Testing facial composite construction under witness stress

Peter J.B. Hancock
Psychology, School of Natural Sciences
University of Stirling
Stirling, FK9 4LA, UK
pjbh1@stir.ac.uk

Karen Burke
Psychology, School of Natural Sciences
University of Stirling
Stirling, FK9 4LA, UK
E-mail

Charlie D Frowd
School of Psychology
Darwin Building
University of Central Lancashire
Preston PR1 2HE
cfrowd@uclan.ac.uk

Abstract

Facial composite systems may be used by police to help a witness to a crime create a likeness of the perpetrator. Evaluation of new facial composite systems in the laboratory allows a measure of experimental control, but lacks the emotional impact of a real crime. As a step towards a more realistic level of stress for our participant witnesses, we presented target face images while they were engaged in playing an action thriller computer game. The quality of the composites they subsequently produced was compared with that of a second 'onlooker' participant, who merely observed the game and had the same view of the target face. Heart rate monitoring confirmed that the players were more stressed than the onlookers while the recognition rate of the onlooker composites was twice as good. We conclude that the method holds some promise as a method for composite system evaluation.

1. Introduction

A witness or victim of a crime may be asked by police to create a likeness of the face of the perpetrator, commonly known as a facial composite because early systems worked by piecing together different features to make up the face. In fact the image may be produced by a trained artist, or by one of the more modern computerized systems that do not work in this feature-driven way. Over the last decade, Frowd, Hancock and Bruce have developed one such 'fourth generation' composite system, called EvoFIT [1-3]. As part of this process, we have established a 'gold-standard' evaluation procedure [4], which involves showing witness-participants a photograph or video of an unfamiliar person and then, 1-2 days later, working with them to create a composite. These composites are later shown to other participants who do know the people depicted, to attempt recognition. This process therefore mimics realistic usage in that neither the witness, nor the composite system operator, know the target, there is

a delay between seeing the target and making the composite, and the ultimate evaluation is whether the composite is recognized. However, the methodology completely fails to mimic the reality of a crime, quite possibly violent and personal. It seems possible that following a real incident, a witness would react very differently to the interviewing process and that therefore, what works in the laboratory might not in real life.

This problem is by no means unique to the evaluation of composite systems. Exactly the same issues apply to eye-witness testimony as a whole, both in terms of reporting accurately what happened, and in attempting to identify the perpetrator, for example in a lineup. Thus, for instance, the aim might be to compare sequential and simultaneous presentation of lineup items. In a laboratory, participants can be shown a video of some incident and later tested on the lineup in one of the two conditions. So, what effect might a real incident have, compared with being shown a video? Intuitively, it would seem that a real incident might lead to better memory: the witness will be more aroused and motivated than a participant sitting in a laboratory. However, if the witness is too scared, perhaps they will remember less. Deffenbacher et al. [5] report the results of a meta-analysis of previous studies that have investigated the role of stress, and concluded that overall, there is evidence for those under stress performing worse. For example, Ihlebæk et al. [6] compared recall for an event seen either live or on video, and found that recall was worse for those who had seen it live, both in terms of number of details and accuracy.

An interesting issue is the effect known as weapons focus. Someone threatened with a weapon may feel their life to be in serious danger and therefore pay extreme attention to the incident. However, the evidence from many studies is that the presence of a weapon worsens recall: it seems attention is paid to the weapon, rather than the face of the person wielding it [7]. However, there is some evidence for a laboratory study effect: Halford and Milne [8] reported an analysis of real incidents and found that those exposed to weapons or violence performed *better* than those who had not been. So it seems that there may indeed be important differences between the kind of stress possible in a laboratory study and that caused by a real assault. Perhaps those involved in real incidents really are better motivated than people in laboratories.

One problem is that it is not generally ethical to stress people in a realistic manner in a laboratory study. It would not be acceptable to make our participants actually fear for their lives. A possible solution is to find situations where people are already under stress, for another reason. A good example is reported by Morgan et al.[9], who studied soldiers enrolled in a military survival school, designed to emulate being held as a prisoner of war and subjected to interrogation. All were subjected to both high and low stress interrogations, by different instructors. Subsequent tests evaluated their ability to recognize the instructors in a variety of ways, such as live or photographic lineups. Recognition of the instructors from the high-stress interrogation was significantly worse: 30% compared to 62% for the live lineup trials.

A somewhat less extreme methodology is reported by Valentine and Mesout [10], who studied visitors to the London Dungeon. This is intended to be entertainingly scary and raises no ethical issues, since attendance is voluntary. During their time in the ‘Horror Labyrinth’, which is a dark, disorienting maze of mirrors, visitors were confronted by a scary-looking actor. After leaving the dungeon, they were asked to try and identify the actor from a lineup. Results showed that those who reported a high level of anxiety were less likely to make an

accurate identification, and produced fewer correct descriptors. A separate study showed that heart rate, measured by portable monitors, was positively related to reported anxiety level, so the labyrinth was successful in raising stress levels.

Both of these studies, while relatively realistic, involve a considerable amount of fieldwork. While we have contemplated carrying out assessment of composite systems under similar circumstances, it would still be helpful to find something based in the laboratory that would produce stress, ethically. Our proposed solution is to use computer games. Some of these are highly immersive and involve fighting for your virtual life. Our ideal answer would be to import the image of our target into the game. Then a player could interact normally with the game and be asked subsequently to try and build a composite of the character that they had been in a fight with. However, the technology to do so is not quite there yet, though close. For this trial, therefore, we used an alternative approach, which was suddenly to switch the video feed from the game to an image of a target person, which is quite startling for the player.

In order to have a comparison group, we had two participants in each session; one played the game, projected onto a wall-screen, the other simply watched. Both therefore had exactly the same view of both the action and the target image, but the player should be more stressed. To confirm whether this was the case, both participants had their heart rate monitored during the game, with the expectation that they would be correlated, as the game action varied, but that the player should be higher than the observer. A day or two after the game session, each returned independently to make a composite of the target. Our prediction was that the observers' composites would be better recognized than the players'.

2. Method, composite construction

2.1. Participants

Twenty-four undergraduates from University of Stirling, aged between 17 and 52, took part either voluntarily or in return for course credit. Four had to be dropped due to software problems.

2.2 Materials

Digital photographs were obtained of 10 members of psychology staff, half male and all unknown to any of the above participants. The photographs were clear, showing full face with a neutral expression, in colour. The game played was 'Alan Wake', a 'psychological action thriller' running on an Xbox 360. A video mixer allowed instant switching between the Xbox and the computer used to display the photographs. These appeared on screen at about four times life size, giving a viewing angle from 4 metres of a person standing one metre away.

2.3 Procedure

Participants took part in pairs, with one randomly allocated to be the player ('victim') while the other was the observer ('bystander'). They sat either side of a table, giving similar views of the projection screen and were connected to a Biopac heart rate monitor. The player was given a sheet of instructions for the game and started to play. The experimenter used a stopwatch to time presentations of the target image, which were shown five times, for a

period of about 1 second each time, approximately 3 minutes apart. The participants were unaware beforehand that face images would be shown during game play, but knew they would be required to make a composite of a face they saw during the session. Both participants' heart rate was recorded for a period of a minute before and after each image display.

The participants returned, separately, one to two days later to create their composite of the target face. The procedure followed that recommended to UK police. They first took part in a cognitive interview, which uses context reinstatement to help with recall. They were first asked to describe the face in as much detail as possible, before being prompted about each feature in turn. They were then asked to think about the face holistically, rating the face on scales such as intelligence and friendliness [11]. Throughout, they were encouraged to use visual imagery to help with recall. The first stage of image construction was to choose the hairstyle. Once chosen, this was blurred, which helps to focus attention on the internal features of the face [12]. In EvoFit, these are generated by first selecting the best shape for the face, using average features, by selecting from screens of 18 randomly generated face images, then repeating this process for the best features, then picking the best combination of shape and features. This process is repeated, with new images being generated by an underlying evolutionary algorithm. Towards the end of the process, holistic tools are available, that can change the overall appearance of the face, for example to make it look older. The process is described in detail elsewhere [12].

3. Method: composite recognition

3.1 Participants

Forty third and fourth year psychology undergraduates, familiar with the staff targets, took part voluntarily.

3.2 Materials and procedure

The ten composites made by 'victims' and the ten made by 'bystanders' were printed and placed in separate folders, along with copies of an extra 10 composites (half male) that were created by the experimenter from unfamiliar faces. The purpose of these was to reduce the possibility of the participants guessing the identity of a composite through process of elimination. Thus if a composite had short, blond hair, the participant might think through which members of staff might fit, but since they knew some composites were not of staff members, they would be forced actively to recognize the face. Participants were tested individually, with half seeing each booklet. The order of composites was randomized each time. After viewing the composites, participants were shown prints of the original photographs of the 10 staff members to check for recognition, since they could not be expected to identify the composite if they did not know the staff member depicted.

3.3 Results

3.3.1 composite naming: On average, 12.5% of the composites produced by the 'victims' (games players) were recognized, compared with 25.5% of those produced by the 'bystanders'. This difference is significant, t_{38}=2.97, p=0.005, with a large effect size, d=0.94. However, because composites vary greatly in naming rates, it as well also to perform

analysis by items, in case the effect derives only from one or two exceptional images. In this case, there are only 10 pairs of composites and the distribution of naming rates is distinctly non-normal, so a t-test is not very reliable. We therefore used resampling statistics, sampling at random with replacement from the 10 pairs of naming scores. Thus the algorithm generates a new set of ten pairs of scores, which could in principle all be sampled from the same original pair of composite scores. The means of the ten 'victim' and 'bystander' composites are then compared. This whole process was repeated 100,000 times and the mean of the 'bystander' items was higher than the mean of the 'victim' items 96,032 times, giving p=.04, one-tailed.

3.3.2. Heart rate data: Three pairs of participants had to be excluded from the heart rate comparison due to erratic data from the recording system. For the remaining pairs, their heart rates were recorded for a period of a minute before and after each image display, giving up to 5 heart rate measures for each of them. Only 'up to' 5, because for all but one of the pairs, at least one measure was missing for one of the participants. This resulted in between 3 and 5 measures for each of the 7 pairs. The overall averages were 84.4 bpm (sd=8.0) for the 'victims' and 81.6 bpm (sd=5.4) for the 'bystanders'. These data cannot readily be analysed by standard statistics, since we have a variable number of repeated measures for each pair of participants. It is important that the data are paired, since at each image display, the two participants were observing the same passage of game play, which might be relatively exciting or dull. We therefore again used resampling statistics, sampling at random with replacement for each participant pair. Thus the first pair of participants had four pairs of data points. The algorithm picks four new pairs of data from the original four with replacement, as above. Repeating this process for each of the 7 pairs of participants gives a new set of data and the means are compared. The whole process is repeated 100,000 times, and the average for the victim group came out higher than that of the bystander group 98,637 times, thus p=.014, one-tailed.

Our prediction was that the heart rates of the paired 'victim' and 'bystander' should be correlated, if the game action was exciting. Thus, if there was something scary on screen, both would see it and respond, though presumably the 'victim' player might respond more. Unfortunately, the patchy data make this prediction hard to test; there are 7 pairs of data, of varying length, from 3 to 5. To get an idea of the correlation, we used the same resampling procedure, but then calculated an average correlation for the 7 pairs (or rather, a root mean square correlation, since correlation scores do not add). This procedure gave an estimate for the correlation of 0.48, which is a medium to large effect size, though not formally significant because the numbers are so low. Overall, then, there is suggestive evidence that the heart rates reflect game action, rather than simply the player moving their arms more through using the controller.

4. Discussion

Our expectation was that the 'victims', who actually played the game, would be under more stress than the 'bystanders' and that as a consequence, their composite would be less well recognized. The heart rate data confirmed our prediction about stress levels, with the victims showing a small but significant increase in average heart rate at the times when the target images were shown. While the difference between the two groups could simply reflect the physical effort of game playing, the evidence for a correlation between the heart rates of victim-bystander pairs suggests that the game action was having some effect. The composite

naming data also confirmed our expectation: the 'bystander' composites were named approximately twice as often as those from the 'victims', with the more powerful subjects analysis confirming a very large effect size.

The finding of an effect of stress is consistent with other memory studies [9,10] and with theoretical models of the processes involved, such as Deffenbacher's catastrophe model [5]. Thus, while mild stress may increase attention and memory, too much will result in decreased accuracy. However, one alternative explanation for the difference in composite quality is that the 'victims', who had to play the game, were not stressed in the sense of being frightened, but simply had a higher cognitive load than the bystanders. Although both saw the image briefly, the bystanders would be more able to detach from the game and think about the face, where the players would be more involved in the game play. The heart rate difference indicates that there was a stress difference but anecdotally, the participants did not often report feeling scared. Future work could involve a more demanding computer game, with participants already experienced in how to play it so that they are able to concentrate more on the game and, hopefully, get more emotionally involved. On the other hand, experienced game players may not be stressed by the action, either.

Another possible explanation is that the 'victims', who had to play the game, happened to be looking either at the game controller or the instruction sheet at the time the target image flashed up. While it seems unlikely that this would be the case on all five presentations, it might have contributed to a lower average exposure to the target for the 'victim' group. On the other hand, it is also possible that the 'victims' playing the game, paid greater attention to the action on the screen, where the 'bystander' was free to let their attention wander. Future work could check this either by giving each participant head-mounted eye-tracking, or more economically by videoing them and checking where they were looking during the image display periods; their reaction to the surprise image should also indicate whether they saw it. More experienced game players should also address this issue.

The target image was only displayed for a total of five seconds, compared with typical viewing times of a minute in previous work. Despite this, and the presence of foil composites in the set presented for recognition, which reduce the chances of identification by elimination, the correct identification rate for the bystander composites was 26%. This is very similar to the 25% reported by Frowd et al. [12] when using EvoFIT in the same way, with hair blurred and holistic tools used during construction. However, in that study, participant witnesses were given the photograph to inspect for a minute and knew beforehand that they should attempt to remember it for later composite construction. Here, our participants knew that they would be required to make a composite of a face seen during the first part of the study, but knew nothing about when it would be shown. The important implication is that it is possible to make highly identifiable composites a couple of days after an unexpected and fleeting view of a face.

Finally, the main aim of this work is to address a concern that lab-based studies may not translate well to the real world. The acid test of whether a composite system works will come from use by the police with real victims. In such use, it is hard to obtain an objective measure of composite quality. Suppose someone is convicted (on other evidence, a composite should not be used as evidence). One measure might be ratings of similarity between the convict and the composite, but if conducted by those unfamiliar with the person concerned they would not be very reliable, because of differences between familiar and unfamiliar face processing. The

best measure would be to see whether those who know the convict recognize the composite, but it would be difficult to do this in an unbiased fashion after the trial. A good measure is whether the composite directly leads to an arrest and initial field trials suggest that this happens for between 25-38% of cases [13]. Therefore it would appear that our lab-based methodologies do translate well to real world use but it remains possible that appropriate stress in the lab would allow us to develop still better techniques.

5. References

[1] P. J. B. Hancock, "Evolving faces from principal components.," Behavior Research Methods, Instruments and Computers 32(2), 2000, 327-333
[2] C. D. Frowd, P. J. B. Hancock, and D. Carson, "EvoFIT: A holistic, evolutionary facial imaging technique for creating composites.," ACM Transactions on Applied Perception 1(1), 2000, 19-39.
[3] C. Frowd, V. Bruce, and P. J. B. Hancock, "Changing the face of criminal identification," Psychologist 21(8), 2008, 668-672
[4] C. D. Frowd et al., "A forensically valid comparison of facial composite systems," Psychology Crime & Law 11(1), 2005, 33-52
[5] K. A. Deffenbacher, B. H. Bornstein, S. D. Penrod, and E. K. McGorty, "A meta-analytic review of the effects of high stress on eyewitness memory," Law and Human Behavior 28(6), 2004, 687–706
[6] C. Ihlebæk, T. Løve, D. E. Eilertsen, and S. Magnussen, "Memory for a staged criminal event witnessed live and on video.," Memory 11(3), 2003, 319-327
[7] N. M. Steblay, "A meta-analytic review of the weapon focus effect," Law and Human Behavior 16(4), 1992, 413–424
[8] P. Halford and R. Milne, "The identification performance of forensic eyewitnesses exposed to weapons and violence," presented at the XVth European Conference on Psychology and Law, Vilnius, 2005.
[9] C. A. Morgan et al., "Accuracy of eyewitness memory for persons encountered during exposure to highly intense stress.," International Journal of Law and Psychiatry 27(3), 2007, 265-279
[10] T. Valentine and J. Mesout, "Eyewitness identification under stress in the London Dungeon.," Applied Cognitive Psychology 23(2), 2009, 151-161
[11] C. D. Frowd, V. Bruce, A. J. Smith, and P. J. B. Hancock, "Improving the quality of facial composites using a holistic cognitive interview," Journal of Experimental Psychology-Applied 14(3), 2008, 276-287
[12] C. D. Frowd et al., "The psychology of face construction: Giving evolution a helping hand.," Applied Cognitive Psychology 25(2), 2011, 195-203
[13] C. D. Frowd et al., "Catching More Offenders with EvoFIT Facial Composites: Lab Research and Police Field Trials," Global Journal of Human Social Science 11(3), 2011, 35-46

Authors

Peter Hancock is professor of Psychology at the University of Stirling. His first degree was chemistry and his PhD in computing science. He conceived the EvoFIT facial composite system and has overseen its development.

Karen Burke has recently completed her BSc in Psychology at the University of Stirling. This work derived from her honours dissertation project.

Dr Charlie Frowd is Senior Lecturer in the School of Psychology at the University of Central Lancashire. He is the principal developer of the EvoFIT system and is responsible for the deployment with police.