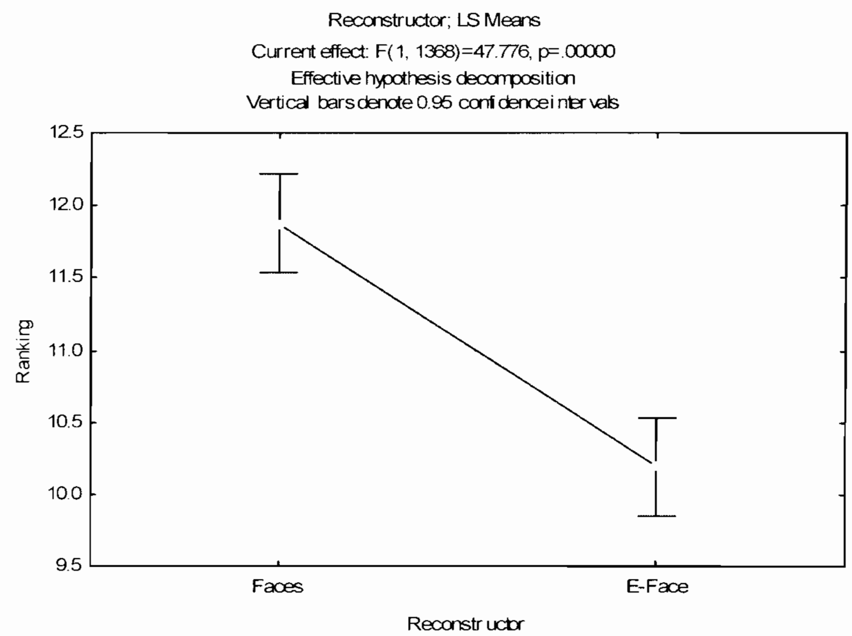
For the ranking task, evaluators were required to scrutinise each facial composite relative to the photograph of the target, and then rank order each of the four stacks of composite cards from most similar-looking to least similar-looking (to the matching target faces). Each composite received a score between 1 and 20 depending on where they were ranked by evaluators, e.g. a composite ranked fifth out of 20 by an evaluator would be given a score of 5. To clarify, the lower the ranking score, the better the quality of the facial composite.

Table 15: Four-way ranking factorial ANOVA results

Source	*SS*	*df*	*MS*	*F*	*p*
Intercept	175319.7	1	175319.7	8224.521	0.000000*
(1)Reconstructor	1018.4	1	1018.4	47.776	0.000000*
(2)Condition	3216.7	2	1608.3	75.450	0.000000*
(3)Target	53.7	3	17.9	0.839	0.472495
(4)C.Number	19.3	2	9.7	0.453	0.635986
Reconstructor*Condition	2372.7	2	1186.3	55.653	0.000000*
Reconstructor*Target	1313.8	3	437.9	20.545	0.000000*
Condition*Target	1323.7	6	220.6	10.349	0.000000*
Reconstructor*C.Number	402.0	2	201.0	9.428	0.000086*
Condition*C.Number	615.6	4	153.9	7.220	0.000009*
Target*C.Number	325.2	6	54.2	2.542	0.018815*
Reconstuctor*Condition*Target	1594.4	6	265.7	12.466	0.000000*
1*2*4	128.2	4	32.1	1.504	0.198731
Reconstructor*Target*C.Number	708.7	6	118.1	5.541	0.000011*
Condition*Target*C.Number	1167.4	12	97.3	4.564	0.000000*
1*2*3*4	1686.4	12	140.5	6.593	0.000000*
Error	29161.3	1368	21.3		

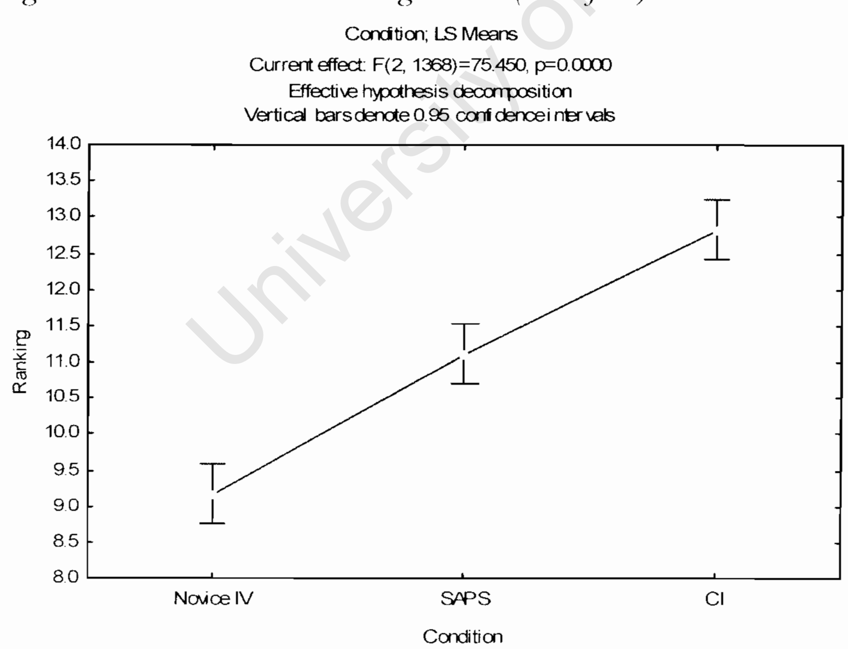
All reconstructions made by participants received an average evaluation ranking of 11.03 out of 20. (See Appendix J for the ranking means of the reconstructions made under the first three experimental conditions). As can be seen in Table 15, the four-way factorial ANOVA results show that the main effect for "Reconstructor" was significant, $F(1, 1368) = 47.78$, $p < 0.01$. Although the main effect for "Composite Number" is an important one, it has been dropped in the ranking results analysis, as it does not aid the interpretation of the key experimental effects. Post-hoc analysis using Tukey HSD comparison tests revealed that E-Face composites ($M = 10.19$) received significantly better mean rankings than Faces composites ($M = 11.88$). Figure 44 below presents the reconstructor ranking means graphically.

Figure 44: "Reconstructor" ranking means (out of 20)



The main effect for "Condition" was significant, F (2, 1368) = 75.45, $p < 0.01$. Post-hoc analysis using Tukey HSD comparison tests revealed that Novice IV composites (M = 9.16) received significantly better mean rankings than SAPS composites (M = 11.11) and CI composites (M = 12.82). SAPS composites (M = 11.11) received significantly better mean rankings than CI composites (M = 12.82). Figure 45 below presents the condition ranking means graphically.

Figure 45: "Condition" ranking means (out of 20)

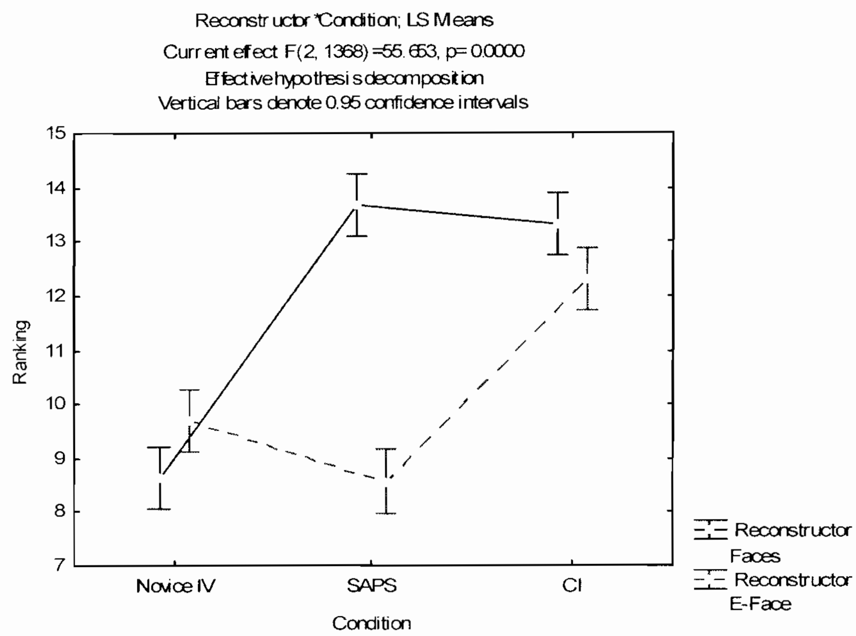


The main effect for "Target" was not significant, $F(3, 1368) = 0.84$, $p > 0.05$. The interaction between "Reconstructor" and "Condition" was significant, $F(2, 1368) = 55.65$, $p < 0.01$. As can be seen in Table 16 below, the simple effects analysis reveals that there was a significant interaction between "Condition" and the E-Face reconstructor, $F(2, 1368) = 41.63$, $p < 0.01$, such that CI composites (M = 12.31) received significantly poorer mean rankings than Novice IV (M = 9.70) and SAPS composites (M = 8.56). There was also a significant interaction between "Condition" and the Faces reconstructor, $F(2, 1368) = 89.57$, $p < 0.01$, such that Novice IV composites (M = 8.63) were significantly better ranked than SAPS (M = 13.67) and CI composites (M = 13.33). In addition, significant interactions were found between "Reconstructor" and the two "from memory" conditions, such that E-Face composites were better ranked than Faces composites under these conditions. Figure 46 below presents the "R x C" ranking means graphically.

Table 16: "R x C" interaction

Source	*SS*	*df*	*MS*	*F*	*p*
R at Novice IV	139.75	1	139.75	6.561	0.027223*
R at SAPS	3126.30	1	3126.30	146.775	0.000000*
R at CI	125.05	1	125.05	5.871	0.028238*
C at E-Face	1773.54	2	886.77	41.632	0.000000*
C at Faces	3815.8	2	1907.9	89.574	0.000000*
Error	29161.3	1368	21.3		

Figure 46: Ranking means for "R x C" (out of 20)

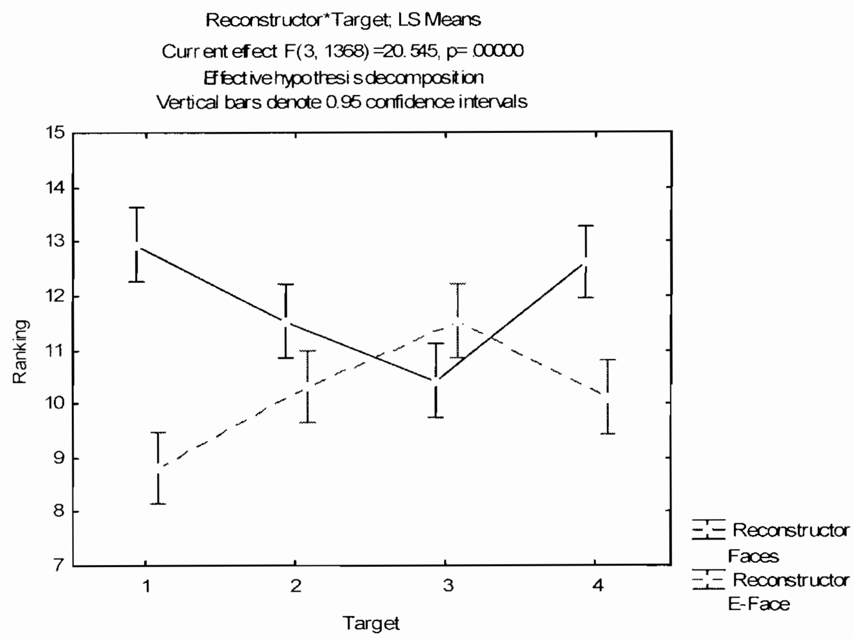


The interaction between "Reconstructor" and "Target" was significant, F (3, 1368) = 20.55, $p < 0.01$. As can be seen in Table 17, the simple effects analysis reveals that there was a significant interaction between "Reconstructor" and Targets 1, 2 and 4, such that E-Face composites received considerably better mean rankings than Faces composites for those three targets. Figure 47 below presents the "R x T" ranking means graphically.

Table 17: "R x T" interaction

Source	*SS*	*df*	*MS*	*F*	*P*
R at T1	1525.23	1	1525.23	71.607	0.000000*
R at T2	132.01	1	132.01	6.198	0.041721*
R at T3	110.00	1	110.00	5.164	0.061641
R at T4	565.00	1	565.00	26.526	0.000008*
T at E-Face	660.22	3	220.07	10.332	0.000031*
T at Faces	707.3	3	235.8	11.068	0.000074*
Error	29161.3	1368	21.3		

Figure 47: Ranking means for "R x T" (out of 20)

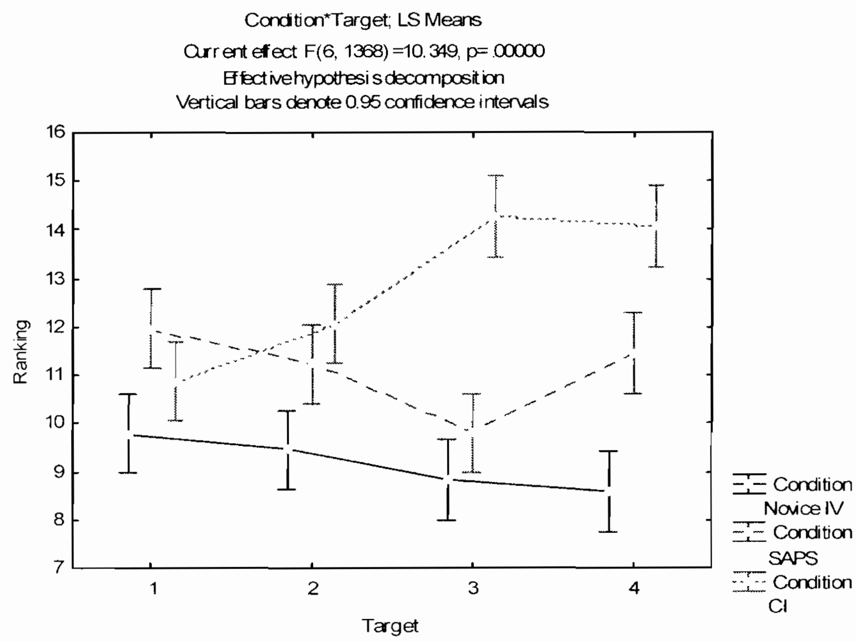


The interaction between "Condition" and "Target" was significant, F (6, 1368) = 10.35, $p < 0.01$. As can be seen in Table 18, the simple effects analysis reveals that there was a significant interaction between "Condition" and all four targets. There was a significant interaction between "Target" and the SAPS condition, F (3, 1368) = 4.80, $p < 0.05$; and between "Target" and the CI condition, F (3, 1368) = 15.00, $p < 0.01$. Figure 48 below presents the "C x T" ranking means graphically.

Table 18: "C x T" interaction

Source	*SS*	*df*	*MS*	*F*	*p*
C at T1	283.84	2	141.92	6.663	0.013336*
C at T2	426.57	2	213.29	10.013	0.001138*
C at T3	2024.82	2	1012.41	47.531	0.000000*
C at T4	1805.11	2	902.55	42.373	0.000000*
T at Novice IV	111.92	3	37.31	1.752	0.273211
T at SAPS	306.66	3	102.22	4.799	0.024218*
T at CI	958.72	3	319.57	15.003	0.000000*
Error	29161.3	1368	21.3		

Figure 48: Ranking means for "C x T" (out of 20)

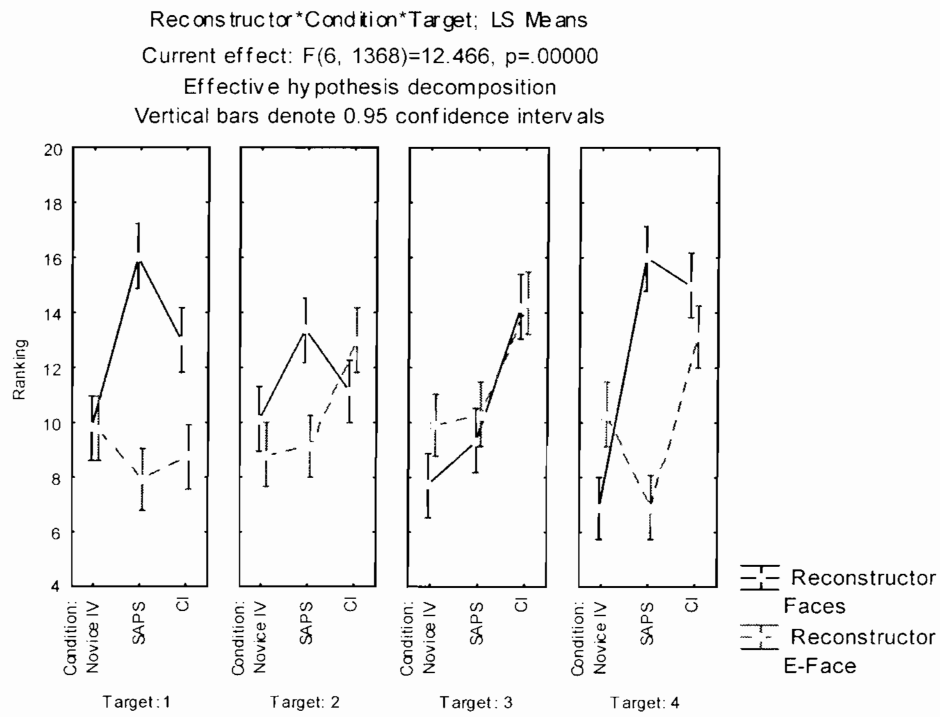


The three-way interaction between "Reconstructor", "Condition" and "Target" was significant, F (6, 1368) = 12.47, $p < 0.01$. As can be seen in Table 19 below, the simple interaction effects analysis reveals that there was a significant interaction between "Reconstructor" and "Condition" for Target 1 composites, F (2, 1368) = 23.02, $p < 0.01$; for Target 2 composites, F (2, 1368) = 13.04, $p < 0.01$; and for Target 4 composites, F (2, 1368) = 55.60, $p < 0.01$. Figure 49 below presents the "RC x T" rating means graphically.

Table 19: "RC x T" interaction

Source	SS	df	MS	F	p
RC at T1	980.72	2	490.36	23.022	0.000000*
RC at T2	555.34	2	277.67	13.036	0.000095*
RC at T3	62.69	2	31.34	1.472	0.297341
RC at T4	2368.37	2	1184.19	55.596	0.000000*
Error	29161.3	1368	21.3		

Figure 49: Ranking means for "RC x T" (out of 20)



The best-ranked composites for Targets 1 and 2 were created using the E-Face system, under the SAPS condition. They received evaluation rankings of 6.35 and 6.65 out of 20 respectively. The best-ranked composites for Targets 3 and 4 were both created using the Faces system, under the Novice IV condition. They received evaluation rankings of 6.85 and 3.55 out of 20 respectively (see Appendix K and Figure 50).

Figure 50: Best ranked composites (bottom) of each target face (top)

Reconstruction Phase (Novice IV v Expert IV)

Upon completion of the reconstructions, the two expert operators were asked on a 7-point Likert-type scale how accurate they thought their reconstructions were to the target face, 1 being "highly inaccurate" and 7 being "highly accurate". All reconstructions made by participants under the last two experimental conditions (i.e. Novice IV and Expert IV) received an average satisfaction rating of 5.38 out of 7, which equates to 77%. Target 3 received the highest satisfaction rating of 6.50 out of 7; Target 2 received the lowest satisfaction rating of 4.50 out of 7. The Faces expert (M = 5.75) made a higher average satisfaction rating than the E-Face expert (M = 5.00). Both expert operators gave Target 3 their highest satisfaction ratings. Table 20 below shows the mean expert satisfaction ratings outlined above.

Table 20: Mean expert satisfaction ratings (out of 7)

T/Face	Reconstructor	Valid N	Mean	Minimum	Maximum	Std. Dev.
		160	5.38	4.00	7.00	0.86
	E-Face	80	5.00	4.00	6.00	0.71
	Faces	80	5.75	5.00	7.00	0.83
Target 1		40	5.00	5.00	5.00	0.00
Target 2		40	4.50	4.00	5.00	0.51
Target 3		40	6.50	6.00	7.00	0.51
Target 4		40	5.50	5.00	6.00	0.51
Target 1	E-Face	20	5.00	5.00	5.00	0.00
Target 2	E-Face	20	4.00	4.00	4.00	0.00
Target 3	E-Face	20	6.00	6.00	6.00	0.00
Target 4	E-Face	20	5.00	5.00	5.00	0.00
Target 1	Faces	20	5.00	5.00	5.00	0.00
Target 2	Faces	20	5.00	5.00	5.00	0.00
Target 3	Faces	20	7.00	7.00	7.00	0.00
Target 4	Faces	20	6.00	6.00	6.00	0.00

Evaluation Phase (Novice IV v Expert IV)

For the evaluation phase, the results of the last two experimental conditions (i.e. Novice IV and Expert IV) are separately reported for the three evaluation tasks of the experiment, i.e. (1) the rating task, (2) sorting task and (3) ranking task.

Task 1: Rating Task

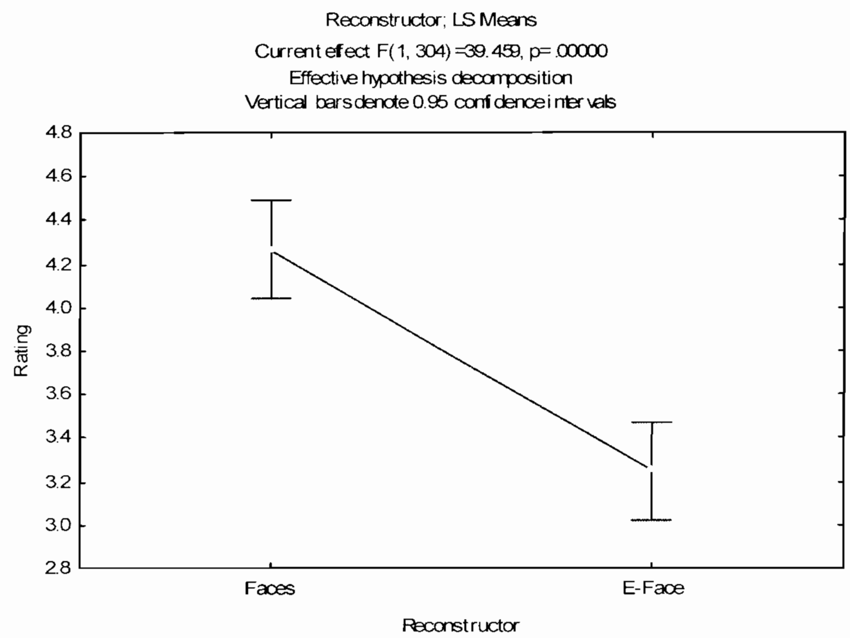
For the rating task, evaluators were asked to use a 7-point Likert-type scale to rate the composites produced by participants in the reconstruction phase against the target faces they were trying to reconstruct. Ratings were made according to similarity, 1 being "extremely dissimilar" and 7 being "extremely similar".

Table 21: Three-Way factorial ANOVA rating results

Source	*SS*	*df*	*MS*	*F*	*p*
Intercept	4510.006	1	4510.006	2152.154	0.000000*
Reconstructor	82.689	1	82.689	39.459	0.000000*
Condition	47.022	1	47.022	22.439	0.000003*
Target	32.831	3	10.944	5.222	0.001576*
Reconstructor*Condition	3.472	1	3.472	1.657	0.198998
Reconstructor*Target	11.575	3	3.858	1.841	0.139663
Condition*Target	14.331	3	4.777	2.279	0.079469
Reconstructor*Condition*Target	1.242	3	0.414	0.198	0.898053
Error	637.056	304	2.096		

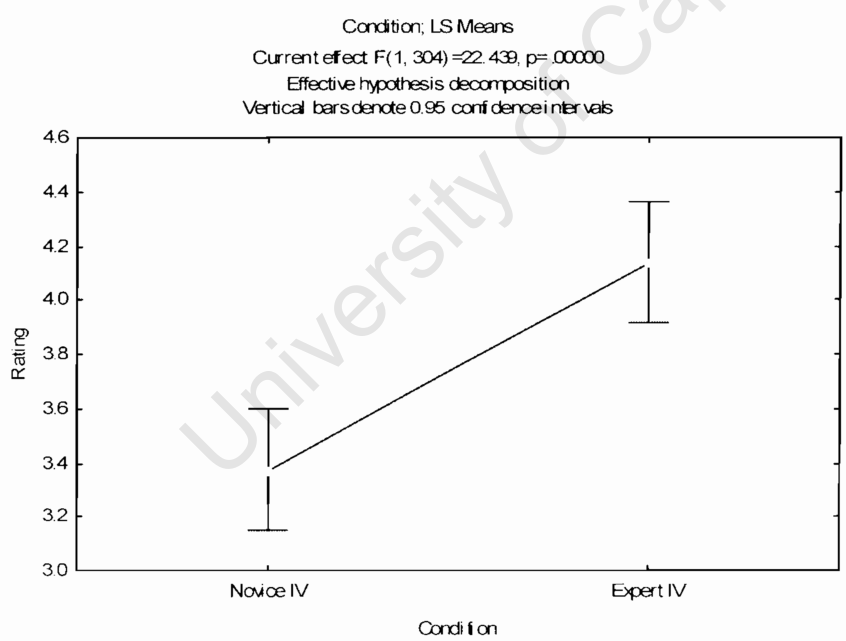
All reconstructions made by participants in the Novice IV and Expert IV conditions received an average rating score of 3.75 out of 7, which equates to 54%. (See Appendix L for the similarity rating means of the reconstructions made under the two IV conditions). As can be seen in Table 21 above, the three-way factorial ANOVA results show that the main effect for "Reconstructor" was significant, $F(1, 304) = 39.46$, $p < 0.01$. Post-hoc analysis using Tukey HSD comparison tests revealed Faces composites ($M = 4.26$) received significantly higher mean ratings than E-Face composites ($M = 3.25$). Figure 51 below presents the reconstructor rating means graphically.

Figure 51: "Reconstructor" rating means (out of 7)



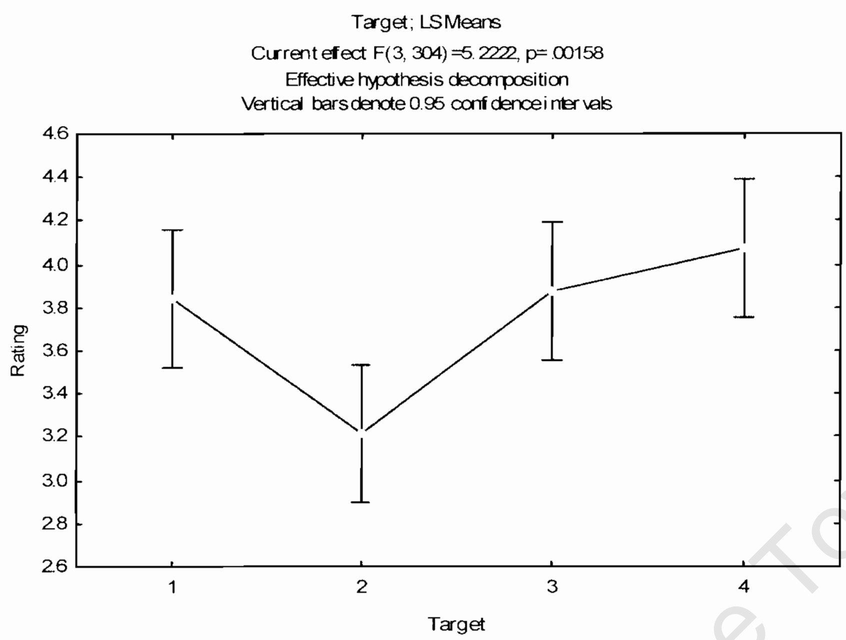
The main effect for Condition was significant, F (1, 304) = 47.02, $p < 0.01$. Post-hoc analysis using Tukey HSD comparison tests revealed Expert IV composites (M = 4.14) performed significantly better than Novice IV composites (M = 3.37), as expected. Figure 52 below presents the two condition rating means graphically.

Figure 52: "Condition" rating means (out of 7)



The main effect for Target was significant, $F (3, 304) = 10.94$, $p < 0.01$. Post-hoc analysis using Tukey HSD comparison tests revealed Target 2 composites (M = 3.22) received the lowest ratings on average, significantly lower than Target 1 composites (M = 3.85), Target 3 composites (M = 3.88) and Target 4 composites (M = 4.08). Figure 53 below presents the four target rating means graphically.

Figure 53: "Target" rating means (out of 7)



All two- and three-way interactions were found to be not significant. The interaction between "Reconstructor" and "Condition" was not significant, $F (1, 304) = 1.66$, $p > 0.05$. The interaction between "Reconstructor" and "Target" was not significant, $F (3, 304) = 1.84$, $p > 0.05$. The interaction between "Condition" and "Target" was not significant, $F (3, 304) = 2.28$, $p > 0.05$. The three-way interaction between "Reconstructor", "Condition" and "Target" was not significant, $F (3, 304) = 0.20$, $p > 0.05$.

The best-rated Expert IV composites of Targets 1 to 4 were all created using the Faces system. They received evaluation ratings of 5.00, 3.65, 4.95 and 5.40 out of 7 respectively (see Figure 54). The best-rated Novice IV composites for Targets 1, 3 and 4 were created using the Faces system. They received evaluation ratings of 4.30, 4.65 and 4.95 out of 7 respectively. The best-rated Novice IV composite for Target 2 was created using the E-Face system, and received an evaluation rating of 3.75 out of 7 (see Appendix M and Figure 55).

Figure 54: Best-rated Expert IV composites (bottom) of each target face (top)

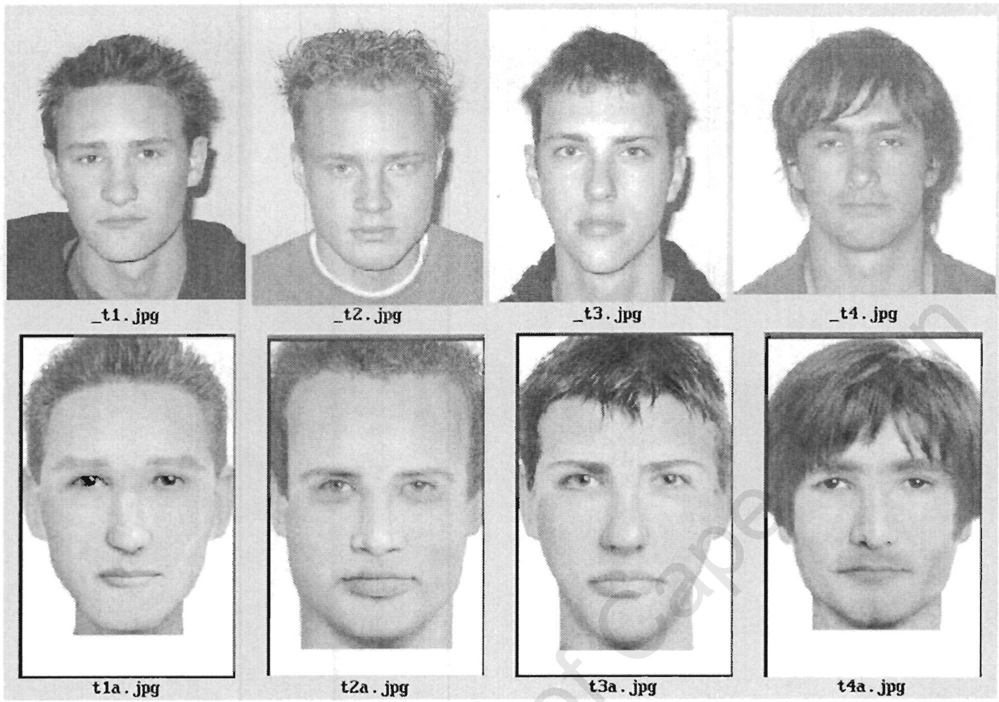


Figure 55: Best-rated Novice IV composites (bottom) of each target face (top)

Task 2: Sorting Task

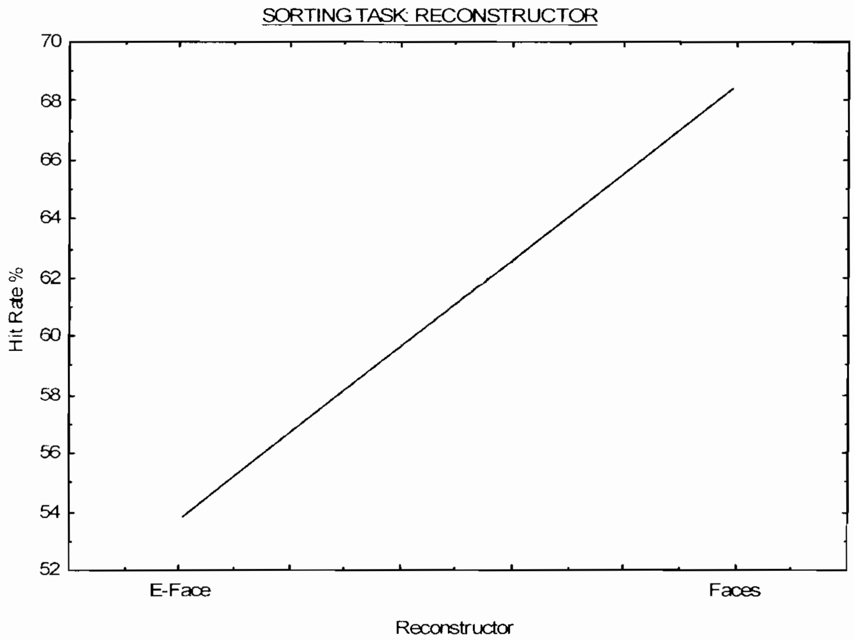
For the sorting task, evaluators were to sort the facial composites produced by participants in the reconstruction phase according to what they perceived to be the matching target face. Each facial composite received a score of either 0 or 1, depending on whether or not they were correctly assigned to the matching target face by the evaluators.

Table 22: Sorting logistic regression results

	χ^2	*df*	*p*	χ^2 difference from full model	*p*
Reconstructor	7.4021	1	0.00652*	26.01573	0.000002*
Condition	13.873	1	0.00020*	19.54444	0.000057*
Target	11.001	1	0.00091*	22.41725	0.000014*
Reconstructor x Condition	21.615	2	0.00002*	11.80323	0.000592*
Reconstructor x Target	18.669	2	0.00009*	14.74892	0.000123*
Condition x Target	25.382	2	0.00000*	8.036147	0.004588*
Reconstructor x Condition x Target	33.418	3	0.00000*	-	-

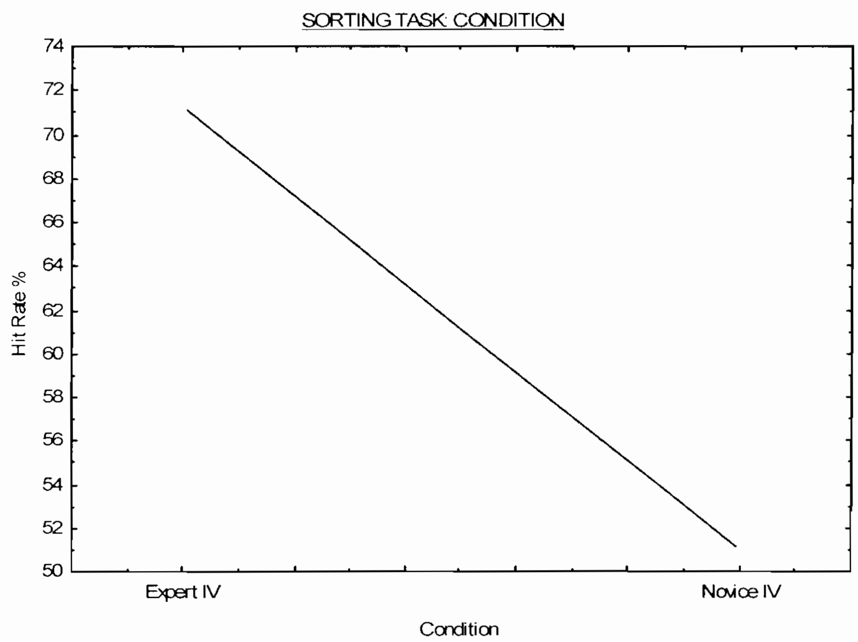
All reconstructions made by participants in the two IV conditions received an average hit rate of 61%. The sorting task results indicate that Faces composites were better than E-Face composites, receiving a hit rate of 69% compared to the E-Face hit rate of 54%. (See Appendix N for the sorting hit rates of the reconstructions made under the two IV conditions). The logistic regression (see Table 22) produced a χ^2 of 7.4021 on 1 *df* for the "Reconstructor" factor, which is significant at 0.00652. This suggests that the type of reconstructor used (i.e. Faces and/or E-face) made a significant contribution to the prediction of the sorting outcome. However, since these comparisons are "untested", such an inference cannot be made. Figure 56 below presents the reconstructor sorting hit rates graphically.

Figure 56: "Reconstructor" sorting hit rates



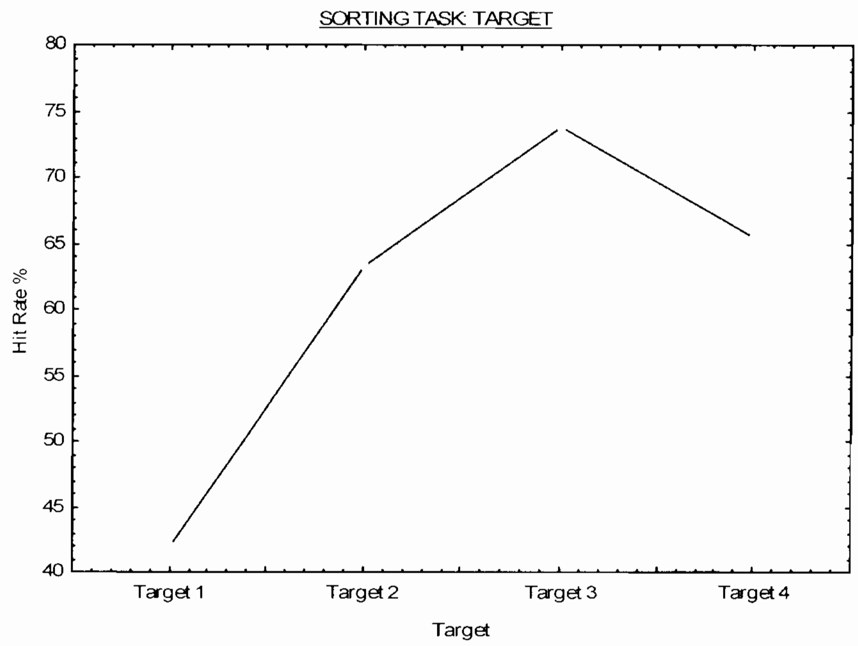
As expected, Expert IV composites (hit rate = 71%) were considerably better than Novice IV composites (hit rate = 51%). The logistic regression produced a significant χ^2 of 13.873 on 1 *df* for the "Condition" factor. This suggests that the condition that participants were put through (i.e. Expert IV and/or Novice IV) made a significant contribution to the prediction of the sorting outcome. However, since these comparisons are "untested", such an inference cannot be made. Figure 57 below presents the condition sorting hit rates graphically.

Figure 57: "Condition" sorting hit rates



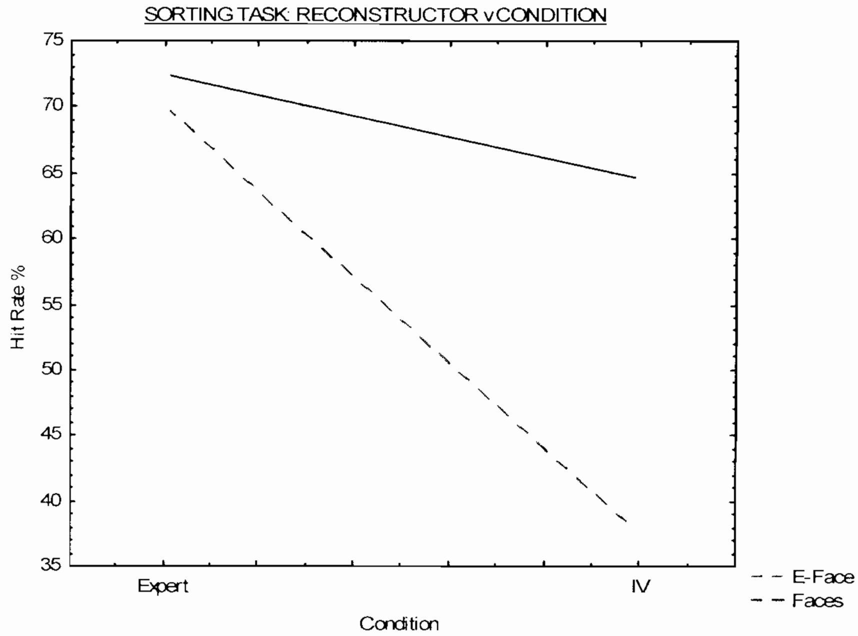
The sorting task results reveal that Target 3 was the target which was easiest to reconstruct, followed by Target 4, 2, and then 1. Their hit rates were 74%, 65%, 63% and 42% respectively. The logistic regression produced a χ^2 of 11.001 on 1 *df* for the "Target" factor, which is significant at 0.00091. This suggests that the target reconstructed by participants (i.e. Targets 1, 2, 3, and/or 4) made a significant contribution to the prediction of the sorting outcome. However, since these comparisons are "untested", such an inference cannot be made. Figure 58 below presents the target sorting hit rates graphically.

Figure 58: "Target" sorting hit rates



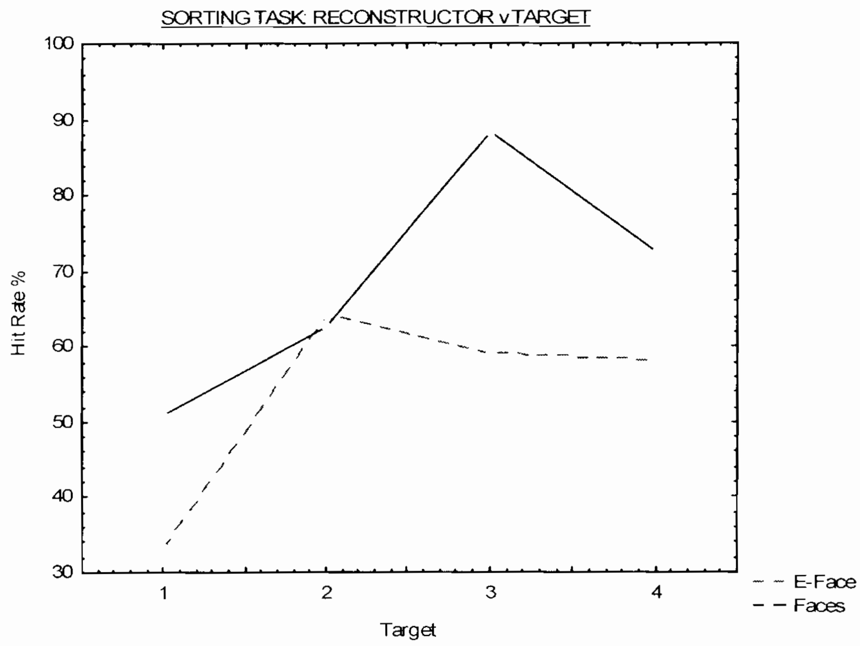
As expected, Expert IV composites were better than Novice IV composites for both reconstructors. The sorting task results indicate that the Faces system produced better composites than the E-Face system under the two IV conditions. The logistic regression results reveal that for the "R x C" interaction, the change in χ^2 from the full model (i.e. the model including all three factors) is 11.80, which is significant at 0.0006. This suggests that the "R x C" model is not the best fit for the sorting data. However, since these comparisons are "untested", such an inference cannot be made. Figure 59 below presents the "Reconstructor x Condition" sorting hit rates graphically.

Figure 59: "R x C" sorting hit rates



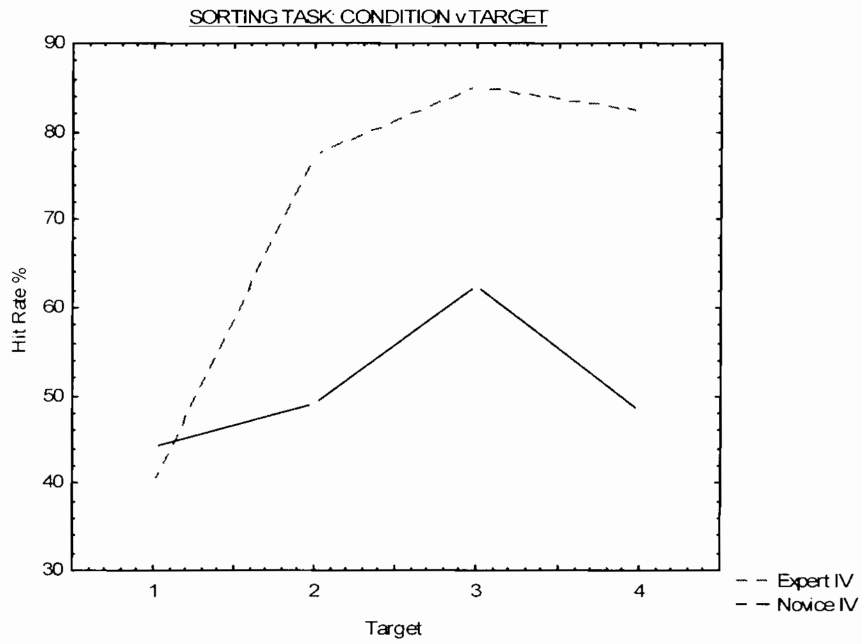
For all targets except Target 2, the Faces system performed better than the E-Face system. For both reconstructors, Target 1 performed the worst with a sorting hit rate of 50% when the Faces system was used and 33% when the E-Face system was used. The logistic regression results reveal that for the "R x T" interaction, the change in χ^2 from the full model (i.e. the model including all three factors) is 14.75, which is significant at 0.0001. This suggests that the "R x T" model is not the best fit for the sorting data. However, since these comparisons are "untested", such an inference cannot be made. Figure 60 below presents the "R x T" sorting hit rates graphically.

Figure 60: "R x T" sorting hit rates



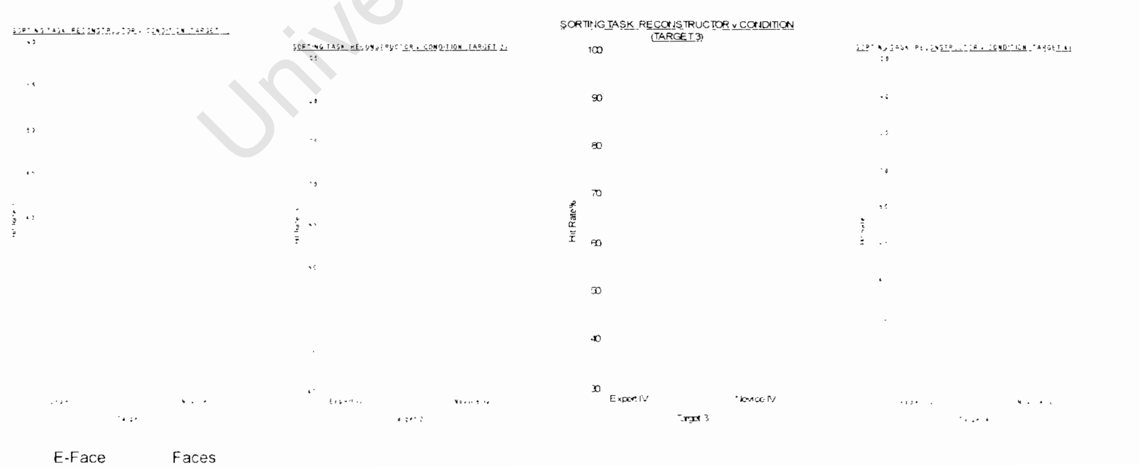
For all targets except Target 1, the experts performed better than the novice users. Target 3 was the top performer for both conditions with sorting hit rates of 85% for the Expert IV condition and 63% for the Novice IV condition. The logistic regression results reveal that for the "C x T" interaction, the change in χ^2 from the full model (i.e. the model including all three factors) is 8.04, which is significant at 0.005. This suggests that the "C x T" model is not the best fit for the sorting data. However, since these comparisons are "untested", such an inference cannot be made. Figure 61 below presents the "C x T" sorting hit rates graphically.

Figure 61: "C x T" sorting hit rates



For all targets and reconstructors, experts performed better than novice users, except when Target 1 was reconstructed using the E-Face system, and when Target 4 was reconstructed using the Faces system. For Targets 1 and 3, the Faces system performed better than the E-Face system for both conditions. The sorting logistic regression results suggest that the full model (i.e. the model including all three factors) has to be accepted since a more parsimonious model cannot be found. However, since these comparisons are "untested", such an inference cannot be made. Figure 62 below presents the "Reconstruction x Condition x Target" sorting hit rates graphically.

Figure 62: "Reconstructor x Condition x Target" sorting hit rates



The best-sorted Expert IV composites of Targets 1 and 3 were made using the Faces system, achieving hit rates of 55% and 95% respectively. The best-sorted Expert IV composites of Targets 2 and 4 were made using the E-Face system, achieving hit rates of 80% and 95% respectively (see Figure 63). The best-sorted Novice IV composites of Targets 2, 3 and 4 were made using the Faces system. Targets 2 and 4 both achieved hit rates of 85%, while Target 3 achieved a hit rate of 95%. The best-sorted Novice IV composite of Target 1 was made using the E-Face system, achieving a hit rate of 55% (see Appendix O and Figure 64).

Figure 63: Best-sorted Expert IV composites (bottom) of each target face (top)

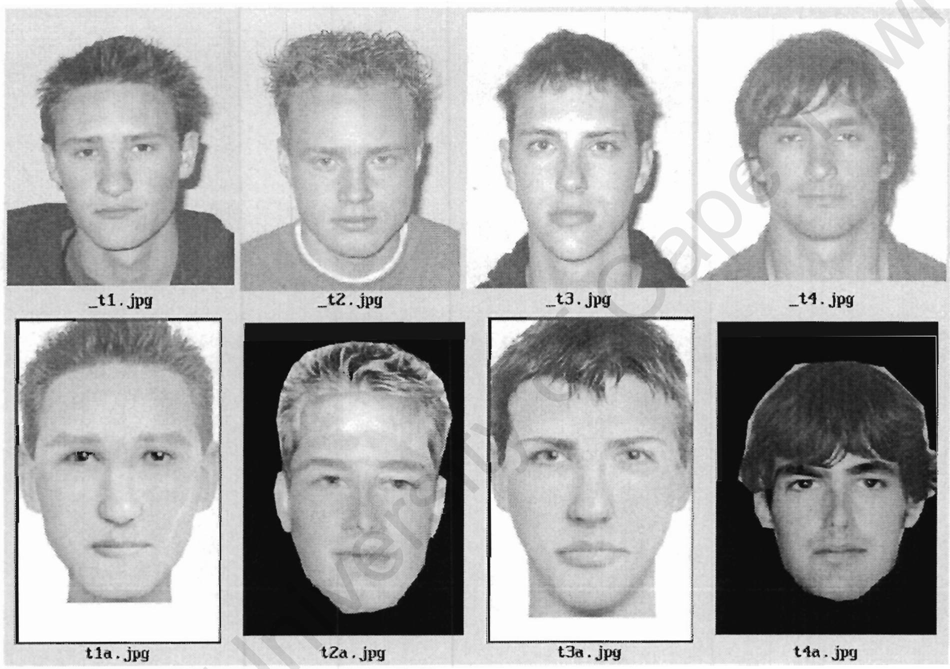
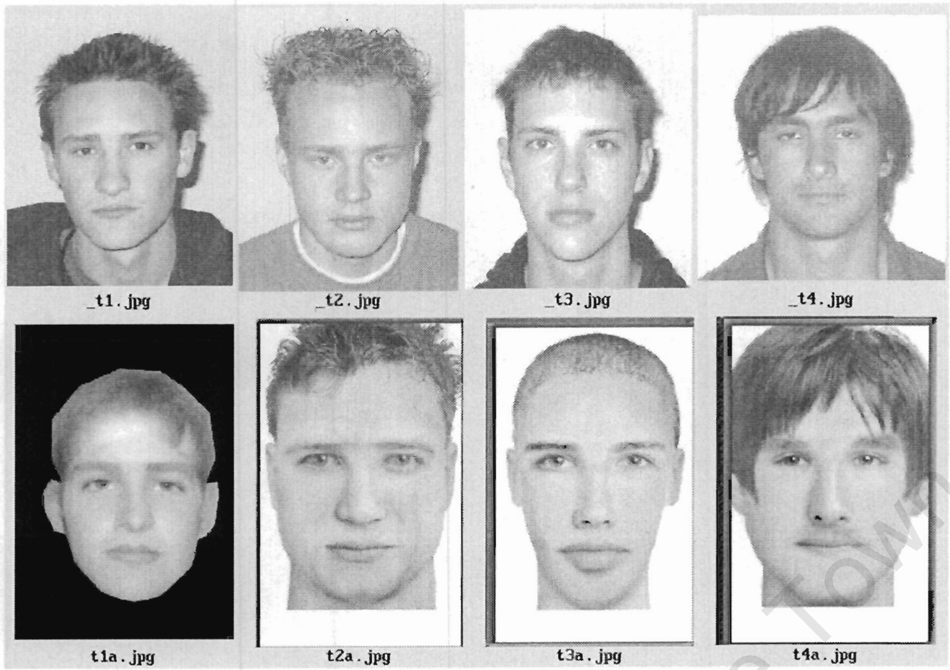


Figure 64: Best-sorted Novice IV composites (bottom) of each target face (top)



Task 3: Ranking Task

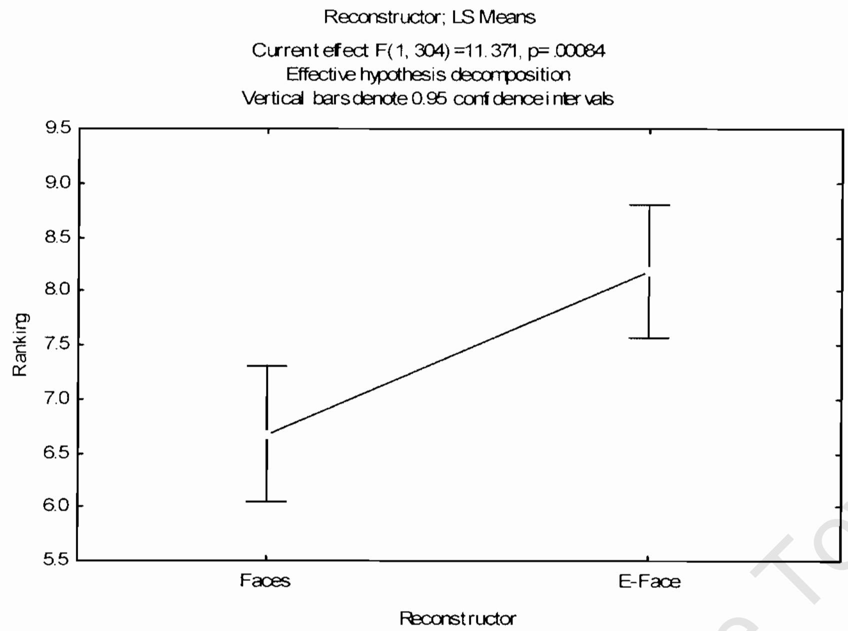
For the ranking task, evaluators were required to scrutinise each composite card relative to the photograph of the target and then rank order each of the four stacks of composite cards from most similar-looking to least similar-looking (to the matching target faces). Each composite received a score between 1 and 20 depending on where they were ranked by evaluators.

Table 23: Three-Way ranking factorial ANOVA results

Source	*SS*	*df*	*MS*	*F*	*p*
Intercept	17661.61	1	17661.61	1103.384	0.000000*
Reconstructor	182.01	1	182.01	11.371	0.000843*
Condition	963.73	1	963.73	60.208	0.000000*
Target	359.12	3	119.71	7.478	0.000076*
Reconstructor*Condition	14.73	1	14.73	0.921	0.338098
Reconstructor*Target	265.08	3	88.36	5.52	0.001056*
Condition*Target	146.16	3	48.72	3.044	0.029142*
Reconstructor*Condition*Target	156.18	3	52.06	3.252	0.022093*
Error	4866.06	304	16.01		

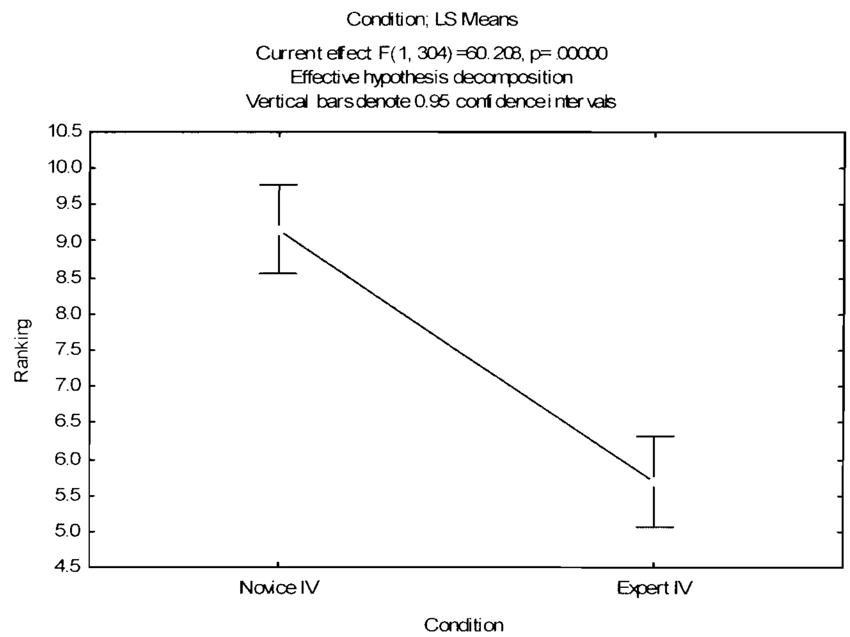
All reconstructions made by participants in the Novice IV and Expert IV conditions received a mean ranking of 7.43 out of 20. (See Appendix P shows the ranking means for the reconstructions made under the last two experimental conditions). As can be seen in Table 23 above, the three-way factorial ANOVA results show that the main effect for "Reconstructor" was significant, F (1, 304) = 11.37, $p < 0.01$. Post-hoc analysis using Tukey HSD comparison tests revealed that Faces composites (M = 6.68) received significantly better mean rankings than E-Face composites (M = 8.18). Figure 65 below presents the reconstructor ranking means graphically.

Figure 65: "Reconstructor" ranking means (out of 20)



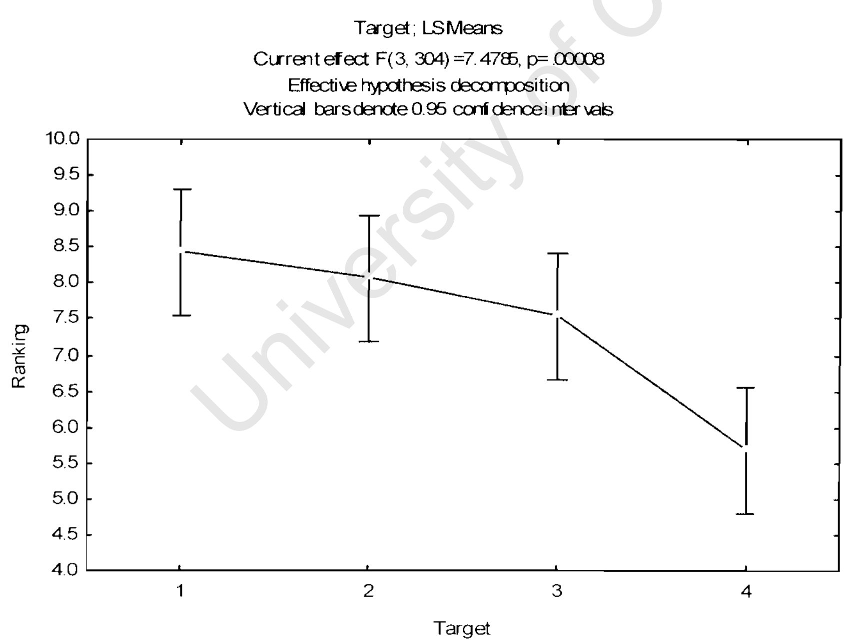
The main effect for "Condition" was significant, F (1, 304) = 60.21, $p < 0.01$. Post-hoc analysis using Tukey HSD comparison tests revealed that Expert IV composites (M = 5.69) received significantly better mean rankings than Novice IV composites (M = 9.16). Figure 66 below presents the condition ranking means graphically.

Figure 66: "Condition" ranking means (out of 20)



The main effect for "Target" was significant, F (3, 304) = 7.48, $p < 0.01$. Post-hoc analysis using Tukey HSD comparison tests revealed that Target 4 (M = 5.68) received significantly higher mean rankings than Targets 1, 2 and 3 (M = 8.43, 8.07 and 7.54 respectively). Figure 67 below presents the target ranking means graphically.

Figure 67: "Target" ranking means (out of 20)

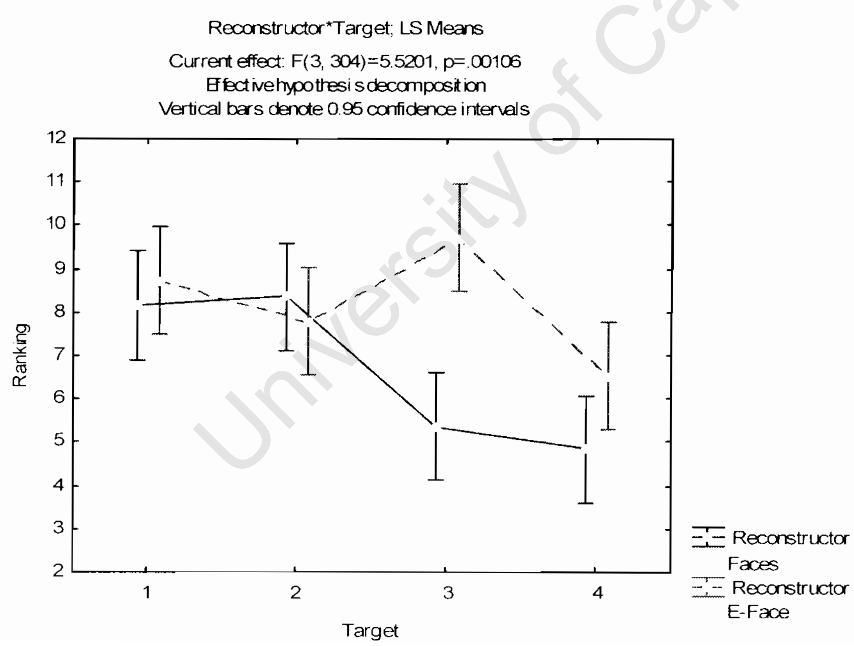


The interaction between "Reconstructor" and "Condition" was not significant, F (1, 304) = 0.92, $p > 0.05$. The interaction between "Reconstructor" and "Target" was significant, F (3, 304) = 5.52, $p < 0.01$. As can be seen in Table 24 below, the simple effects analysis reveals that there was a significant interaction between "Reconstructor" and Target 3, F (1, 304) = 23.55, $p < 0.01$, such that Faces composites (M = 5.37) received considerably better mean rankings than E-Face composites (M = 9.71) of Target 3. There was also a significant interaction between "Target" and E-Face, F (3, 304) = 4.60, $p < 0.01$; and between Target and Faces, F (3, 304) = 8.40, $p < 0.01$. Figure 68 below presents the "R x T" ranking means graphically.

Table 24: "R x T" interaction

Source	*SS*	*df*	*MS*	*F*	*P*
R at T1	6.422	1	6.422	0.4011	0.588343
R at T2	6.422	1	6.422	0.4011	0.588294
R at T3	377.001	1	377.001	23.5479	0.000028*
R at T4	57.235	1	57.235	3.5749	0.064594
T at E-Face	220.81	3	73.60	4.5972	0.007857*
T at Faces	403.389	3	134.463	8.3987	0.000478*
Error	4866.06	304	16.01		

Figure 68: Ranking means for "R x T" (out of 20)

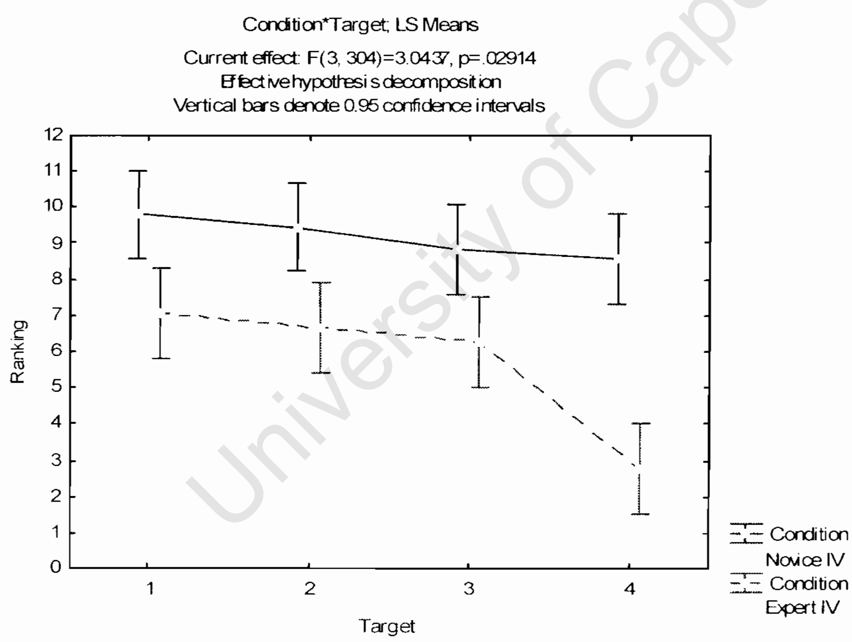


The interaction between "Condition" and "Target" was significant, F (3, 304) = 3.04, $p < 0.05$. As can be seen in Table 25 below, the simple effects analysis reveals that there was a significant interaction between "Condition" and all four targets, such that Expert IV composites received considerably better rankings than Novice IV composites. There was also a significant interaction between "Target" and the Expert IV condition, F (3, 304) = 9.74, $p < 0.01$, such that Target 4 composites (M = 2.78) received considerably better rankings than Targets 1, 2 and 3 composites (M = 7.08, 6.68 and 6.25 respectively) produced by experts. Figure 69 below presents the "C x T" ranking means graphically.

Table 25: "C x T" interaction

Source	SS	df	MS	F	p
C at T1	147.606	1	147.606	9.2196	0.008016*
C at T2	154.939	1	154.939	9.6776	0.006529*
C at T3	132.613	1	132.613	8.2831	0.016716*
C at T4	674.735	1	674.735	42.1446	0.000000*
T at Expert IV	467.969	3	155.990	9.7433	0.000211*
T at Novice IV	37.31	3	12.44	0.7768	0.400669
Error	4866.06	304	16.01		

Figure 69: Ranking means for "C x T" (out of 20)

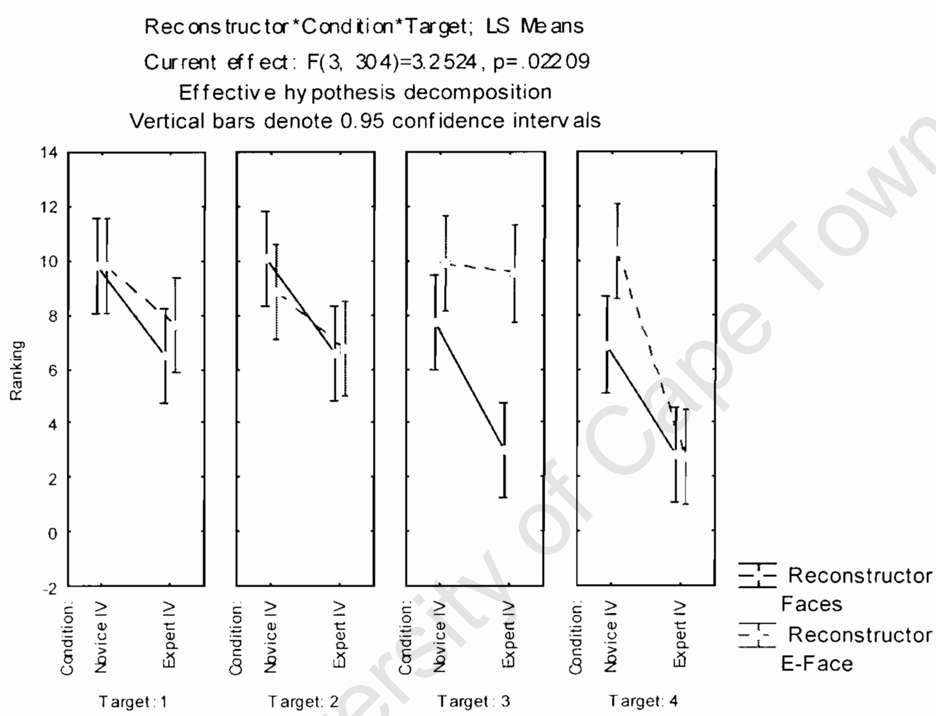


The three-way interaction between "Reconstructor", "Condition" and "Target" was significant, F (3, 304) = 3.25, $p < 0.05$. As can be seen in Table 26 below, the simple interaction effects analysis reveals there was a significant interaction between "Reconstructor" and "Condition" for Target 3 composites, F (1, 304) = 5.82, $p < 0.05$; and for Target 4 composites, F (1, 304) = 3.79, $p < 0.01$. Figure 70 below presents the "RC x T" rating means graphically.

Table 26: "RC x T" interaction

Source	*SS*	*df*	*MS*	*F*	*p*
RC at T1	6.806	1	6.806	0.4251	0.564130
RC at T2	10.272	1	10.272	0.6416	0.477251
RC at T3	93.168	1	93.168	5.8194	0.020264*
RC at T4	60.668	1	60.668	3.7894	0.004401*
Error	4866.06	304	16.01		

Figure 70: Ranking means for "RC x T" (out of 20)



The best-ranked Expert IV composites of Targets 1, 2 and 3 were created using the Faces system. They were ranked 6.50, 6.60 and 3.00 out of 20 respectively. The best-ranked Expert IV composite of Target 4 was created using the E-Face system, and achieved a ranking of 2.75 out of 20 (see Figure 71). The best-ranked Novice IV composites of Targets 1, 3 and 4 were created using the Faces system. They achieved hit rates of 8.30, 6.85 and 3.55 out of 20 respectively. The best-ranked Novice IV composite of Target 2 was created using the E-Face system, and achieved a ranking of 7.25 out of 20 (see Appendix Q and Figure 72).

Figure 71: Best-ranked Expert IV composites (bottom) of each target face (top)

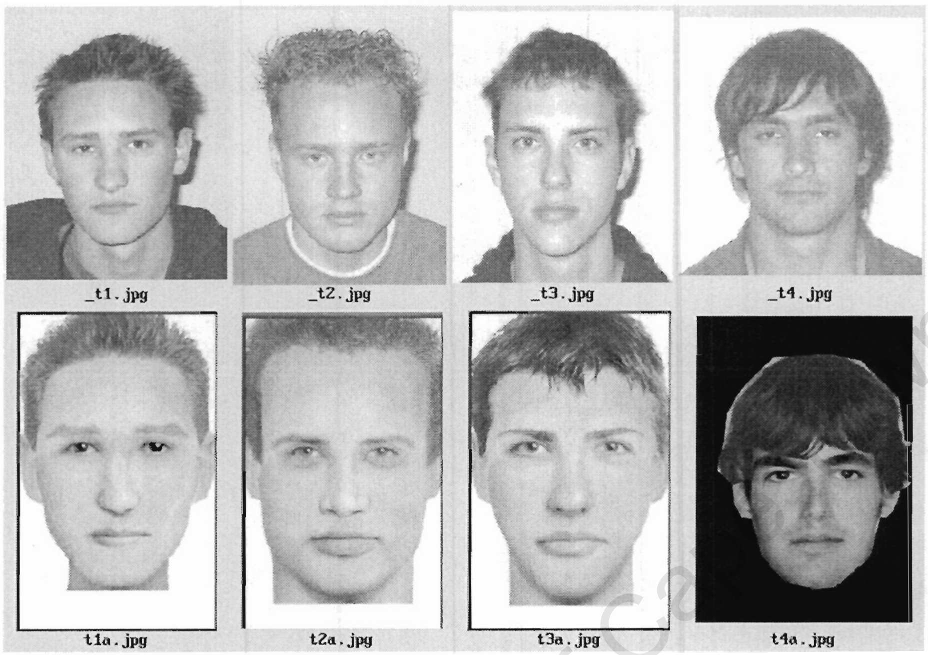


Figure 72: Best-ranked Novice IV composites (bottom) of each target face (top)

7.3. Discussion

The results of Experiment 2 suggest that it takes a shorter period of time to produce a facial composite using the E-Face system than it does using the Faces system, although the difference in time is not significant. It took an average of approximately 25 minutes to produce an E-Face reconstruction, and approximately 27 minutes to produce a Faces reconstruction.

Despite making several positive comments about both the E-Face and Faces composite systems, novice operators did not seem very satisfied with the composites they produced. Novice E-Face operators made an average satisfaction rating of 54%, slightly more than the 51% made by novice Faces operators (see Table 7). Expert operators, on the other hand, were quite satisfied with the "in view" reconstructions they produced, giving them an average satisfaction rating of 77%. The expert Faces operator was 11% more satisfied with his reconstructions than the expert E-Face operator (see Table 20). Participants in the Novice IV condition were generally more satisfied with the composites they produced than those in the two "from memory" conditions. Since these comparisons are "untested", the above inferences cannot be made.

All composites produced by participants in three conditions of the reconstruction phase (i.e. SAPS Interview, CI and Novice IV) were given an average rating score of 41%, an average identification rate of 42%, and an average ranking score of 11.03 (out of 20) by independent evaluators. Although E-Face operators were more satisfied with their reconstructions than Faces operators, rating scores and identification rates given to Faces reconstructions were 42% and 45% respectively, which is slightly greater (1% and 6% greater) than those given to E-Face reconstructions. These results are contradicted by the ranking results though, which indicate that E-Face composites were of significantly better quality than Faces composites. Despite this, further support for the Faces system having the edge over the E-Face system, comes from the fact that the Faces system was used in the production of the majority of the highest performing composites in all the respective conditions of this experiment. Also, when the results of the two "in view" conditions (i.e. Novice IV and Expert IV) were merged, the Faces system was found to be significantly better than the E-Face system in all of the three evaluation tasks.

Composites created "in view" performed significantly better than those produced "from memory" in all of the three evaluation tasks. Novice IV reconstructions received an average rating score of 48%, an average identification rate of 51%, and an average ranking score of 9.16 (out of 20). "From memory" reconstructions received an average rating score of 39%, an average identification rate of 38%, and an average ranking score of 11.97 (out of 20). This anticipated result is consistent with the general finding, in several previously conducted experiments evaluating a range of manual and computerised facial composite systems, that "in view" reconstructions are more accurate than "from memory" ones, e.g. Davies et al.'s (2000) evaluation of E-fit; Ellis et al.'s (1975) evaluation of Photofit; Frowd's (2001) evaluation of EvoFIT; Tredoux et al.'s (2000) evaluation of the latest version of E-Face; Tredoux et al.'s, (1999) evaluation of the first version of E-Face; and Wogalter and Marwitz's (1991) evaluation of the Mac-a-Mug Pro system.

Interestingly, the Novice IV results of this experiment almost mirror the results of Experiment 1, as the average rating score is just 1% higher and the identification rate (following a thresholding of the rating results) is identical to that found in Experiment 1. The same identification rate of 51% was also found for Tredoux et al.'s (1999) Novice IV reconstructions produced using the first version of E-Face. Very similar results were found in Prag's (2000) study (in which I evaluated the black and coloured E-Face databases) where Novice IV composites received a rating score of 46%. In general, the Novice IV results compare very favourably against those found by Frowd (2001) and Brace et al. (2000), who discovered naming rates of 25% for both EvoFIT and E-fit reconstructions; and Davies and Oldman (1999) who found a low 10% naming rate for E-fit reconstructions.

The Novice IV composites in this experiment does not, however, compare too well against those produced in Tredoux et al.'s (1999) evaluation of the first version of E-Face, as the average rating score was 21% less than that found in Tredoux et al.'s (1999) experiment. The Novice IV results also did not compare well against the 83% matching accuracy found when "in view" E-fit reconstructions were made of familiar faces (Davies et al., 2000), and the 81% matching accuracy found when "in view" Whatsisface reconstructions were made (Gillenson and Chandrasekeran, 1975).

As expected, Expert IV composites were significantly better rated, better identified, and better ranked than Novice IV composites. In fact, Expert IV composites performed better than Novice IV composites for both facial composite systems under evaluation, and for all of the four targets in two of the three evaluation tasks. Average rating scores were 48% and 59% for the Novice IV and Expert IV conditions respectively. Average identification rates were 51% and 71% for the Novice IV and Expert IV conditions respectively. Average ranking scores were 9.16 and 5.69 (out of 20) for the Novice IV and Expert IV conditions respectively.

The Expert IV results of this experiment compare very favourably against those of Gibling and Bennet (1994), who used experienced Photofit operators and found "in view" identification rates of 54% when Photofits were enhanced (by removing feature demarcation lines and adding elaborative details), and 15% for Photofits that were not enhanced. They also compare well against the results of Cutler et al. (1988)'s study, in which a 49% identification rate was found after getting an experienced operator to make "in view" reconstructions of 10 target faces. However, the outcome of Koehn and Fisher's (1997) experiment revealed a better identification rate of 77% when an experienced Mac-a-Mug Pro operator made "in view" reconstructions. Similarly, in Tredoux et al.'s (2002) evaluation of the impact of the incorporation of shape into the latest version of E-Face, an impressive average identification rate of 86% was found when an experienced E-Face operator was used.

The "from memory" results of this experiment are quite comparable to those found by Tredoux et al. (1999), as they discovered that "from memory" E-Face reconstructions of two faces (one familiar and one unfamiliar), following a 15-second exposure, received an average rating score of 60% and an identical average identification rate of 38%. The average identification rate of 43% found for their "from memory" E-Face reconstructions of two unfamiliar white faces (Tredoux et al. 1999); was also not too far off from that found in this experiment. Neither was the 42% "from memory" rating score found for E-Face composites, produced after exposure to a two-minute video-clip, in Prag's (2000) study. When testing the impact of the incorporation of shape into the latest version of E-Face, Tredoux et al. (2002) found a "from memory" identification rate of 50%, which again is not that much different to the results found in this experiment.

Other studies in which the "from memory" results of this experiment are better or quite comparable to include studies carried out by Christie and Ellis (1981), who found a "from memory" identification rate of 23% when Photofit likenesses were produced of a target face following a 60-second exposure; Wogalter and Marwitz (1991), who found a "from memory" matching score of 40% when Mac-a-Mug Pro composites were produced of a target face following a brief 8-second exposure; Christie et al. (1981), who found a "from memory" matching accuracy of 28% for CADC composites and 18% for Photofit composites; Davies et al. (2000), who found a "from memory" naming rate of 17% and a matching accuracy of 63% (though this was carried out without distractors) in their comparison between the Photofit and E-fit systems; Davies and Oldman (1999), who found a "from memory" naming rate of just 6% when participants were asked to produce E-fits of famous faces that they strongly liked or disliked; and Frowd (2001), who found a "from memory" naming rate of 10% for EvoFIT composites and 17% for E-fit composites.

All three evaluation tasks indicate that "in view" Faces composites are significantly better than "from memory" ones. The same cannot be said for E-Face composites. While the "in view" E-Face composites received better ratings than the "from memory" ones, the difference between their rating scores were found to be not significant. Even more intriguingly, both the sorting and ranking task results show that "in view" E-Face composites are less accurate than those produced during one of the "from memory" conditions, that being the SAPS condition. Another finding worthy of note is that the average of the two "from memory" identification rates is slightly higher than the average "in view" identification rate for the E-Face system. On the basis of these findings, it could be argued that the configural E-Face system is better able to reconstruct faces from memory than in view. However, this argument could easily be overruled by the fact that both the rating and ranking task results contradictorily show that "in view" E-Face reconstructions are more accurate than "from memory" ones. Also, the above inferences cannot be made from non significant results and "untested" comparisons.

An important finding from this experiment, which needs to be highlighted, is that the E-Face system performs considerably better than the Faces system under "from memory" conditions, on average. This is certainly the case for the SAPS condition, as was deduced from all three evaluation task results. For the CI condition, E-Face was found to be more effective than the Faces system for two of the three evaluation tasks (i.e. the rating and ranking tasks), but not significantly more effective.

Upon review of the results of this experiment, it cannot be disputed that the featural Faces system is better able to reconstruct faces in view than from memory. From this most interesting aspect of the results of this experiment, one could infer that the featural method of composite reconstruction is not best suited to real police investigations, and that it supports the existing evidence to suggest that individual features are not processed independently from other face information, but as the sum of its parts instead. This requires further investigation, but could point the way forward for a more rigorous comparison of the theoretical approaches (i.e. featural, configural and dual) with composite systems.

With the E-Face system, Pike (personal communication, June 20, 2002) believes that with its album approach, where operators are required to select the most similar-looking face from a selection of faces during each generation, there is likely to be the problem of "visual overshadowing", i.e. the idea that exposure to several faces will distort and eventually "overshadow" the perpetrator's face stored in the witness's memory. This needs to taken into consideration with the factors that influence the E-Face operators during their composite production process.

In general, research shows a reduction in the likelihood of an accurate identification as the number of mugshots viewed by the witness increases. For example, in an experiment conducted by Lindsay, Nosworthy, Martin and Martynuck (1994), participants were exposed to a staged theft. Thereafter, they were required to try to pick out the perpetrator from 100, 300, 500 or 700 mugshots. Participants' ability to correctly identify the perpetrator decreased as the numbers of mugshots increased (from 36%, 30%, 18% and 18% respectively). Although the mugshot search is a valued police procedure, some researchers have speculated that a sequential search through a large pool of mugshot photos is likely to degrade or confuse the witness's mental image of the perpetrator's face (Baker, 1999; Davies, Shepherd & Ellis, 1979; Pryke, Lindsay & Pozzulo, 2000). Therefore several researchers concluded that examining mugshots is a dangerous procedure likely to result in accusations against innocent people (Wells, 1988).

This suggests that E-Face operators who progressed through a large number of generations to arrive at their finally chosen composite, and as a result were exposed to a larger number of faces, are more likely to make an inaccurate choice. On the other hand, those who used a small number of generations, and as a results were exposed to a fewer number of faces, are more likely to make an accurate choice. This has not been sufficiently explored in this experiment, but would be an interesting matter of investigation in future experiments.

Boylan (2000) would also not favour the mugshot approach of the E-Face system, as she believes that "by showing eyewitnesses books of facial photos from which to choose look-alike images, the well-meaning police or police artists are literally handing to the eyewitness - at the very height of that person's vulnerability to suggestion and influence - the visual tools to effectively discard, distort or further entomb the actual image that created the trauma" (Boylan, 2000, p.15). Another potential disadvantage of the mugshot approach of the E-Face system is the problem of "criminal stereotyping", where witnesses may select the eigenfaces in each generation based on what they think a rapist, for example, looks like (Bull, personal communication, June 18, 2002).

An important advantage of the mugshot approach adopted by E-Face, which should not be overlooked, is that "it uses the natural human ability to recognise faces and thus enables specification of the query without requiring the user to articulate or even be consciously aware of what specific facial features are being sought" (Baker, 1999, p.18).

With the Faces system, Pike (personal communication, June 20, 2002) argues, features used to build the facial composites are not added and altered in the context of an average face, which is disadvantageous as it goes against face perception theory. With the E-fit system, for example, one inputs a verbal description, which yields a general face. Thereafter, each feature can be added and altered in the context of this face, which is more in line with face perception theory (Pike, personal communication, June 20, 2002).

This matter has been further investigated in an experiment conducted by Turner et al. (2000), who suspected that the quality of E-fits is affected by a form of "overshadowing", as a result of similar but incorrect feature exemplars being exposed to the witness during the composite production process. They suggested that one way of overcoming this would be to construct the E-fit within a "minimal face", which would provide a face-context without the possibility of overshadowing. For their experiment, each participant studied one of nine "front-on" target photographs for 2½ minutes. Thereafter participants proceeded in their attempts to produce "from memory" reconstructions of the target with the assistance of an E-fit operator (who had no exposure to the target). Prior to this, the operator briefed the participants on the capabilities of the system and the nature of their impending tasks. The operator also took a description from the participants and entered it into E-fit.

Turner et al. (2000) decided to use a schematic "minimal face", consisting of a simple oval with two filled circles for eyes and short lines for the nose and mouth. This minimal face was used in two reconstruction conditions, i.e. (1) "Piecemeal" where the participants selected each feature individually within the context of the minimal face, and (2) "Jigsaw" where the participants began as in the "Piecemeal" condition, but each feature was left on view while subsequent features were added. For the third reconstruction condition, the minimal face was not used and participants selected and/or manipulated the facial features within the context of the whole facial image, as is normally practiced by E-fit operators. Independent judges were required to evaluate each of the composites produced during the reconstruction phase, by giving them similarity ratings to their original targets (Turner et al., 2000).

The results revealed that minimal "Jigsaw" composites were rated significantly higher than the "normal E-fits", and that the minimal "Piecemeal" composites were rated slightly lower than the "normal E-fits". Turner et al. (2000) concluded that although the minimal face is sufficient to elicit a "face schema", it is not enough in itself to outweigh the benefits of seeing a whole facial image. They surmise that in the minimal "Piecemeal" condition, there may simply not be enough information available to the witness (e.g. configural or relational information), to produce an accurate likeness, and so the minimal face is not beneficial. However, in the "Jigsaw" condition this information becomes more and more available, which possibly explains the relative superiority of minimal "Jigsaw" composites. With the Faces system, facial features are not selected and/or manipulated within the context of a whole facial image. However, facial composites are compiled in a "Jigsaw" manner, which seems to be advantageous, as Turner et al.'s (2000) discovered. The effect of compiling facial composites within the context of a whole facial image has not been thoroughly investigated in this experiment, but should definitely be a matter of further exploration in future studies.

Regarding the comparison between the two "from memory" conditions, there is the suggestion, from the overall results of this experiment, that the CI is less useful a technique than the SAPS Interview in helping witnesses recall the face of the target. While there were no significant differences between their average rating scores, SAPS composites did fair 1% greater than CI composites. SAPS composites were also identified better (in the sorting task) and ranked significantly higher than CI composites. For the E-Face system, composites produced under the SAPS condition were superior to those produced under the CI condition for all of the three evaluation tasks. For the Faces system, while their identification rates were identical, CI composites were rated and ranked just 1% higher than SAPS ones. For all targets except Target 1 (in the rating and ranking tasks) and Target 2 (in the sorting task), SAPS composites performed better than CI ones.

A possible explanation for the poorer performance of the CI, specially developed for the purpose of facial composite production, relates to debate surrounding the incorporation of "imagery" as part of the CI technique, i.e. getting witnesses to visualise, acquire and maintain a mental image of their best view of the perpetrator. According to Stevenage (1997), the use of imagery in the CI was promoted because it was believed to facilitate the retrieval of information. For example, when encouraged to imagine a crime scene, witnesses' recollections of that scene are typically richer than if no imagery is used (Bekerian, Dennett, Hill & Hitchcock, 1990; Dobson & Markham, 1993). Also, Pavio (1971) clearly demonstrated that one's memory for word pairs was enhanced when participants were encouraged to picture the two objects together.

On the other hand, there are some researchers, like Clark (2000) and Pike (personal communication, June 20, 2002) for example, who strongly oppose the incorporation of imagery as part of the CI. In Clark's (2000) experiment, participants were shown a one-minute video-clip of a criminal incident, followed by a ten-second still of the perpetrator's face at the end of the video. Approximately 22 hours later, the participants were asked to produce a composite of the perpetrator using the E-fit composite system with the aid of a skilled operator. A total of 64 composites were produced using eight methods, consisting of eight combinations of three specific techniques.

An overall identification rate of 38% was found for all composites, which were evaluated by displaying them where people who knew the perpetrators would clearly see them. When the number of composites correctly identified for each of the eight methods was compared, Clark (2000) found that the four lowest scoring methods all used one common technique, i.e. the instruction from the operator to the witness, requesting the witness to image the suspect in their "minds eye". An identification rate of only 18% was found when witnesses were instructed to image, compared to an identification rate of 57% found when no instruction was given. As can be seen, Clark (2000) found a clear association between the instruction to image and the amount of composites identified, and therefore recommends that witnesses should not be instructed to image during the composite production process. He also recommends that further research be carried out to establish the root cause of the effects of imaging on the identification rates of composites and the duration of the effect.

Another factor that could play a part in the accuracy of CI composites is the fact that "from memory" reconstructions were made between one and fifteen day(s) after participants were exposed to a target in the live staged events of this experiment. The resultant "delay effect" was not taken into consideration in the results analysis of this experiment, as it would increase the number of factors in the design, thereby complicating it and sacrificing power. It could be argued that the problem of delay could have been eliminated in this experiment by replacing the live staged event with a videotaped criminal incident, thereby promoting standardisation (Pike, personal communication, June 20, 2002). However, a videotaped incident would not be an ecologically valid method of target exposure, particularly because of the CI reconstruction condition, in which it is essential for witnesses to be exposed to the target in a "live" setting. Staging live events also have the advantage of evaluating the facial composite systems under realistic circumstances.

While previous research provides evidence that the perceived quality of composites does not decline with a delay between exposure to the face and the reconstruction process (McNeil, Wray, Hibler, Foster, Rhyne & Thibault, 1987; Davies, Ellis & Shepherd, 1978; Ellis, Davies & Shepherd, 1978), there are several researchers who would promote a reasonable delay between target exposure and facial composite production. It has been argued that a critical variable that determines the effectiveness of the CI is the delay between the incident and test. Malpass (1996) has pointed out that if the eyewitness has a clear, accurate memory for the "focal" element of the event (i.e. the part of the event about which the interviewer is trying to elicit information), then additional contextual cues are not likely to be useful (Smith, 1988). However, if memory for the "focal" element is weak (which is likely to happen as the delay progresses) then contextual reinstatement should be beneficial (Malpass, 1996). Green and Geiselman (1989) similarly advise, on the basis of the results of their study, that a reasonable delay between exposure to the face and reconstruction should occur to achieve best results. While it would have been useful to explore the effect of delay on the facial composites produced, it was unfortunately not a practical option in this experiment.

Besides the "delay effect", another confounding variable in this experiment relates to the performance of the interviewer. According to Memon, Wark, Holley, Bull and Koehnken (1996, p.153), "interviewer behaviour in the CI needs careful monitoring and there should be opportunities to provide regular feedback on performance so that interviewers can become more proficient and comfortable in the use of CI techniques. Unless this is taken on board, the outcome of using a CI is likely to vary from one interviewer to the next". This particular confounding variable is applicable to the SAPS reconstruction condition as well.

Of all composites produced by participants in three conditions of the reconstruction phase (i.e. SAPS Interview, CI and Novice IV), composites of Target 3 received the highest average rating score, a score that was significantly better than that of two other targets. Target 3 also received the highest identification rate of 54%, which suggests that it was the easiest of the four targets to reconstruct. In fact, both expert operators gave Target 3 their highest satisfaction ratings, with the E-Face expert giving it 86% and the Faces expert giving it a perfect 100% satisfaction rating. While Target 3 may have performed the best in all evaluation tasks when the Faces system was used, this was certainly not the case when the E-Face system was used. When the E-Face system was used, inconsistent results were revealed as to which target was easiest or most difficult to reconstruct.

Similar inconsistent results were revealed when all composites produced by participants in the two "in view" conditions (i.e. Novice IV and Expert IV) were evaluated. This suggests that the two facial composite systems vary in their ability to produce accurate representations of the same target face. Nonetheless, there appears to be the underlying implication that certain faces are more difficult to reconstruct than others, irrespective of which facial composite system is used. In order to minimise and/or eliminate this problem, the faces / facial features that make up the respective databases of the two systems need to be expanded so that they are more representative of all types of faces.

However even if this problem is resolved, as has been mentioned previously, people differ quite markedly in their ability to reconstruct faces (Ellis et al., 1975). The variation in the accuracy of the composites produced indicates that certain participants have the ability to reconstruct faces better than others using this composite system, which suggests a clear witness effect. Some achieved reasonable likenesses to the original faces while others made quite poor reconstructions.

With regards to who should have control over the facial composite system during the facial reconstruction process, there are many who emphatically believe that participants, used during experiments conducted to test the effectiveness of facial composite systems, should not have full control over the facial composite system(s) being evaluated. Instead, they advise that an expert operator take full control of the facial composite system(s) being evaluated, while the participants act as "describers" to the expert operator (e.g. Pike, personal communication, June 20, 2002). Rios (personal communication, May 2, 2002) claims that if eyewitnesses were to take full control of the facial composite system, they are likely to look to the investigator for approval, and thus produce of an inaccurate facial composite.

According to forensic artist, Taylor (2000, p.204), "some vendors of computer-generating softwares have even suggested that crime victims prepare their own composite images. Such a practice may be particularly unwise, especially if done outside the oversight of law enforcement". The designers of E-fit do not advocate that novice users operate their system. Even when inputting the facial description into E-fit, the designers of E-fit have suggested that this be done in front of the witness or with the witness's direct assistance. This is because the witness may feel obligated to complete description boxes that they are unsure about, as a result of being faced with a set of forced option questions. The designers claim that these should be avoided at all costs as they may lead to misinformation and cause the witness to confuse the memory of the perpetrator with the memory of giving the description (Clark, 2000). Taylor (2000, p.205) goes on to say that every effort should be made "to ensure that sensitive, skilled interviewers operate computer composite systems. Ideally, those who operate computer composite-generating systems should have training in interviewing specifically for the purpose of producing a facial image".

Levi (personal communication, June 7, 2002), from the Israeli police department, also disagrees with leaving the critical job of constructing the composite in the sole hands of the witness. He advises that experiments conducted to test facial composite systems should represent reality as much as is possible, and therefore suggests that either the experimenter becomes trained in the system being evaluated or a trained operator be used. This has, however, not been practical for this experiment, and for several experiments conducted in the past, due to large sample sizes used. Hence, participants have been given full control of the facial composite system(s) during experimentation, despite the unfortunate fact that it does not represent reality. Further research should be conducted to compare composites produced using "describers" (participants who work in collaboration with an expert facial composite operator to compile the facial composite) with those produced by participants who take sole control of the facial composite system during the composite production process.

While we have attempted to conduct this experiment under realistic, ecologically valid conditions, there appears to be a few ways of making it even more realistic and ecologically valid. In real police investigations, composites are most useful when the police officer that looks at a composite knows the target. The composite may remind a police investigator of a particular suspect should there be any similarities, or exclude a known suspect from the investigation should there be major differences. Rarely can a police officer make use of a composite when he or she is unfamiliar with the suspect. Koehn et al. (1999) therefore suggest that, in laboratory experiments, judges be asked to evaluate the composites by comparing them against people whom they are familiar with, not against people whom they do not know.

Another suggested offered by Koehn et al. (1999) to improve the ecological validity of laboratory experiments, which would apply to this experiment, is to make a comparison between a verbal description only and a verbal-description-plus-composite, instead of asking judges to evaluate the facial composite in isolation. This is because when facial composites are used in a police investigation, they typically accompany a verbal description. According to Koehn et al. (1999, p.21), "the value of the facial composite then is the degree to which it enhances the investigation beyond that of using only the verbal description, not whether the facial composite, by itself, can lead to an identification".

Chapter 8: Conclusion

Experiment 1 demonstrated that the E-Face system compares quite well against previously conducted studies, where novice operators have used various facial composite systems to produce facial composites under "in view" conditions. Experiment 2 showed that the E-Face system is marginally poorer than the Faces system, a system that is currently being used by law enforcement agencies worldwide, under varying conditions. Even though the difference between the effectiveness of the two systems was not significant, much optimism can come out of the overall results of these two experiments for the E-Face system, especially considering that it is still in the process of development, and that further refinements made to improve the system, are yet to be evaluated.

One of the refinements, referred to above, is the "feature builder". Although the original starting point of the E-Face system was to try to avoid featural reconstruction, witness feedback and direct observation has persuaded the need for the addition of some sort of featural search into the system. This is the reason behind the recent introduction of the "feature builder" to the configural basis of the E-Face system, which is indicative of the "dual" face processing theory. According to Frowd (2001, p.27), "...since an exclusive holistic bias may not be the best system for analytical encoders, a hybrid holistic-componential photofitting approach may be optimal for a witness..."

The feature builder allows individual features of the finally chosen face to be altered and manipulated. Therefore, if the lips of the finally chosen face needed to be altered, for example, the feature builder would keep that face constant but expand the range of lips (extracted from the relevant E-Face database) on that face. To clarify, E-Face would produce a new generation of 18 faces; the lips on each of these 18 faces would vary, while the rest of the features on each of these faces would remain constant. The witness would then select the face with the most similar-looking lips to that of the perpetrator. Other facial features would be altered in much the same manner (Tredoux, personal communication, December 6, 2004). The incorporation of the feature builder responds to the verbal feedback, provided by many participants in the reconstruction phase of both experiments, that the E-Face system could be improved by being able to manipulate and refine the finally chosen eigenface.

In addition to the feature builder, another recent development to E-Face is the "accelerator", which allows for the search gradient to be controlled. To clarify, during the composite production process, if the witness finds a face in a certain E-Face generation that looks quite similar to that of the perpetrator, then the accelerator can be used to produce other faces similar to that particular face. This useful feature speeds up the composite production process, and minimises the likelihood of "visual overshadowing" as the witness is likely to be exposed to less faces before making his/her final choice. The accelerator can also be used to make the spread of eigenfaces in a certain generation look less similar if the witness feels that they are getting too similar. This particular new feature would resolve a complaint, lodged by several participants in both experiments of this study, that the eigenfaces of the E-Face system become too similar too quickly when progressing from one generation to the next. At present, a beta version of the new refinements made to the E-Face system exists, which will go into testing in due course (Tredoux, personal communication, December 6, 2004).

A key finding from Experiment 1 is that it seemed to disprove the hypothesis that the E-Face system would be improved if the Indian faces were removed from the coloured database, and used to create a new Indian database. In a measurement of the "other-race effect" in this experiment, the results did not provide clarity as to which of the three races (black, white or coloured) is better at reconstructing coloured and Indian faces.

In Experiment 2, E-Face was found to perform slightly poorer than Faces. It could therefore be inferred that a featural method of facial reconstruction should be recommended over a configural method. However, since the difference between the effectiveness of the two systems was not significant under the varying conditions they were put through, such a suggestion may not be substantiated. The most interesting aspect to come out of the results of Experiment 2 is that the Faces system performed considerably better under "in view" conditions than under "from memory" ones. From this, one could infer that the featural method of composite reconstruction is not best suited to real police investigations, and that it supports the existing evidence to suggest that individual features are not processed independently from other face information. This could point the way forward for a more rigorous comparison of the theoretical approaches (i.e. featural, configural and dual) with composite systems. Another important finding to come out of this experiment is that, on average, the E-Face system performs considerably better than the Faces system under "from memory" conditions; though not significantly better for all conditions.

Since the SAPS Interview was found to be only marginally more useful a technique than the CI in helping witnesses recall the face of the target, it would be difficult to advise with confidence which interview technique should be adopted by police investigators. Nonetheless, the fact that the CI technique, (formulated for the specific task of facial composite production) compared well against the SAPS Interview, suggests that this technique has potential to serve as an alternative to the standard police interview, particularly if it is further enhanced.

Koehn et al. (1999) believe that the utility of the interview strategy should depend upon the specific composite system being used, and that it should be modified according to the structural constraints of the system. They also claim that the interview technique should be adapted according to the type of face being reconstructed. For example, "a featurally-distinctive face may be better described when questions focus on specific features, whereas a more typical face may be better described from a more holistically-oriented interviewing strategy" (Koehn et al., 1999, p.20). It goes without saying that this should be a matter for further investigation.

The overall results of both experiments of this study are in line with the bulk of similar previously conducted studies. These studies have revealed that the efficacy of composite systems as tools to promote recognition of suspects in criminal contexts is questionable. Extensive empirical examination, including the two experiments conducted as part of this thesis, indicate that composite systems produce poor quality composites, which are difficult to match to target faces (Christie & Ellis, 1981; Duggal, Mickus, Daneker & Kassin, 1992; Davies, Van der Willik & Morrison, 2000;

Ellis, Shepherd & Davies, 1975; Koehn et al., 1999; Koehn & Fisher, 1997; Kovera, Penrod, Pappas & Thill, 1997; Laughery & Fowler, 1980; Yount & Laughery, 1982).

Facial composite systems, including the ones evaluated in this thesis, have generally been found to be more effective under non-ecologically valid, unrealistic experimental conditions, such as when the target was in full view whilst novice and/or expert operators constructed the facial composite, or when witnesses are exposed to the target and then anticipate the future task of constructing a facial composite (e.g. Cutler, Stocklein & Penrod, 1988; Koehn et al., 1999; Penrod & Stocklein, 1992; Wells & Hyrciw, 1984; Wogalter & Marwitz, 1991).

Many composites indeed look like the perpetrator, but the problem is that they also look like many other people. Still, from a practical point of view, facial composite and generation systems are invaluable in police investigations, in spite of their shortcomings. Very often in police investigations a composite of a suspect's face is the only lead at hand. As Brignull (1998) asserts, until a practical alternative has been developed, these systems are simply the best available.

References

Ainsworth, P. B. (1998). *Psychology, law and eyewitness testimony*. New York: John Wiley & Sons.

Aspley Ltd. (2000). *E-Fit for windows: on-line help file*. Hatfield, UK: Aspley Ltd.

Baddeley, A. & Woodhead, M. (1983). Improving face recognition ability. In S. M. A. Lloyd-Bostock & B. R. Clifford (Eds.), *Evaluating eyewitness evidence: Recent psychological research and new perspectives* (pp.125-136). Chichester: John Wiley & Sons.

Baker, E. (1999). *The mug-shot search problem: a study of the eigenface metric, search strategies, and interfaces in a system for searching facial image data.* Unpublished Manuscript, Harvard University, Massachusetts.

Barber, P. (1988). *Applied cognitive psychology*. London: Methuen.

Bartlett, J. C., & Searcy, J. H. (1993). Inversion and configuration of faces. *Cognitive Psychology, 25*, 281-316.

Bekerian, D. A., & Dennett, J. L. (1995). Developmental differences in eyewitness suggestibility and memory for source. *Journal of Experimental Child Psychology, 60*, 57-83.

Bekerian, D. A., & Dennet, J. L. (1993). The cognitive interview technique: Reviving the issues. *Applied Cognitive Psychology, 7*, 275-297.

Bennett, P. (1986). Face recall: a police perspective. *Human Learning, 5*, 197-202.

Bothwell, R. K., Brigham, J. C., & Malpass, R. S. (1989). Cross-racial identification. *Personality and Social Psychology Bulletin, 15*, 19-25.

Bower, B. (1997). Thanks for the memories. *Science News, 151*, 246-247.

Boylan, J. (2000). *Portraits of guilt: The woman who profiles the faces of America's deadliest criminals*. USA: Pocket Books.

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

Bradshaw, J. L., & Wallace, G. (1971). Models for the processing and identification of faces. *Perception and Psychophysics, 9*, 443-448.

Brignull, H. (1998). *Why do face reconstruction systems give such poor results and how might they be improved?* [On-Line]. Available: http://neocortex.co.uk/essays/applied/identikit.htm [2000, September 6].

Bruce, V. (1988). *Recognising faces*. London: Lawrence Erlbaum Associates Publishers.

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

Bruyer, R., & Coget, M. C. (1987). Features of laterally displayed faces: Saliency or top-down processing? *Acta Psychologica, 66*, 103-114.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, B. (1999). Face recognition in poor quality video: evidence from security surveillance. *Psychological Science, 10* (3), 243-248.

Cabeza, R., & Kato, T. (2000). Features are also important: contributions of featural and configural processing to face recognition. *Psychological Science, 11*(5). 429-433.

Caldwell, C., & Johnston, V.S. (1991). Tracking a criminal suspect through "face-space" with a genetic algorithm. *Proceedings of the Fourth International Conference on Genetic Algorithms*, 416-421. Morgan Kaufmann Publishers.

Campbell, R., Coleman, M., Walker, J., Benson, P. J., Wallace, S., Michelotti, J., & Baron-Cohen, S. (1999). When does the inner-face advantage in familiar face recognition arise and why? *Visual Cognition, 6*(2), 197-216.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science, 195*, 312-313.

Chance, J. E., & Goldstein, A. G. (1996). The other-race effect and eyewitness identification. In S. L. Sporer, R. S. Malpass & G. Koehnken (Eds.), *Psychological issues in eyewitness identification*. Hillsdale, New Jersey: Lawrence Erlbaum.

Chance, J. E., & Goldstein, A. G. (1976). Recognition of faces and verbal labels. *Bulletin of the Psychonomic Society, 7*, 384-386.

Christie, D., Davies, G., Shepherd, J., & Ellis, H. (1981). Evaluating a new computer-based system for face recall. *Law and Human Behaviour, 5*(2), 209-218.

Christie, D., & Ellis, H. D. (1981). Photofit construction versus verbal descriptions of faces. *Journal of Applied Psychology, 66*, 358-363.

Clark, C. (2001). *Interviewing for facial identification: part II.* Home Office PRCU Publication.

Clark, C. (2000). *Interviewing for facial identification.* Home Office PRCU Publication.

Clifford, B. R., & Davies, G. M. (1989). Procedures for obtaining identification evidence. In D. C. Rankin, *Psychological methods in criminal investigation and evidence* (pp. 47-96). New York: Springer.

Collishaw, S. M., & Hole, G. J. (2000). Featural and configural processes in the recognition of faces of different familiarity. *Perception, 29*, 893-909.

Cootes, T. F., & Taylor, C. J. (2001). *Statistical models of appearance for computer vision.* Unpublished manuscript, University of Manchester.

Cutler, B. L., Stocklein, C. J., & Penrod, S. D. (1988). Empirical examination of a computerized facial composite production system. *Forensic Reports, 1*, 207-218.

Davies, G. M. (1983). Forensic face recall: the role of visual and verbal information. In S. M. A. Lloyd-Bostock & B. R. Clifford (Eds.), *Evaluating eyewitness evidence: Recent psychological research and new perspectives* (pp.103-123). Chichester: John Wiley & Sons.

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

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

Davies, G., & Milne, A. (1985). Eyewitness composite production a function of mental or physical reinstatement of context. *Criminal Justice and Behaviour, 12*, 209-220.

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

Davies, G., Shepherd, J., Ellis, H. (1979). Effects of interpolated mugshot exposure on accuracy of eyewitness identification. *Journal of Applied Psychology, 64*, 232-237.

Davies, G. M., 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.

Deffenbacher, K. A. (1988). Eyewitness research: the next ten years. In M. M. Gruneberg, P. E. Morris & R. N. Sykes (Eds.), *Practical aspects of memory: current research and issues* (pp. 20-26). Chichester, England: John Wiley & Sons.

Dobson, M., & Markham, R. (1983). Imagery ability and source monitoring: implication for eyewitness memory. *British Journal of Psychology, 84*(1), 111-118.

Dodson, C. S., Johnson, M. K., & Schooler, J. W. (1997). The verbal overshadowing effect: Why descriptions impair face recognition. *Memory and Cognition, 25*(2), 129-139.

Duggal, S., Mickus, L., Daneker, E., & Kassin, S. (1992). *Computerized facial composites and eyewitness recognition*. Poster presented at the meetings of the Eastern Psychological Association in Boston, Williams College.

Ellis, H. D. (1986). Face recall: a psychological perspective. *Human Learning, 5*, 1-8.

Ellis, H. D. (1984). Practical aspects of face memory. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness Testimony* (pp.12-37). Cambridge: Cambridge University Press.

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

Ellis, H., Davies, G. M., & Shepherd, J. (1978b). 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. (1992). Face memory - theory and practice. In M.M. Gruneberg & P.E. Morris (Eds.), *Aspects of Memory, Vol. 1, The practical aspects* (2nd Edition). London: Routledge.

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.

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. (1975). An investigation of the use of the Photofit technique for recalling faces. *British Journal of Psychology, 66*(1), 29-37.

Endo, M., Takahashi, K., & Maruyama, K. (1984). Effects of observer's attitude on the familiarity of faces: using the difference in cue value between central and peripheral facial elements as an index of familiarity. *Tohoku Psychologica Folia, 43*, 23-24.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105*, 482-498.

Finger, K., & Pezdek, K. (1999). The effect of the cognitive interview on face identification accuracy:

release from overshadowing. *Journal of Applied Psychology, 84*(3), 340-348.

Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing techniques for investigative interviewing.* Springfield: Charles C. Thomas.

Fisher, R. P., Geiselman, R. E., & Amador, M. (1989). Field test of the cognitive interview: Enhancing the recollection of actual victims and witnesses of crime. *Journal of Applied Psychology, 74*, 722-727.

Fisher, R. P., Geiselman, R. E., & Raymond, D. S. (1987a). Critical analysis of police interview techniques. *Journal of Police Science and Administration, 15*, 177-185.

Fisher, R. P., Geiselman, R. E., Raymond, D. S., Jurkevich, L. M., & Warhaftig, M. L. (1987). Enhancing eye-witness memory: refining the cognitive interview. *Journal of Police Science and Administration, 15*, 291-297.

Fisher, R. P., & McCauley, M. R. (1995). Information retrieval: interviewing witnesses. In N. Brewer & C. Wilson (Eds.), *Psychology and Policing*. Hillsdale, NJ: Lawrence Erlbaum.

Fisher, R. P., McCauley, M. R., & Geiselman, R. E. (1994). Improving eyewitness testimony with the cognitive interview. In D. F. Ross, J. D. Read & M. P. Toglia (Eds.), *Adult eyewitness testimony* (pp. 245-268). Cambridge: Cambridge University Press.

Fraser, I., & Parker, D. (1986). Reaction time measures of feature saliency in perceptual integration task. In H. Ellis, M. Jeeves, F. Newcombe & A. Young (Eds.), *Aspects of face processing.* Dordrecht: Martinus Nijhoff.

Friedman, N. P., Reed, S. K., & Carterette, E. C. (1971). Feature saliency and recognition memory for faces. *Perception and Psychophysics, 10*, 47-50.

Frowd, C. (2001). *EvoFit: a holistic, evolutionary facial imaging system.* Unpublished Manuscript, University of Stirling.

Frowd, C., McQuiston-Surret, D., Ness, H., Hancock, P. (2004). *The process of facial composite production.* Paper presented at the European Conference of Psychology and Law, Krakow, Poland.

George, R. C., & Clifford, B. R. (1996). The cognitive interview - does it work? In G. Davies, S. Lloyd-Bostock, M McMurran & C. Wilson (Eds.), *Psychology, law and criminal justice: international developments in research and practice* (pp.146-154). Berlin: Walter de Gruyter.

Geiselman, R. E. (1999). Commentary on recent research with cognitive interview. *Psychology, Crime and Law, 5*, 197-202.

Geiselman, R. E., Fisher, R. P., Firstenberg, I., Hutton, L. A., Sullivan, S., Avetissian, I., & Prosk, A. (1984). Enhancement of eyewitness memory: an empirical evaluation of the cognitive interview. *Journal of Police Science and Administration, 12*, 74-80.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1986). Enhancement of eye-witness memory with the cognitive interview. *American Journal of Psychology, 99*, 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 and Law, 1*, 93-100.

Gillenson, M. L., & Chandrasekeran, B. (1975). A heuristic strategy for developing human facial images on a CRT. *Pattern Recognition, 7*, 187-196.

Goldstein, A. G., & Mackenberg, E. J. (1966). Recognition of human faces from isolated facial features: A developmental study. *Psychonomic Science, 6*, 149-150.

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

Haig, N. D. (1986). Exploring recognition with interchanged features. *British Journal of Psychology, 76*, 156-170.

Haig, N. D. (1984). The effect of feature displacement on face recognition. *Perception, 13*, 505-512.

Hancock, P. B. J., & Frowd, C. D. (2002). Evolutionary generation of faces. In P. J. Bentley and D. W. Corne (Eds), *Creative Evolutionary Systems*. San Francisco: Morgan Kaufmann.

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

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

Hines, D., & Braun, J. A. (1990). Order of feature recognition in familiar and unfamiliar faces. *Brain and Cognition, 14*, 165-184.

Hole, G. J. (1994). Configurational factors in the perception of unfamiliar faces. *Perception, 23*, 65-74.

Homa, D., Haver, B., & Schwartz, T. (1976). Perceptibility of schematic face stimuli: Evidence of a perceptual gestalt. *Memory and Cognition, 4*, 176-185.

Hosie, J. A., Ellis, H. D., & Haig, N. D. (1988). The effect of feature displacement on the perception of well-known faces. *Perception, 17*, 461-474.

Interquest Ltd. (1998). *Faces 3.0 user's guide*. Quebec, Canada: Interquest Ltd.

Kirby, M., & Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterisation of human faces. *IEEE Transactions of Pattern Analysis and Machine Intelligence, 12*(1).

Koehn, C. E. (1995). *Theoretical and applied evaluations of facial composite systems*. Unpublished Manuscript, Florida International University, Florida.

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

Koehn, C. E., Fisher, R. P., & Cutler, B. L. (1999). Using cognitive interviewing to construct facial composites. In D. Canter and L. Alison (Eds.), *Interviewing and Deception* (Vol. 2). Ashgate: Dartmouth.

Köhnken, G. (1995). Interviewing adults. In R. Bull and D. Carson (Eds.), *Handbook of psychology in legal contexts*. West Sussex, England: John Wiley & Sons.

Köhnken, G., Milne, R. Memon, A., & Bull, R. (1999). A meta-analysis on the effects of the Cognitive Interview. *Special Issue of Psychology, Crime & the Law, 5*, 3-27.

Kovera, M. B., Penrod, S. D., Pappas, C., & Thill, D. L. (1997). Identification of computer-generated facial composites. *Journal of Applied Psychology, 82*(2), 235-246.

Krafka, C., & Penrod, S. (1985). Reinstatement of context in a field experiment on eyewitness identification. *Journal of Personality and Social Psychology, 49*(1), 58-69.

Laughery, K. R., & Fowler, R. H. (1980). Sketch artist and Identikit procedures for recalling faces. *Journal of Applied Psychology, 65*, 307-316.

Laughery, K. R., Alexander, J. F., & Lane, A. B. (1971). Recognition of human faces: effects of target exposure time, target position, pose position, and type of photograph. *Journal of Applied Psychology, 51*, 477-483.

Leder, H., & Bruce, V. (2000). When inverted faces are recognised: the role of configural information in face recognition. *Quarterly Journal of Experimental Psychology, 53A*, 513-536.

Light, L. L., Kayra-Stuart, F., & Hollander, S., (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory, 5(3)*, 212-228.

Lindsay, R. C. L., Nosworthy, G. J., Martin, R., & Martynuck, C. (1994). Using mug shots to find suspects. *Journal of Applied Psychology, 79*, 121-130.

Lindsay, R. C. L., & Wells, G. L. (1983). What do we really know about cross-race eyewitness identification? In S. M. A. Lloyd-Bostock & B. R. Clifford (Eds.), *Evaluating witness evidence: recent psychological research and new perspectives*. New York: Wiley.

Luu, T. N., & Geiselman, R. E. (1993). Cognitive retrieval techniques and order of feature construction in the formation of composite facial images. *Journal of Police and Criminal Psychology, 9*, 34-39.

Malpass, R. S., & Devine, P. G. (1981). Guided memory in eyewitness identification. *Journal of Applied Psychology, 66*(3), 343-350.

Matthews, M. L. (1978). Discrimination of Identikit construction of faces: Evidence for a dual processing strategy. *Perception and Psychophysics, 26*, 472-480.

McKelvie, S. J. (1973). The meaningfulness and meaning of schematic faces. *Perception and Psychophysics, 14*, 343-348.

McKelvie, S. J. (1976). The role of eyes and mouth in the memory of a face. *American Journal of Psychology, 89*, 311-323.

McNeil, J. E., Wray, J. L., Hibler, N. S., Foster, W. D., Rhyne, C. E., & Thibault, R. (1987). Hypnosis and Identi-kit: a study to determine the effect of using hypnosis in conjunction with the making of identikit composites. *Journal of Police Science and Administration, 15*, 63-67.

McQuiston, D. E., & Malpass, R. S. (2000). *Use of facial composite systems in U.S. law enforcement agencies*. Poster presented at the meetings of the American Psychology-Law Society in New Orleans, L.A.

Memon, A., & Bull, R. (1991). The cognitive interview: its origins, empirical support, evaluation and practical implications. *Journal of Community Psychology, 1*, 291-307.

Memon, A., Cronin, O., Eaves, R., & Bull, R. (1996). An empirical test of the mneumonic components of the cognitive interview. In G. Davies, S. Lloyd-Bostock, M. McMurran & C. Wilson. (Eds.), *Psychology, law and criminal justice: international developments in research and practice*. Berlin: Walter de Gruyter.

Memon, A., & Higham, P. (1999). A review of the cognitive interview. *Psychology, Crime and Law (Special Issue), 5*, 177-196.

Memon, A., Vrij, A., & Bull, R. (2003). *Psychology and law: truthfulness, accuray and credibility of* *victims, witnesses and suspects* (2nd Edition). Chichester: John Wiley & Sons.

Memon, A., Wark, L., Holley, A., Bull, R., & Koehnken, G. (1996). Interviewer behaviour in investigative interviews. *Psychology, Crime and Law, 3,* 135-155.

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

Meissner, C. A., Brigham, J. C., & Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. *Memory & Cognition, 29,* 176-186.

Milne, R., & Bull, R. (1999). *Investigative interviewing: psychology and practice*. New York: John Wiley & Sons.

Osterburg, J. W., & Ward, R. H. (1997). *Criminal investigation: a method of reconstructing the past.* Cincinnati: Anderson Publishing.

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 406-417.

Pavio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart and Winston.

Penrod, S., & Stocklein, J. (1992). *Assessing the impact of production methods on the recognizability of computer-generated facial composites*. Unpublished Manuscript, University of Minnesota.

Penry, J. (1974). Photo-fit. *Forensic Photography, 3*(7), 4-10.

Penry, J. (1971). *Looking at faces and remembering them*. London: Elek Books.

Penry, J. (1970). Photofit. *Police Journal*, 302-316.

Prag, B. (2000). *Investigating the effectiveness of a computer-driven system for reconstructing faces.* (Unpublished honour's thesis, University of Cape Town).

Pryke, S., Lindsay, R. C. L., & Pozzulo, J. D. (2000). Sorting mug shots: methodological issues. *Applied Cognitive Psychology, 14*, 81-96

Purcell, D. G., & Stewart, A. L. (1986). The face-detection effect. *Bulletin of Psychodynamic Society, 24*(2), 118-120.

Reynolds, J. K., & Pezdek, K. (1992). Face recognition memory: The effects of exposure duration and encoding instruction. *Applied Cognitive Psychology, 6*, 279-292.

Rosenthal, Y., de Jager, G., & Greene, J. (1998). *A computerised face recall system using eigenfaces.* Unpublished Manuscript, University of Cape Town, South Africa.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36-71.

Sergent, J. (1984). An investigation into component and configural processes underlying face recognition. *British Journal of Psychology, 75*, 221-242.

Shapiro, P. N., & Penrod, S.D. (1986). Meta-analysis of facial identification rates. *Psychological Bulletin, 100*, 139-156.

Shepherd, J., Davies, G. M., & Ellis, H.D. (1981). Studies of cue saliences. In G. M. Davies, H. D. Ellis, & J. W. Shepherd (Eds.), *Perceiving and remembering faces* (pp. 105-131). London: Academic Press.

Shepherd, J., Davies, G. M., & Ellis, H.D. (1978). How best shall a face be described? In M. M.

Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory*. New York: Academic Press.

Shepherd, J., & Ellis, H. (1996). Face recall – methods and problems. In S. L. Sporer, & R. S. Malpass G. Koehnken (Eds.), *Psychological issues in eyewitness identification*. Hillsdale, New Jersey: Lawrence Erlbaum.

Smith, E. E., & Nielsen, G. D. (1970). Representations and retrieval processes in short-term memory: recognition and recall of faces. *Journal of Experimental Psychology*, *85*, 397-405.

Stevenage, S. V., & Memon, A. (1997). On the importance of imagery in the cognitive interview: a reply to Roberts. *Psycholoquy*, *7*(6), witness memory.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, *46A*, 225-245.

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

Taylor, K. T. (2000). *Forensic art and illustration*. USA: CRC Press

Thompson, P. (1980). Margaret Thatcher: a new illusion. *Perception*. *9*, 483-484.

Tredoux, C., Nunez, D., Oxtoby, O., Rosenthal, Y., da Costa, L., Prag, B., & Pooley, S. (2002). *E-Face – an eigenface compositor*. Symposium presented at the meetings of the International Association of Applied Psychology, Singapore.

Tredoux, C., da Costa, L., Nunez, D., & Rosenthal, Y. (1999). *Face reconstruction using a configural, eigenface-based composite system*. Paper presented at SARMAC III, Boulder, Colorado.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*, 71-86.

Turner, J., Bennett, P., Pike, G., Towell, N., & Kemp. R. (2000). *Applying perceptual research to E-FIT construction: a minimal face experiment*. Paper presented at the European Conference of Psychology and Law, Cyprus.

Valentine, T. & Endo, M. (1992). Towards an exemplar model of face processing: the effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology*, *44A*, 671-703.

Walker-Smith, G. J. (1978). The effects of delay and exposure duration in a face recognition task. *Perception*, *6*, 63-70.

Walker-Smith, G. J., Gale, A. G., & Findlay, J. M. (1977). Eye movement strategies involved in face perception. *Perception and Psychophysics*, *24*, 63-70.

Wells, G. L. (1988). *Eyewitness identification*. Toronto: Ontario, Canada: Carswell

Wells, G. L., & Hryciw, B. (1984). Memory for faces: Encoding and retrieval operations. *Memory & Cognition*, *12*, 338-344.

Wogalter, M. S., & Marwitz, D. B. (1991). Face composite construction: in-view and from-memory quality and improvement with practice. *Ergonomics*, *34*(4), 459-468.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*, 747-759.

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

Appendices

Appendix A: MENSA Intelligence Test

You have to work out what the letters mean. See No 0 as an example.
According to MENSA, if you get 19 + of these, you are a "genius".
Only 2 MENSA members achieved full marks. See how well you do.
Scoring:1 to 5 is Average, 6 - 11 Somewhat Intelligent, 12 to 18 Intelligent, 19 + Genius

No.	Cryptic	Answer	
0	24 H in a D	24 HOURS IN A DAY	TRUE
1	26 L of the A		
2	7 D of the W		
3	7 W of the W		
4	12 S of the Z		
5	66 B of the B		
6	52 C in a P (WJs)		
7	13 S in the USF		
8	18 H on a G C		
9	39 B of the O T		
10	5 T on a F		
11	90 D in a R A		
12	3 B M (S H T R)		
13	32 is the T in D F at which W F		
14	15 P in a R T		
15	3 W on a T		
16	100 C in a R		
17	11 P in a F (S) T		
18	12 M in a Y		
19	13=UFS		
20	8 T on a O		
21	29 D in F in a L Y		
22	27 B in the N T		
23	365 D in a Y		
24	13 L in a B D		
25	52 W in a Y		
26	9 L of a C		
27	60 M in a H		
28	23 P of C in the H B		
29	64 S on a C B		
30	9 P in S A		
31	6 B to an O in C		
32	1000 Y in a M		
33	15 M on a D M C		

YOU SCORED 0

Appendix B: Results of cumulative binomial test (probability values in ascending order for each target)

Evaluator	T/Face	Hits	Misses	Cumulative Binomial *p*
9	cm1	12	0	0.000371*
7	cm1	10	2	0.000371*
10	cm1	9	3	0.000371*
12	cm1	9	3	0.000371*
16	cm1	9	3	0.000371*
14	cm1	8	4	0.000371*
15	cm1	8	4	0.000371*
19	cm1	8	4	0.000371*
20	cm1	8	4	0.000371*
2	cm1	7	5	0.000371*
8	cm1	6	6	0.000371*
11	cm1	6	6	0.000371*
1	cm1	5	7	0.000371*
3	cm1	5	7	0.000371*
5	cm1	5	7	0.000371*
13	cm1	5	7	0.000371*
18	cm1	5	7	0.000371*
17	cm1	4	8	0.010406*
6	cm1	3	9	0.085642
4	cm1	0	12	0.931281
9	cm2	11	1	0.000371*
1	cm2	9	3	0.000371*
10	cm2	9	3	0.000371*
16	cm2	9	3	0.000371*
19	cm2	9	3	0.000371*
20	cm2	9	3	0.000371*
2	cm2	8	4	0.000371*
6	cm2	8	4	0.000371*
14	cm2	7	5	0.000371*
7	cm2	6	6	0.000371*
8	cm2	6	6	0.000371*
11	cm2	6	6	0.000371*
17	cm2	6	6	0.000371*
3	cm2	5	7	0.000371*
13	cm2	5	7	0.000371*
15	cm2	5	7	0.000371*
18	cm2	5	7	0.000371*
5	cm2	3	9	0.085642
12	cm2	3	9	0.085642
4	cm2	2	10	0.322222
8	im1	10	2	0.000371*
18	im1	10	2	0.000371*
1	im1	9	3	0.000371*
7	im1	9	3	0.000371*
13	im1	9	3	0.000371*
16	im1	9	3	0.000371*
19	im1	8	4	0.000371*

20	im1	8	4	0.000371*
3	im1	7	5	0.000371*
12	im1	7	5	0.000371*
14	im1	7	5	0.000371*
17	im1	7	5	0.000371*
11	im1	5	7	0.000371*
2	im1	4	8	0.010406*
6	im1	3	9	0.085642
10	im1	3	9	0.085642
15	im1	3	9	0.085642
9	im1	2	10	0.322222
4	im1	1	11	0.677877
5	im1	1	11	0.677877
7	im2	9	3	0.000371*
16	im2	9	3	0.000371*
17	im2	9	3	0.000371*
2	im2	8	4	0.000371*
10	im2	8	4	0.000371*
9	im2	7	5	0.000371*
1	im2	6	6	0.000371*
8	im2	6	6	0.000371*
18	im2	6	6	0.000371*
3	im2	5	7	0.000371*
19	im2	5	7	0.000371*
6	im2	4	8	0.010406*
11	im2	4	8	0.010406*
12	im2	4	8	0.010406*
13	im2	4	8	0.010406*
4	im2	3	9	0.085642
5	im2	3	9	0.085642
14	im2	3	9	0.085642
15	im2	3	9	0.085642
20	im2	3	9	0.085642

Appendix C: Mean "other-race" reconstruction identification rates

Operator Race	T/Face	D/Base	Valid N	Hits	Identification Rate %
black			320	171	53.44
coloured			320	171	53.44
white			320	152	47.50
black	cm1		100	59	59.00
black	cm2		60	26	43.33
black	im1		120	63	52.50
black	im2		40	23	57.50
coloured	cm1		60	33	55.00
coloured	cm2		100	58	58.00
coloured	im1		80	40	50.00
coloured	im2		80	40	50.00
white	cm1		80	40	50.00
white	cm2		80	47	58.75

white	im1		40	19	47.50
white	im2		120	46	38.33
black		cm	140	85	60.71
black		im	100	49	49.00
black		ic	80	37	46.25
coloured		cm	60	36	60.00
coloured		im	100	50	50.00
coloured		ic	160	85	53.13
white		cm	120	56	46.67
white		im	120	64	53.33
white		ic	80	32	40.00

Appendix D: Mean "other-race" evaluation identification rates

Evaluator Race	T/Face	D/Base	Valid N	Hits	Identification Rate %
Black			336	163	48.51
coloured			336	163	48.51
White			288	168	58.33
Black	cm1		84	41	48.81
Black	cm2		84	43	51.19
Black	im1		84	42	50.00
Black	im2		84	37	44.05
coloured	cm1		84	52	61.90
coloured	cm2		84	45	53.57
coloured	im1		84	31	36.90
coloured	im2		84	35	41.67
White	cm1		72	39	54.17
White	cm2		72	43	59.72
White	im1		72	49	68.06
White	im2		72	37	51.39
Black		cm	112	58	51.79
Black		im	112	53	47.32
Black		ic	112	52	46.43
coloured		cm	112	63	56.25
coloured		im	112	53	47.32
coloured		ic	112	47	41.96
White		cm	96	56	58.33
White		im	96	57	59.38
White		ic	96	55	57.29

Appendix E: Individual composite rating means (out of 7, in descending order for each target)

T/Face	D/Base	Valid N	Mean	Minimum	Maximum	Std. Dev.
cm1	Ic	20	3.95	2	7	1.50
cm1	Cm	20	3.55	1	7	1.76
cm1	Cm	20	3.45	1	7	1.90
cm1	Im	20	3.40	1	6	1.76

cm1	Im	20	3.35	1	6	1.63
cm1	cm	20	3.30	1	7	2.00
cm1	cm	20	3.20	1	7	1.85
cm1	im	20	3.15	1	7	1.95
cm1	ic	20	3.00	1	6	1.52
cm1	im	20	3.00	1	7	1.72
cm1	ic	20	2.90	1	7	2.13
cm1	ic	20	2.90	1	6	1.55
cm2	ic	20	4.20	1	7	1.96
cm2	cm	20	3.95	1	7	2.04
cm2	im	20	3.90	1	7	2.00
cm2	cm	20	3.80	1	7	2.14
cm2	im	20	3.55	1	7	1.85
cm2	cm	20	3.45	1	7	1.85
cm2	ic	20	3.40	1	6	2.11
cm2	cm	20	2.85	1	6	2.08
cm2	im	20	2.80	1	6	1.70
cm2	ic	20	2.65	1	6	1.69
cm2	ic	20	2.60	1	7	1.96
cm2	im	20	2.50	1	5	1.54
im1	cm	20	3.85	1	7	1.95
im1	im	20	3.65	1	7	2.08
im1	ic	20	3.55	1	7	2.24
im1	im	20	3.55	1	7	1.70
im1	im	20	3.25	1	7	1.71
im1	cm	20	3.15	1	7	2.13
im1	im	20	3.05	1	7	1.93
im1	ic	20	3.00	1	7	1.95
im1	cm	20	3.00	1	7	1.86
im1	ic	20	3.00	1	7	2.10
im1	ic	20	2.90	1	7	2.05
im1	cm	20	2.75	1	7	1.74
im2	im	20	3.95	1	6	1.67
im2	ic	20	3.90	1	7	1.89
im2	cm	20	3.80	1	7	1.82
im2	im	20	3.70	1	7	2.15
im2	im	20	3.60	1	7	2.14
im2	cm	20	3.50	1	7	2.12
im2	im	20	3.45	1	7	1.85
im2	ic	20	3.20	1	7	1.61
im2	ic	20	3.15	1	7	2.18
im2	cm	20	2.95	1	7	1.96
im2	cm	20	2.35	1	7	1.81
im2	ic	20	1.95	1	5	1.15

Appendix F: Rating means (out of 7)

Reconstructor	Condition	Target	Valid N	Mean	Minimum	Maximum	Std.Dev.
			1440	2.90	1	7	1.68
Faces			720	2.93	1	7	1.76
E-Face			720	2.88	1	7	1.60
	SAPS		480	2.72	1	7	1.64
	CI		480	2.63	1	7	1.62
	Novice IV		480	3.37	1	7	1.69
		1	360	2.92	1	7	1.59
		2	360	2.77	1	7	1.66
		3	360	3.13	1	7	1.71
		4	360	2.80	1	7	1.75
Faces	SAPS		240	2.48	1	7	1.66
Faces	CI		240	2.54	1	7	1.59
Faces	Novice IV		240	3.78	1	7	1.72
E-Face	SAPS		240	2.95	1	7	1.60
E-Face	CI		240	2.72	1	7	1.64
E-Face	Novice IV		240	2.97	1	7	1.57
Faces		1	180	2.98	1	7	1.67
Faces		2	180	2.69	1	7	1.60
Faces		3	180	3.40	1	7	1.81
Faces		4	180	2.66	1	7	1.86
E-Face		1	180	2.85	1	7	1.52
E-Face		2	180	2.86	1	7	1.71
E-Face		3	180	2.86	1	7	1.55
E-Face		4	180	2.95	1	7	1.64
	SAPS	1	120	2.38	1	7	1.46
	SAPS	2	120	2.68	1	7	1.71
	SAPS	3	120	3.24	1	7	1.64
	SAPS	4	120	2.56	1	7	1.65
	CI	1	120	2.88	1	7	1.41
	CI	2	120	2.52	1	7	1.60
	CI	3	120	2.64	1	7	1.65
	CI	4	120	2.48	1	7	1.77
	Novice IV	1	120	3.49	1	7	1.71
	Novice IV	2	120	3.12	1	7	1.62
	Novice IV	3	120	3.50	1	7	1.72
	Novice IV	4	120	3.38	1	7	1.72
Faces	SAPS	1	60	2.22	1	7	1.50
Faces	SAPS	2	60	2.30	1	7	1.52
Faces	SAPS	3	60	3.48	1	7	1.69
Faces	SAPS	4	60	1.93	1	7	1.52
Faces	CI	1	60	2.78	1	7	1.44
Faces	CI	2	60	2.62	1	7	1.50
Faces	CI	3	60	2.70	1	7	1.80
Faces	CI	4	60	2.05	1	7	1.53
Faces	Novice IV	1	60	3.95	1	7	1.60
Faces	Novice IV	2	60	3.15	1	7	1.69
Faces	Novice IV	3	60	4.02	1	7	1.72
Faces	Novice IV	4	60	3.98	1	7	1.77

E-Face	SAPS	1	60	2.55	1	6	1.42
E-Face	SAPS	2	60	3.07	1	7	1.80
E-Face	SAPS	3	60	3.00	1	7	1.56
E-Face	SAPS	4	60	3.18	1	7	1.55
E-Face	CI	1	60	2.97	1	6	1.39
E-Face	CI	2	60	2.42	1	7	1.71
E-Face	CI	3	60	2.58	1	7	1.51
E-Face	CI	4	60	2.90	1	7	1.89
E-Face	Novice IV	1	60	3.03	1	7	1.70
E-Face	Novice IV	2	60	3.08	1	6	1.57
E-Face	Novice IV	3	60	2.98	1	6	1.58
E-Face	Novice IV	4	60	2.77	1	6	1.43

Appendix G: Individual composite rating means (out of 7, in descending order for each target)

Reconstructor	Condition	Target	C. Number	Valid N	Mean	Minimum	Maximum	Std.Dev.
Faces	Novice IV	1	3	20	4.30	2	7	1.87
Faces	Novice IV	1	1	20	4.15	1	6	1.42
E-Face	Novice IV	1	3	20	3.55	1	7	1.67
Faces	Novice IV	1	2	20	3.40	1	7	1.39
E-Face	Novice IV	1	1	20	3.30	1	6	1.69
E-Face	CI	1	3	20	3.25	1	6	1.41
Faces	CI	1	3	20	3.15	1	7	1.5
E-Face	CI	1	2	20	3.10	1	6	1.52
E-Face	SAPS	1	2	20	3.00	2	6	1.45
Faces	CI	1	1	20	3.00	1	5	1.38
Faces	SAPS	1	2	20	2.80	1	7	1.67
E-Face	SAPS	1	1	20	2.60	1	6	1.39
E-Face	CI	1	1	20	2.55	1	6	1.19
E-Face	Novice IV	1	2	20	2.25	1	6	1.52
Faces	CI	1	2	20	2.20	1	5	1.32
Faces	SAPS	1	3	20	2.15	1	4	0.99
E-Face	SAPS	1	3	20	2.05	1	6	1.32
Faces	SAPS	1	1	20	1.70	1	7	1.59
E-Face	SAPS	2	1	20	3.75	1	7	1.62
E-Face	Novice IV	2	1	20	3.55	1	6	1.76
E-Face	SAPS	2	3	20	3.55	1	7	1.7
E-Face	Novice IV	2	3	20	3.40	1	6	1.39
Faces	Novice IV	2	1	20	3.30	1	7	1.66
Faces	Novice IV	2	3	20	3.25	1	7	1.86
E-Face	CI	2	1	20	3.15	1	7	1.79
Faces	CI	2	2	20	3.00	1	6	1.45
Faces	Novice IV	2	2	20	2.90	1	6	1.59
Faces	SAPS	2	3	20	2.60	1	7	1.64
Faces	SAPS	2	2	20	2.50	1	6	1.47
Faces	CI	2	1	20	2.45	1	6	1.5
Faces	CI	2	3	20	2.40	1	7	1.54
E-Face	Novice IV	2	2	20	2.30	1	5	1.26
E-Face	CI	2	2	20	2.30	1	7	1.69

E-Face	SAPS	2	2	20	1.90	1	6	1.55
E-Face	CI	2	3	20	1.80	1	7	1.44
Faces	SAPS	2	1	20	1.80	1	7	1.4
Faces	Novice IV	3	3	20	4.65	2	7	1.73
Faces	Novice IV	3	2	20	3.85	1	6	1.63
Faces	SAPS	3	3	20	3.70	2	6	1.45
Faces	Novice IV	3	1	20	3.55	1	6	1.7
E-Face	SAPS	3	2	20	3.50	1	7	1.76
Faces	SAPS	3	1	20	3.40	1	7	1.7
Faces	SAPS	3	2	20	3.35	1	7	1.95
E-Face	Novice IV	3	2	20	3.20	1	6	1.74
Faces	CI	3	1	20	3.15	1	6	1.84
E-Face	CI	3	1	20	3.10	1	6	1.62
E-Face	Novice IV	3	1	20	2.95	1	6	1.67
Faces	CI	3	2	20	2.95	1	7	1.9
E-Face	Novice IV	3	3	20	2.80	1	6	1.36
E-Face	SAPS	3	1	20	2.80	1	6	1.47
E-Face	SAPS	3	3	20	2.70	1	6	1.38
E-Face	CI	3	2	20	2.35	1	5	1.27
E-Face	CI	3	3	20	2.30	1	7	1.56
Faces	CI	3	3	20	2.00	1	6	1.49
Faces	Novice IV	4	2	20	4.95	2	7	1.43
E-Face	CI	4	2	20	4.00	1	6	1.81
Faces	Novice IV	4	3	20	3.85	1	7	1.69
E-Face	SAPS	4	1	20	3.55	1	7	1.57
E-Face	SAPS	4	3	20	3.20	1	6	1.51
Faces	Novice IV	4	1	20	3.15	1	6	1.76
E-Face	Novice IV	4	3	20	2.95	1	6	1.43
E-Face	SAPS	4	2	20	2.80	1	7	1.54
E-Face	CI	4	1	20	2.80	1	7	1.77
E-Face	Novice IV	4	1	20	2.70	1	6	1.49
E-Face	Novice IV	4	2	20	2.65	1	5	1.42
Faces	SAPS	4	3	20	2.35	1	7	1.53
Faces	CI	4	3	20	2.25	1	6	1.74
Faces	CI	4	1	20	2.15	1	6	1.39
E-Face	CI	4	3	20	1.90	1	7	1.55
Faces	SAPS	4	2	20	1.80	1	7	1.67
Faces	CI	4	2	20	1.75	1	7	1.48
Faces	SAPS	4	1	20	1.65	1	6	1.31

Appendix H: Sorting hit rates

Reconstructor	Condition	Target	Valid N	Hits	Hit Rate %
			1440	606	42.08
Faces			720	323	44.86
E-Face			720	283	39.31
	SAPS		480	211	43.96
	CI		480	150	31.25
	Novice IV		480	245	51.04
		1	360	121	33.61
		2	360	166	46.11
		3	360	194	53.89
		4	360	125	34.72
Faces	SAPS		240	84	35.00
Faces	CI		240	84	35.00
Faces	Novice IV		240	155	64.58
E-Face	SAPS		240	127	52.92
E-Face	CI		240	66	27.50
E-Face	Novice IV		240	90	37.50
Faces		1	180	60	33.33
Faces		2	180	76	42.22
Faces		3	180	129	71.67
Faces		4	180	58	32.22
E-Face		1	180	61	33.89
E-Face		2	180	90	50.00
E-Face		3	180	65	36.11
E-Face		4	180	67	37.22
	SAPS	1	120	37	30.83
	SAPS	2	120	42	35.00
	SAPS	3	120	84	70.00
	SAPS	4	120	48	40.00
	CI	1	120	31	25.83
	CI	2	120	65	54.17
	CI	3	120	35	29.17
	CI	4	120	19	15.83
	Novice IV	1	120	53	44.17
	Novice IV	2	120	59	49.17
	Novice IV	3	120	75	62.50
	Novice IV	4	120	58	48.33
Faces	SAPS	1	60	13	21.67
Faces	SAPS	2	60	10	16.67
Faces	SAPS	3	60	53	88.33
Faces	SAPS	4	60	8	13.33
Faces	CI	1	60	19	31.67
Faces	CI	2	60	36	60.00
Faces	CI	3	60	24	40.00
Faces	CI	4	60	5	8.33
Faces	Novice IV	1	60	28	46.67
Faces	Novice IV	2	60	30	50.00
Faces	Novice IV	3	60	52	86.67
Faces	Novice IV	4	60	45	75.00

E-Face	SAPS	1	60	24	40.00
E-Face	SAPS	2	60	32	53.33
E-Face	SAPS	3	60	31	51.67
E-Face	SAPS	4	60	40	66.67
E-Face	CI	1	60	12	20.00
E-Face	CI	2	60	29	48.33
E-Face	CI	3	60	11	18.33
E-Face	CI	4	60	14	23.33
E-Face	Novice IV	1	60	25	41.67
E-Face	Novice IV	2	60	29	48.33
E-Face	Novice IV	3	60	23	38.33
E-Face	Novice IV	4	60	13	21.67

Appendix I: Individual composite sorting hit rates (in descending order for each target)

Reconstructor	Condition	Target	C. Number	Valid N	Sum	Hit Rate %
E-Face	Novice IV	1	1	20	11	55
E-Face	SAPS	1	2	20	10	50
Faces	Novice IV	1	3	20	10	50
Faces	CI	1	3	20	10	50
Faces	Novice IV	1	1	20	9	45
Faces	Novice IV	1	2	20	9	45
E-Face	SAPS	1	1	20	8	40
E-Face	Novice IV	1	2	20	7	35
E-Face	Novice IV	1	3	20	7	35
Faces	SAPS	1	2	20	7	35
Faces	CI	1	1	20	7	35
E-Face	SAPS	1	3	20	6	30
E-Face	CI	1	3	20	6	30
Faces	SAPS	1	3	20	5	25
E-Face	CI	1	1	20	4	20
E-Face	CI	1	2	20	2	10
Faces	CI	1	2	20	2	10
Faces	SAPS	1	1	20	1	5
E-Face	SAPS	2	3	20	20	100
Faces	Novice IV	2	3	20	17	85
E-Face	CI	2	1	20	14	70
Faces	CI	2	1	20	14	70
Faces	CI	2	3	20	13	65
E-Face	Novice IV	2	2	20	11	55
E-Face	SAPS	2	1	20	11	55
E-Face	Novice IV	2	1	20	9	45
E-Face	Novice IV	2	3	20	9	45
E-Face	CI	2	2	20	9	45
Faces	CI	2	2	20	9	45
Faces	SAPS	2	2	20	8	40
Faces	Novice IV	2	2	20	7	35
E-Face	CI	2	3	20	6	30
Faces	Novice IV	2	1	20	6	30

E-Face	SAPS	2	2	20	1	5
Faces	SAPS	2	1	20	1	5
Faces	SAPS	2	3	20	1	5
Faces	SAPS	3	1	20	20	100
Faces	Novice IV	3	1	20	19	95
Faces	Novice IV	3	2	20	18	90
Faces	SAPS	3	2	20	17	85
Faces	SAPS	3	3	20	16	80
E-Face	SAPS	3	2	20	15	75
Faces	Novice IV	3	3	20	15	75
E-Face	Novice IV	3	2	20	13	65
E-Face	SAPS	3	3	20	13	65
Faces	CI	3	1	20	13	65
Faces	CI	3	2	20	10	50
E-Face	Novice IV	3	1	20	7	35
E-Face	CI	3	1	20	4	20
E-Face	CI	3	3	20	4	20
E-Face	Novice IV	3	3	20	3	15
E-Face	SAPS	3	1	20	3	15
E-Face	CI	3	2	20	3	15
Faces	CI	3	3	20	1	5
Faces	Novice IV	4	2	20	17	85
E-Face	SAPS	4	1	20	16	80
E-Face	SAPS	4	3	20	16	80
Faces	Novice IV	4	1	20	14	70
Faces	Novice IV	4	3	20	14	70
E-Face	SAPS	4	2	20	8	40
E-Face	CI	4	2	20	7	35
E-Face	Novice IV	4	2	20	6	30
Faces	SAPS	4	1	20	6	30
E-Face	Novice IV	4	3	20	5	25
E-Face	CI	4	1	20	4	20
E-Face	CI	4	3	20	3	15
Faces	CI	4	3	20	3	15
E-Face	Novice IV	4	1	20	2	10
Faces	SAPS	4	2	20	2	10
Faces	CI	4	2	20	2	10
Faces	SAPS	4	3	20	0	0
Faces	CI	4	1	20	0	0

Appendix J: Ranking means (out of 20)

Reconstructor	Condition	Target	Valid N	Mean	Minimum	Maximum	Std.Dev.
			1440	11.03	1	20	5.60
Faces			720	11.88	1	20	5.74
E-Face			720	10.19	1	20	5.33
	SAPS		480	11.11	1	20	5.72
	CI		480	12.82	1	20	5.10
	Novice IV		480	9.16	1	20	5.36
		1	360	10.88	1	20	5.75
		2	360	10.92	1	20	5.65
		3	360	10.97	1	20	5.61
		4	360	11.36	1	20	5.39
Faces	SAPS		240	13.67	1	20	5.40
Faces	CI		240	13.33	1	20	4.82
Faces	Novice IV		240	8.63	1	20	5.54
E-Face	SAPS		240	8.56	1	20	4.83
E-Face	CI		240	12.31	1	20	5.33
E-Face	Novice IV		240	9.70	1	20	5.13
Faces		1	180	12.94	1	20	5.73
Faces		2	180	11.53	1	20	5.32
Faces		3	180	10.42	1	20	5.73
Faces		4	180	12.62	1	20	5.85
E-Face		1	180	8.82	1	20	5.00
E-Face		2	180	10.32	1	20	5.91
E-Face		3	180	11.52	1	20	5.45
E-Face		4	180	10.11	2	20	4.56
	SAPS	1	120	11.97	1	20	6.28
	SAPS	2	120	11.24	1	20	5.69
	SAPS	3	120	9.81	1	20	5.02
	SAPS	4	120	11.44	2	20	5.66
	CI	1	120	10.88	1	19	5.35
	CI	2	120	12.07	1	20	5.43
	CI	3	120	14.28	2	20	4.92
	CI	4	120	14.07	2	20	3.79
	Novice IV	1	120	9.79	1	20	5.42
	Novice IV	2	120	9.46	1	20	5.54
	Novice IV	3	120	8.83	1	20	5.36
	Novice IV	4	120	8.58	1	20	5.09
Faces	SAPS	1	60	16.02	3	20	4.90
Faces	SAPS	2	60	13.35	1	20	4.94
Faces	SAPS	3	60	9.32	1	20	4.98
Faces	SAPS	4	60	15.98	5	20	3.80
Faces	CI	1	60	13.00	1	19	5.18
Faces	CI	2	60	11.13	1	20	5.23
Faces	CI	3	60	14.20	3	20	4.83
Faces	CI	4	60	15.00	7	20	2.85
Faces	Novice IV	1	60	9.80	1	18	5.40
Faces	Novice IV	2	60	10.10	1	19	5.34
Faces	Novice IV	3	60	7.73	1	20	5.34
Faces	Novice IV	4	60	6.87	1	20	5.52

E-Face	SAPS	1	60	7.92	1	20	4.72
E-Face	SAPS	2	60	9.13	1	20	5.65
E-Face	SAPS	3	60	10.30	1	19	5.05
E-Face	SAPS	4	60	6.90	2	16	2.86
E-Face	CI	1	60	8.77	1	19	4.67
E-Face	CI	2	60	13.00	2	20	5.51
E-Face	CI	3	60	14.35	2	20	5.05
E-Face	CI	4	60	13.13	2	20	4.37
E-Face	Novice IV	1	60	9.78	1	20	5.47
E-Face	Novice IV	2	60	8.82	1	20	5.71
E-Face	Novice IV	3	60	9.92	1	20	5.19
E-Face	Novice IV	4	60	10.30	2	20	3.99

Appendix K: Individual composite ranking means (out of 20, in descending rank order for each target)

Reconstructor	Condition	Target	C. Number	Valid N	Mean	Minimum	Maximum	Std.Dev.
E-Face	SAPS	1	2	20	6.35	1	14	3.75
E-Face	SAPS	1	1	20	7.70	1	20	5.52
E-Face	CI	1	1	20	8.30	1	16	4.67
Faces	Novice IV	1	1	20	8.30	1	17	5.50
Faces	Novice IV	1	3	20	8.30	2	18	5.44
E-Face	CI	1	3	20	8.45	1	19	4.98
E-Face	Novice IV	1	1	20	9.50	2	17	5.06
E-Face	Novice IV	1	3	20	9.55	1	18	5.75
E-Face	CI	1	2	20	9.55	1	18	4.48
E-Face	SAPS	1	3	20	9.70	3	18	4.33
E-Face	Novice IV	1	2	20	10.30	1	20	5.82
Faces	CI	1	1	20	11.60	5	18	4.87
Faces	CI	1	3	20	11.80	1	18	5.23
Faces	Novice IV	1	2	20	12.80	3	18	4.07
Faces	SAPS	1	2	20	13.25	3	20	5.15
Faces	SAPS	1	3	20	15.05	5	20	4.90
Faces	CI	1	2	20	15.60	4	19	4.64
Faces	SAPS	1	1	20	19.75	18	20	0.55
E-Face	SAPS	2	3	20	6.65	1	18	5.34
E-Face	Novice IV	2	1	20	7.25	1	20	5.47
E-Face	SAPS	2	1	20	7.95	1	16	4.32
E-Face	CI	2	1	20	8.10	2	19	4.56
E-Face	Novice IV	2	3	20	8.70	3	17	4.58
Faces	Novice IV	2	3	20	8.80	1	19	6.75
Faces	Novice IV	2	2	20	9.80	1	17	4.89
Faces	CI	2	2	20	9.85	1	19	5.37
E-Face	Novice IV	2	2	20	10.50	1	20	6.68
Faces	SAPS	2	2	20	11.60	2	19	4.73
Faces	CI	2	3	20	11.65	3	19	4.67
Faces	Novice IV	2	1	20	11.70	3	18	3.80
Faces	CI	2	1	20	11.90	3	20	5.62
E-Face	SAPS	2	2	20	12.80	3	20	5.45
Faces	SAPS	2	3	20	13.10	1	20	5.62

E-Face	CI	2	3	20	15.05	5	20	4.52
Faces	SAPS	2	1	20	15.35	7	20	3.80
E-Face	CI	2	2	20	15.85	9	20	3.86
Faces	Novice IV	3	2	20	6.85	1	18	5.22
Faces	Novice IV	3	3	20	6.90	1	17	5.13
E-Face	Novice IV	3	2	20	7.80	1	20	5.04
Faces	SAPS	3	1	20	8.70	1	16	5.10
E-Face	Novice IV	3	1	20	9.25	2	16	4.85
E-Face	SAPS	3	2	20	9.35	1	19	5.55
Faces	Novice IV	3	1	20	9.45	1	20	5.49
Faces	SAPS	3	2	20	9.60	2	18	5.01
Faces	SAPS	3	3	20	9.65	3	20	5.02
E-Face	SAPS	3	3	20	9.70	3	16	4.29
E-Face	CI	3	1	20	10.30	2	18	4.26
E-Face	SAPS	3	1	20	11.85	4	19	5.11
Faces	CI	3	2	20	12.65	3	20	5.02
E-Face	Novice IV	3	3	20	12.70	5	19	4.62
Faces	CI	3	1	20	12.95	4	19	5.13
E-Face	CI	3	3	20	14.55	5	19	4.51
Faces	CI	3	3	20	17.00	10	20	2.90
E-Face	CI	3	2	20	18.20	11	20	2.80
Faces	Novice IV	4	2	20	3.55	1	9	2.28
E-Face	SAPS	4	1	20	5.50	2	10	1.76
E-Face	SAPS	4	3	20	6.30	2	10	2.60
Faces	Novice IV	4	1	20	7.20	1	20	6.12
E-Face	Novice IV	4	3	20	8.30	3	16	2.98
E-Face	SAPS	4	2	20	8.90	3	16	2.99
E-Face	Novice IV	4	2	20	9.25	2	20	3.91
Faces	Novice IV	4	3	20	9.85	3	20	5.55
E-Face	CI	4	2	20	11.25	2	20	4.53
E-Face	CI	4	1	20	12.85	7	20	3.18
E-Face	Novice IV	4	1	20	13.35	9	19	3.18
Faces	CI	4	3	20	13.55	7	18	3.02
Faces	SAPS	4	3	20	14.05	5	19	3.71
E-Face	CI	4	3	20	15.30	7	20	4.46
Faces	CI	4	1	20	15.35	11	19	2.18
Faces	CI	4	2	20	16.10	9	20	2.79
Faces	SAPS	4	2	20	16.55	5	20	3.61
Faces	SAPS	4	1	20	17.35	8	20	3.44

Appendix L: Rating means (out of 7)

Reconstructor	Condition	Target	Valid N	Mean	Minimum	Maximum	Std.Dev.
			320	3.75	1.00	7.00	1.61
Faces			160	4.26	1.00	7.00	1.61
E-Face			160	3.25	1.00	7.00	1.46
	Expert IV		160	4.14	1.00	7.00	1.79
	Novice IV		160	3.37	1.00	6.67	1.32
		1	80	3.85	1.00	7.00	1.62
		2	80	3.22	1.00	6.67	1.27
		3	80	3.88	1.00	7.00	1.69
		4	80	4.08	1.00	7.00	1.73
Faces	Expert IV		80	4.75	1.00	7.00	1.73
Faces	Novice IV		80	3.78	1.33	6.67	1.31
E-Face	Expert IV		80	3.53	1.00	7.00	1.63
E-Face	Novice IV		80	2.97	1.00	6.33	1.21
Faces		1	40	4.48	1.67	7.00	1.54
Faces		2	40	3.40	1.00	6.67	1.34
Faces		3	40	4.48	1.33	7.00	1.59
Faces		4	40	4.69	1.67	7.00	1.66
E-Face		1	40	3.22	1.00	6.33	1.45
E-Face		2	40	3.04	1.00	5.00	1.20
E-Face		3	40	3.27	1.00	7.00	1.57
E-Face		4	40	3.46	1.00	7.00	1.59
	Expert IV	1	40	4.20	1.00	7.00	1.81
	Expert IV	2	40	3.33	1.00	6.00	1.38
	Expert IV	3	40	4.25	1.00	7.00	1.85
	Expert IV	4	40	4.78	1.00	7.00	1.80
	Novice IV	1	40	3.49	1.33	6.33	1.32
	Novice IV	2	40	3.12	1.33	6.67	1.16
	Novice IV	3	40	3.50	1.33	6.33	1.43
	Novice IV	4	40	3.38	1.00	6.00	1.35
Faces	Expert IV	1	20	5.00	2.00	7.00	1.72
Faces	Expert IV	2	20	3.65	1.00	6.00	1.42
Faces	Expert IV	3	20	4.95	2.00	7.00	1.67
Faces	Expert IV	4	20	5.40	2.00	7.00	1.70
Faces	Novice IV	1	20	3.95	1.67	5.33	1.17
Faces	Novice IV	2	20	3.15	1.33	6.67	1.23
Faces	Novice IV	3	20	4.02	1.33	6.33	1.40
Faces	Novice IV	4	20	3.98	1.67	6.00	1.32
E-Face	Expert IV	1	20	3.40	1.00	6.00	1.57
E-Face	Expert IV	2	20	3.00	1.00	5.00	1.30
E-Face	Expert IV	3	20	3.55	1.00	7.00	1.79
E-Face	Expert IV	4	20	4.15	1.00	7.00	1.73
E-Face	Novice IV	1	20	3.03	1.33	6.33	1.34
E-Face	Novice IV	2	20	3.08	1.33	5.00	1.12
E-Face	Novice IV	3	20	2.98	1.33	5.67	1.31
E-Face	Novice IV	4	20	2.77	1.00	4.67	1.11

Appendix M: Individual composite rating means (out of 7, in descending order for each target)

Reconstructor	Condition	Target	Valid N	Mean	Minimum	Maximum	Std.Dev.
Faces	Expert IV	1	20	5.00	2.00	7.00	1.72
Faces	Novice IV	1	20	3.95	1.67	5.33	1.17
E-Face	Expert IV	1	20	3.40	1.00	6.00	1.57
E-Face	Novice IV	1	20	3.03	1.33	6.33	1.34
Faces	Expert IV	2	20	3.65	1.00	6.00	1.42
Faces	Novice IV	2	20	3.15	1.33	6.67	1.23
E-Face	Novice IV	2	20	3.08	1.33	5.00	1.12
E-Face	Expert IV	2	20	3.00	1.00	5.00	1.30
Faces	Expert IV	3	20	4.95	2.00	7.00	1.67
Faces	Novice IV	3	20	4.02	1.33	6.33	1.40
E-Face	Expert IV	3	20	3.55	1.00	7.00	1.79
E-Face	Novice IV	3	20	2.98	1.33	5.67	1.31
Faces	Expert IV	4	20	5.40	2.00	7.00	1.70
E-Face	Expert IV	4	20	4.15	1.00	7.00	1.73
Faces	Novice IV	4	20	3.98	1.67	6.00	1.32
E-Face	Novice IV	4	20	2.77	1.00	4.67	1.11

Appendix N: Sorting Hit Rates

Reconstructor	Condition	Target	Valid N	Hits	Hit Rate %
			320	195.67	61.15
Faces			160	109.67	68.54
E-Face			160	86.00	53.75
	Expert IV		160	114.00	71.25
	Novice IV		160	81.67	51.04
		1	80	33.67	42.08
		2	80	50.67	63.33
		3	80	59.00	73.75
		4	80	52.33	65.42
Faces	Expert IV		80	58.00	72.50
Faces	Novice IV		80	51.67	64.58
E-Face	Expert IV		80	56.00	70.00
E-Face	Novice IV		80	30.00	37.50
Faces		1	40	20.33	50.83
Faces		2	40	25.00	62.50
Faces		3	40	35.33	88.33
Faces		4	40	29.00	72.50
E-Face		1	40	13.33	33.33
E-Face		2	40	25.67	64.17
E-Face		3	40	23.67	59.17
E-Face		4	40	23.33	58.33
	Expert IV	1	40	16.00	40.00
	Expert IV	2	40	31.00	77.50
	Expert IV	3	40	34.00	85.00
	Expert IV	4	40	33.00	82.50
	Novice IV	1	40	17.67	44.17
	Novice IV	2	40	19.67	49.17

	Novice IV	3	40	25.00	62.50
	Novice IV	4	40	19.33	48.33
Faces	Expert IV	1	20	11.00	55.00
Faces	Expert IV	2	20	15.00	75.00
Faces	Expert IV	3	20	18.00	90.00
Faces	Expert IV	4	20	14.00	70.00
Faces	Novice IV	1	20	9.33	46.67
Faces	Novice IV	2	20	10.00	50.00
Faces	Novice IV	3	20	17.33	86.67
Faces	Novice IV	4	20	15.00	75.00
E-Face	Expert IV	1	20	5.00	25.00
E-Face	Expert IV	2	20	16.00	80.00
E-Face	Expert IV	3	20	16.00	80.00
E-Face	Expert IV	4	20	19.00	95.00
E-Face	Novice IV	1	20	8.33	41.67
E-Face	Novice IV	2	20	9.67	48.33
E-Face	Novice IV	3	20	7.67	38.33
E-Face	Novice IV	4	20	4.33	21.67

Appendix O: Individual composite sorting hit rates (in descending order for each target)

Reconstructor	Condition	Target	Valid N	Hits	Hit Rate %
Faces	Expert IV	1	20	11.00	55.00
Faces	Novice IV	1	20	9.33	46.67
E-Face	Novice IV	1	20	8.33	41.67
E-Face	Expert IV	1	20	5.00	25.00
E-Face	Expert IV	2	20	16.00	80.00
Faces	Expert IV	2	20	15.00	75.00
Faces	Novice IV	2	20	10.00	50.00
E-Face	Novice IV	2	20	9.67	48.33
Faces	Expert IV	3	20	18.00	90.00
Faces	Novice IV	3	20	17.33	86.67
E-Face	Expert IV	3	20	16.00	80.00
E-Face	Novice IV	3	20	7.67	38.33
E-Face	Expert IV	4	20	19.00	95.00
Faces	Novice IV	4	20	15.00	75.00
Faces	Expert IV	4	20	14.00	70.00
E-Face	Novice IV	4	20	4.33	21.67

Appendix P: Ranking means (out of 20)

Reconstructor	Condition	Target	Valid N	Mean	Minimum	Maximum	Std.Dev.
			320	7.43	1.00	20.00	4.67
Faces			160	6.68	1.00	20.00	4.85
E-Face			160	8.18	1.00	18.00	4.36
	Expert IV		160	5.69	1.00	20.00	5.01
	Novice IV		160	9.16	2.33	16.67	3.55
		1	80	8.43	1.00	18.00	4.64
		2	80	8.07	1.00	20.00	4.64
		3	80	7.54	1.00	18.00	4.85
		4	80	5.68	1.00	15.67	4.10
Faces	Expert IV		80	4.73	1.00	20.00	5.06
Faces	Novice IV		80	8.63	2.33	15.67	3.76
E-Face	Expert IV		80	6.66	1.00	18.00	4.79
E-Face	Novice IV		80	9.70	3.67	16.67	3.26
Faces		1	40	8.15	1.00	18.00	5.15
Faces		2	40	8.35	1.00	20.00	5.12
Faces		3	40	5.37	1.00	17.00	4.35
Faces		4	40	4.83	1.00	15.67	3.75
E-Face		1	40	7.08	1.00	18.00	5.23
E-Face		2	40	6.68	1.00	20.00	5.38
E-Face		3	40	6.25	1.00	18.00	5.44
E-Face		4	40	2.78	1.00	8.00	2.09
	Expert IV	1	40	7.08	1.00	18.00	5.23
	Expert IV	2	40	6.68	1.00	20.00	5.38
	Expert IV	3	40	6.25	1.00	18.00	5.44
	Expert IV	4	40	2.78	1.00	8.00	2.09
	Novice IV	1	40	9.79	3.00	16.67	3.54
	Novice IV	2	40	9.46	3.67	15.67	3.28
	Novice IV	3	40	8.83	3.33	15.67	3.84
	Novice IV	4	40	8.58	2.33	15.67	3.52
Faces	Expert IV	1	20	6.50	1.00	18.00	6.02
Faces	Expert IV	2	20	6.60	1.00	20.00	6.17
Faces	Expert IV	3	20	3.00	1.00	17.00	3.66
Faces	Expert IV	4	20	2.80	1.00	8.00	2.17
Faces	Novice IV	1	20	9.80	3.00	14.00	3.52
Faces	Novice IV	2	20	10.10	4.33	15.67	3.06
Faces	Novice IV	3	20	7.73	3.33	15.67	3.70
Faces	Novice IV	4	20	6.87	2.33	15.67	3.94
E-Face	Expert IV	1	20	7.65	1.00	18.00	4.37
E-Face	Expert IV	2	20	6.75	1.00	15.00	4.61
E-Face	Expert IV	3	20	9.50	1.00	18.00	5.02
E-Face	Expert IV	4	20	2.75	1.00	7.00	2.07
E-Face	Novice IV	1	20	9.78	4.33	16.67	3.66
E-Face	Novice IV	2	20	8.82	3.67	14.67	3.44
E-Face	Novice IV	3	20	9.92	3.67	14.33	3.74
E-Face	Novice IV	4	20	10.30	7.33	13.33	1.91

Appendix Q: Individual composite ranking means (out of 20, in descending rank order for each target)

Reconstructor	Condition	Target	Valid N	Mean	Minimum	Maximum	Std.Dev.
Faces	Expert IV	1	20	6.50	1.00	18.00	6.02
E-Face	Expert IV	1	20	7.65	1.00	18.00	4.37
E-Face	Novice IV	1	20	9.78	4.33	16.67	3.66
Faces	Novice IV	1	20	9.80	3.00	14.00	3.52
Faces	Expert IV	2	20	6.60	1.00	20.00	6.17
E-Face	Expert IV	2	20	6.75	1.00	15.00	4.61
E-Face	Novice IV	2	20	8.82	3.67	14.67	3.44
Faces	Novice IV	2	20	10.10	4.33	15.67	3.06
Faces	Expert IV	3	20	3.00	1.00	17.00	3.66
Faces	Novice IV	3	20	7.73	3.33	15.67	3.70
E-Face	Expert IV	3	20	9.50	1.00	18.00	5.02
E-Face	Novice IV	3	20	9.92	3.67	14.33	3.74
E-Face	Expert IV	4	20	2.75	1.00	7.00	2.07
Faces	Expert IV	4	20	2.80	1.00	8.00	2.17
Faces	Novice IV	4	20	6.87	2.33	15.67	3.94
E-Face	Novice IV	4	20	10.30	7.33	13.33	1.91