A Comparison of Individual and Morphed Facial Composites Created Using Different Systems

Conference Paper · September 2010

DOI: 10.1109/EST.2010.29

CITATIONS
6

READS
59

4 authors, including:

Josh P. Davis
University of Greenwich
34 PUBLICATIONS **328** CITATIONS
SEE PROFILE

Stuart Gibson
University of Kent
32 PUBLICATIONS **311** CITATIONS
SEE PROFILE

A Comparison of Individual and Morphed Facial Composites Created Using Different Systems

Josh P Davis, Lucy Sulley
Psychology and Counselling Department,
University of Greenwich
London, UK
e-mail: j.p.davis@gre.ac.uk

Chris Solomon, Stuart Gibson
School of Physical Sciences
University of Kent
Canterbury, Kent, UK
e-mail: s.j.gibson@gre.ac.uk

Abstract

An evaluation of individual and morphed composites created using the E-FIT and EFIT-V production systems was conducted. With the assistance of trained police staff, composites of unfamiliar targets were constructed from memory following a Cognitive Interview. EFIT-V composite production followed either a two-day delay, or on the same day as viewing a video of the target. E-FIT composites were created on the same day as viewing the target video. Morphs were produced from merging either two, or three composites created by the same witness, but with the assistance of a different operator. Participants familiar with the targets supplied similarity-to-target photograph ratings. No differences were found in the rated quality of composites created using E-FIT or EFIT-V, although a two-day delay in production resulted in inferior images. Morphs were rated as better likenesses than individual composites, although the benefits were greater with EFIT-Vs. Encouraging witnesses to create more than one composite image for subsequent morphing might enhance the likelihood of recognition of facial composites of criminals.

1. Introduction

If there is no suspect in a criminal investigation, witnesses from memory, may work with the police to create a facial composite of the offender. In the past, this required a police artist. More recently, mechanical (e.g., Identi-Kit, Photo-Fit) and computerized systems (e.g., E-FIT, FACES) were developed. However, recognition by those familiar with the targets tended to be poor, and under forensically realistic conditions fewer than 3% make a correct identification [1,2]. Davies and Valentine [2] suggest this is due to the requirement for witnesses to select individual facial features to create an image, a difficult task perceptually, and it does not match the holistic processes known to direct human face recognition. We are better at processing whole faces, not their constituent parts.

Therefore, contemporary systems incorporate genetic algorithms to evolve artificial but highly realistic full-face composites, designed to actively tap into these processes (e.g., EFIT-V; [3]), EvoFIT; [4]). For instance, with EFIT-V, witnesses are presented with a series of arrays of nine artificially generated random faces, and they select the one most resembling their memory of the offender. Selections inform the generation of further arrays, so that faces in each appear more similar to one another and ideally the actual target, continuing in an iterative manner until the witness is satisfied with the final image. Additional alterations are possible at any time, so that individual features can be re-positioned or altered in size or shape and images can be 'aged' (for further information see [3,5]).

Some studies have found an advantage for 'holistic' systems over feature-based systems [6,7]. In contrast, the most forensically valid comparison study found an advantage for E-FIT over a number of alternative feature-based systems, as well as EvoFIT, a 'holistic' system [1]. The majority of police forces in the UK currently use E-FIT, and the research reported in this paper was the first to compare composites produced by E-FIT with EFIT-V, a 'holistic' system. This was the first study to compare EFIT-V with any other system, and yet it is already in use by at least 16 UK police forces [8]. A fair evaluation of different systems can be hard to accomplish, as it partly depends on the experience and capabilities of the operator. However, the current research was designed to follow as closely as possible ecologically valid 'gold' standards for composite system comparison studies [1]. As such, police staff were recruited, contributing as a final component of E-FIT and EFIT-V training courses, thus ensuring consistent operator skill levels. As in a real investigation, each administered the Cognitive Interview [9] and then worked with participant-witnesses to create the composites.

In an investigation, knowledge of crime type, locality etc. would be publicized, and providing contextual cues appears to assist composite naming [8], particularly if faces are distinctive [1]. Other 'post-production' manipulations are also advantageous (e.g., presenting

pairs of composites: [10]; caricaturing: [11]), and these include morphing composites made by different witnesses ('between-witness morphs' [8,12,13]), or by the same witness ('within-witness morphs' [8]). A second aim of the experiment reported in this paper was to further investigate this morphing effect.

It has been suggested that the morph advantage is derived from processes important in unfamiliar face processing [8]. Firstly, distinctive faces; and distinctive facial features are better remembered than typical features [14], and are therefore more likely to be better reproduced in composites. Secondly, the external features of unfamiliar faces (e.g., hairstyle, face shape) are better recognised than internal features (e.g., eyes, nose, mouth; [15]) and are also more accurately replicated in composites [16]. As such, composites of the same offender tend to be correct in the representation of external and distinctive features. Morphed images may therefore enhance correct common distinctive features across multiple images, while reducing the influence of infrequently presented incorrect features. Nevertheless, because external features tend to be better reproduced in composites, morphing enhances internal features to a greater extent [8]. Unlike with unfamiliar faces, familiar faces are primarily processed using internal features [15]. Thus, improvements from morphing improve the likelihood of recognition by someone familiar to the target.

The rational for the current experiment was to compare composites produced by E-FIT and EFIT-V, and to investigate whether morphing would benefit both systems. Witnesses unfamiliar with the celebrity targets created multiple consecutive composites of that target. To ensure no operator bias, a different operator worked with the witness on each procedure. Four-morphs of EFIT-Vs (merging four individual composites produced by the same witness) are named more often than individual composites [8]. However, asking witnesses to produce that number of images may be impractical in operational settings, and therefore the current research examined whether morphs made up of fewer composites had a similar effect (i.e. two- and three-morphs).

Due to the large number of final images, a naming study, the most forensically valid assessment method, would have been impractical, as for sufficient statistical power many hundreds of participants would have been required. Instead, participants familiar with the celebrities rated the individual composites for their similarity to photos. Previous research has found that these similarity-to-target likeness ratings result in a similar outcome pattern to the method of naming composites [8].

Some of the EFIT-V composites used in the present study were created using a forensically valid two-day delay after viewing a video of the target. The remainder were created on the same day as the witnesses viewed the target videos. Such a short delay would be unlikely in a real investigation. However, it did allow assessment of the influence of a time interval and it was predicted that because delay has a generalized adverse influence on face recognition [17], composites created with a brief delay would be rated as possessing a better likeness than those created after a two-day delay. However, because of the different pattern of results found in past research when comparing composite systems, no specific predictions were made concerning the ratings of images produced by EFIT and EFIT-V. Finally, based on previous research, morphed images were expected to receive higher likeness ratings than individual composites.

2. Method

2.1. Participants

Twelve police operators, required to create the composites as a final component of separate E-FIT and EFIT-V training courses, worked alongside twelve student participant-witnesses from the University of Kent, unfamiliar with the male targets who were actors from the popular TV soap *Eastenders*. Forty student participant-judges from the University of Greenwich rated the likeness of the composites and morphs to photographs. All were regular viewers of *Eastenders*, and familiar with those depicted.

2.2. Design

Participant-judges viewed a series of PowerPoint slides on a computer monitor, each depicting a facial composite and two photographs of that target actor and provided a similarity rating. A 3 (system/delay: E-FIT no delay, EFIT-V no delay, EFIT-V two-days delay) x 3 (image type: individual composites, two-morphs, three-morphs) repeated-measures design was employed. The dependent variable was the similarity rating (using a scale of 1–10).

2.3. Materials and procedure

Fifty-six individual composites were obtained. These comprised 14 composites of four different actors, all created from memory by participant-witnesses after viewing a 2-min video of one of the actors. In each case, the police operator administered the Cognitive Interview with the witness before working with the composite production system. Using E-FIT, witnesses produced three individual composites of the target, each time working with a different police operator. Using EFIT-V, witnesses produced four individual composites (the procedure is quicker). However, only the first three were used in the current research. This procedure took approximately 5 hours per witness. Once produced,

witnesses simultaneously viewed all their composites and ranked them for similarity to their memory of the target. For half of the witnesses using EFIT-V, there was a delay of two days between video viewing and composite creation. For the remaining witnesses the composites were created on the same day as viewing the target video.

Morph-studio software (The Learning Company, Cambridge, MA) was used to merge composites. Twelve two-morphs were generated by morphing the two highest ranked individual composites produced by the same witness, with each contributing 50% to the morph. Twelve three-morphs were also constructed with each composite contributing 33% to the morph. The final images (screen size: 14cm x 14cm) and two additional photographs of the actors (5.5cm x 5.5cm) were prepared for display.

Participant-judges were sequentially presented with the final images in a randomized and counterbalanced order on a computer monitor with one composite image on the right of the screen and two high quality photographs on the left for comparison. Working at their own pace, they rated each image for similarity to the photographs on a scale of 1-10.

3. Results

A significance level of 0.05 is reported throughout. Table 1 displays the mean similarity-to-target likeness ratings for the combined individual composites created by each participant-witness, together with the mean ratings given to the individual composites ranked first by their creators as well as the mean ratings to the two- and three-morphs.

Table 1: Mean Similarity Rating by Participant-Judges

	EFIT-V	**EFIT-V**	**E-FIT**
Variable	M (SD)	M (SD)	M (SD)
Mean individual composites	2.12 (1.03)	2.60 (1.28)	2.79 (1.41)
Best individual composites	2.28 (1.33)	2.50 (1.36)	2.79 (1.40)
Two-morphs	2.53 (1.26)	3.64 (1.73)	3.29 (2.00)
Three-morphs	2.68 (1.27)	3.18 (1.60)	3.23 (1.81)

The first analysis compared the mean similarity ratings given to the E-FIT and EFIT-V individual composites produced with no delay, or a delay of two-days. A one-way repeated-measures ANOVA was significant, $F(2, 78) = 6.49$, $p < .002$, $\eta^2 = .14$, Bonferonni-corrected paired comparisons found that the EFIT-Vs created after a two-day delay received lower ratings than both the EFIT-Vs and E-FITs created with no delay ($p < .05$ both comparisons). There was no significant difference between the latter conditions ($p > .05$).

A second similar ANOVA compared the ratings given to the individual composites ranked first only by their creator. The EFIT-Vs created with a two-day delay received the lowest ratings. However, this analysis was not significant, $F(2, 78) = 2.78$, $p = .068$, $\eta^2 = .07$.

The third analysis compared the ratings provided to the mean individual composites, two-morphs and three-morphs produced using EFIT-V and E-FIT. A 3 (image type: individual composites, two-morphs, three-morphs) x 3 (session/delay: E-FIT no delay, EFIT-V no delay, EFIT-V two-days delay) repeated-measures ANOVA was conducted on the similarity ratings.

The main effect of image type was significant, $F(2, 78) = 33.26$, $p < .001$, $\eta^2 = .46$, individual composite ratings ($M = 2.50$, $SD = 1.04$) were lower than two-morphs ($M = 3.15$, $SD = 1.28$) and three-morphs ($M = 3.03$, $SD = 1.22$). The main effect of session/delay, Mauchly's $W = .59$, $\chi^2 (2) = 19.77$, $p < .001$, $\varepsilon = .71$ violated the assumptions of sphericity, so that the Greenhouse-Geisser results are reported. This was also significant, $F(1.42, 55.49) = 5.78$, $p = .011$, $\eta^2 = .13$, ratings were lower for the EFIT-Vs made after a two-day delay than the EFIT-Vs and the E-FITs made with no delay.

These effects were mediated by a significant two-way interaction, $F(4, 156) = 3.40$, $p = .011$, $\eta^2 = .08$. Bonferonni-corrected post-hoc analyses found that for both the EFIT-Vs created with no delay, and those created with a two-day delay, the differences were consistent with the main effect of image type. The three- and two-morphs received higher ratings than the individual composites (both comparisons $p < .05$). There were no differences between the two- and three-morphs ($p > .05$). For the E-FITs, the highest ratings were given to the two-morphs, the lowest ratings to the individual composites, and only this difference was significant ($p < .05$).

After the creation of a series of composites by the same witness, it might be considered prudent in an investigation to release the image rated by that witness as best, rather than a morph. Therefore, a second 3 (image type) x 3 (system/delay) repeated-measures ANOVA was conducted on the similarity ratings, but, only the individual composites rated as best by each creator were included.

The main effect of image type was significant, $F(2, 78) = 27.34$, $p < .001$, $\eta^2 = .41$, ratings were lower for individual composites than both two-morphs and three-morphs. The main effect of session/delay, Mauchly's $W = .52$, $\chi^2 (2) = 24.91$, $p < .001$, $\varepsilon = .68$ violated the assumptions of sphericity, so that the Greenhouse-Geisser results are reported. This was also significant, $F(1.35, 52.67) = 4.58$, $p = .027$, $\eta^2 = .11$, ratings were lower for the EFIT-Vs made after a two-day delay than the EFIT-Vs and the E-FITs made with no delay. There was

significant no difference between the latter two conditions ($p > .05$).

These effects were mediated by a significant two-way interaction, $F(4, 156) = 4.27$, $p = .003$, $\eta^2 = .10$. Bonferonni-corrected post-hoc analyses found that with the EFIT-Vs created with no delay, the differences were consistent with the main effect of image type. The three- and two-morphs had significantly higher ratings than the individual composites (both comparisons $p < .05$). For the EFIT-Vs created with a two-day delay, the three-morphs received the highest ratings, the individual composites the lowest and only this difference was significant ($p < .05$). Finally, with E-FITs, even though higher ratings with given to morphs, no differences were significant ($p > .05$). No other differences were significant ($p > .05$).

4. Discussion

In this experiment, participant-judges, familiar with the celebrity targets assessed the similarity to photographs of facial composites produced by E-FIT, a system in which witnesses build the composite from individual facial features, with EFIT-V, a recently introduced genetic algorithm system designed to exploit the holistic processes known to direct face recognition. This was the first study to compare the quality of composites produced by EFIT-V with any other system. As in a real investigation, people unfamiliar with the targets created the composites after the administration of the Cognitive Interview by a trained police operator. The participant-judges were familiar with the targets. No consistent differences in the perceived quality of individual composites produced by the two systems were found when composite creation was on the same day as viewing a target video. However, with a forensically valid two-day delay, similarity-to-target likeness ratings were lower, an expected consequence of delay on memory for faces.

Previous research has found performance improvements from morphing composites created by the same or by different witnesses [8,12,13]. In the current experiment, morphs were merged from individual composites created by the same witness, but with the assistance of different operator each time. Likeness ratings of EFIT-V two-morphs and three-morphs were higher than the original unmanipulated images. However, there were no consistent differences in the ratings of two-morphs and three-morphs. This may a consequence of using equal weightings in morph creation, so that the two individual composites rated by the creator as best each contributed 50% to the two-morphs. The contribution of each of the three composites in three-morphs was 33%. This meant that those images rated to be the worst likeness to their memory of the target by the witness had a similar influence on the final morph as those rated as best. An approach that is perhaps more prudent has been used [8], so that in morph production, more weight was placed on composites rated by their creators as best. Further research could be conducted to investigate optimal weightings in this context.

In contrast to the EFIT-V images, morphing had a less consistent positive impact on E-FIT composites. The reason that morphing benefits one system and not the other may be a consequence of the different production procedures. With EFIT-V, the emphasis is on selecting whole faces from a series of arrays, even though the witness selects a specific hairstyle early on, and can also use a number of tools at any time to manipulate and save individual facial features. In comparison to strongly defined external features, this may result in less well-defined internal features. Indeed, morphing mainly enhances the internal facial configurations of EFIT-V composites [8]. With E-FIT, witnesses, in selecting specific features are required to concentrate on these internal facial aspects. As such, morphing will have less of an impact. However, further research is required to clarify this explanation.

Consistent with recent research, these results suggest that in the context of a real investigation in which time will be at a premium, it would appear expedient to encourage a witness to create at least two composite images for later morphing as it may enhance detection rates. Indeed, regardless of system and weighting, morphs are at least as good a likeness to the target as the best individual composite.

Two notes of concern are relevant. Hasel and Wells [13] suggest that morphs are more average in appearance in comparison to individual composites. These averaging effects mean that as well as morphs being closer in similarity to targets, they become more similar to non-targets. In an investigation, this could result in additional suspects being identified, although in most cases these could be easily ruled out and it is likely that the advantages of increased correct identifications outweigh the disadvantages. In addition, creating a composite using FACES, a feature-based system, reduces the likelihood of a witness identifying the target in a subsequent lineup (identity parade) [18]. It is unclear whether lineup performance would be similarly impaired if a holistic system such as EFIT-V is used in composite production. Furthermore, it is unclear whether creating multiple composites as in the experiment reported here would have an additional adverse influence. However, before recommending that morphing of composites produced by the same witness is introduced as standard practice, additional research is required.

In summary, composites produced using EFIT-V were found to be as least as good in their similarity to targets as those produced using E-FIT, and consistent with previous research, a clear advantage was found for EFIT-V morphs over individual composites. The quality of composites created after a two-day delay was lower. This highlights

the importance of encouraging a speedy response by a composite operator in an investigation, when the memory trace of the offender will be strongest.

5. Declaration

Chris Solomon and Stuart Gibson are faculty members of the University of Kent and directors of Vision Metric Ltd. Vision Metric Ltd. market E-FIT and EFIT-V, commercially available software packages and organise the training sessions for these systems. Solomon and Gibson played no part in the collection, analysis, or interpretation of data. Their contribution as authors is the intellectual contribution in developing the software and providing software support.

This research was supported by a Research and Enterprise Grant from the University of Greenwich (HSE007-09).

6. References

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

[2] G. M. Davies, and T. Valentine, "Facial composites: Forensic utility and psychological research," in *Handbook of eyewitness psychology, vol. 2, memory for people*, R. C. L. Lindsay, D. F. Ross, J. D. Read, and M. P. Toglia, Eds. Mahwah, NJ: Erlbaum, 2007, pp. 59–83.

[3] S. J. Gibson, C. J. Solomon, M. I. S. Maylin, & C. Clark, "New methodology in facial composite construction: From theory to practice," *International Journal of Electronic Security and Digital Forensics*, vol. 2, 2009, pp. 156–168.

[4] C. Frowd, P. J. B. Hancock, and D. Carson, "EvoFIT: A holistic evolutionary facial imaging technique for creating composites," *Association for Computing Machinery Transactions on Applied Psychology*, vol. 1, 2004, pp. 19–39.

[5] S. J. Gibson, A. Pallares Bejarano, and C. Solomon, "Synthesis of photographic quality facial composites using evolutionary algorithms," In *Proceedings of the British Machine Vision Conference*, R. Harvey & J. A. Bangham, Eds. London: British Machine Vision Association, 2003, pp. 221–230.

[6] C. Frowd, V. Bruce, H. Ness, C. Thomson-Bognor, J. Paterson, A. McIntyre, and P. J. B. Hancock, "Parallel approaches to composite production," *Ergonomics, vol. 50, 2007, pp.* 562–585.

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

[8] T. Valentine, J. P. Davis, K. Thorner, C. Solomon, and S. Gibson, "Evolving and Combining Facial Composites: Between-Witness and Within-Witness Morphs Compared," *Journal of Experimental Psychology: Applied*, vol. 16 2010, pp. 72–86.

[9] G. Wells, A. Memon, and S. D. Penrod, "Eyewitness evidence: improving its probative value," *Psychological Sciences in the Public Interest*, vol. 7, 2007, pp. 45-75.

[10] C. Frowd, V. Bruce, Y. Plenderleith, and P. J. B. Hancock, "Improving target identification using pairs of composite faces constructed by the same person," In *IEE Conference on Crime and Security London: Institute of Engineering and Technology*. 2006 (pp. 386–395).

[11] C. Frowd, V. Bruce, D. Ross, A. McIntyre, and P. J. B. Hancock, "An application of caricature: How to improve the recognition of facial composites," *Visual Cognition*, vol. 15, 2007, pp. 954–984.

[12] V. Bruce, H. Ness, P. J. B. Hancock, C. Newman, and J. Rarity, "Four heads are better than one: Combining face composites yields improvements in face likeness," *Journal of Applied Psychology*, vol. 87, 2002, pp. 894–902.

[13] L. E., Hasel, and G. L. Wells, "Catching the bad guy: Morphing composite faces helps," *Law and Human Behavior*, vol. 31, 2007, 193–207.

[14] L. L. Light, F. Kayra-Stuart, and S. Hollander, "Recognition memory for typical and unusual faces," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 5, 1979, 212–228.

[15] A. W. Young, D. C. Hay, K. H. McWeeny, B. M. Flude, and A. W. Ellis, "Matching familiar and unfamiliar faces from internal and external features," *Perception*, vol. 14, 1985, pp. 737–746.

[16] C. Frowd, V. Bruce, A. McIntyre, and P. J. B. Hancock, "The relative importance of external and internal features of facial composites," *British Journal of Psychology*, vol. 98, 2007, pp. 61–77.

[17] P. N. Shapiro, and S. D. Penrod, "Meta-analysis of facial identification rates," *Psychological Bulletin*, vol. 100, 1986, pp. 139-156.

[18] G. L.Wells, S. D. Charman, and E. A. Olson, (2005). Building facial composites can harm lineup identification performance. *Journal of ExperimentalPsychology: Applied*, 11, 147–156.