Perception, 2014, volume 43, pages 214–218

doi:10.1068/p7676

SHORT REPORT

Evaluating training methods for facial image comparison: The face shape strategy does not work

Alice Towler, David White, Richard I Kemp
School of Psychology, The University of New South Wales, Sydney, NSW 2052, Australia;
e-mail: a.towler@unsw.edu.au

Received 18 December 2013, in revised form 12 February 2014

Abstract. Human performance on unfamiliar face matching is known to be highly error prone. However, in organisations where staff are required to perform this task as part of their daily work, attempts are often made to mitigate risk by providing training. Importantly, the methods used in these training courses have not been subjected to empirical validation. In this study we evaluate a common component of many training programmes which encourages viewers to classify face shape. Our results show very low agreement in face shape classification, both within and between participants, and across repeated presentations of a single image to a single participant. Furthermore, face shape classification training did not improve face matching accuracy, suggesting that the face shape strategy does not facilitate identification.

Keywords: unfamiliar face matching, face perception, face recognition, identity, photography

1 The face shape strategy

It is often necessary to establish the identity of unfamiliar people, particularly in security-based and forensic-based practice. Importantly, in these situations identity is frequently verified on the basis of facial images, requiring someone to determine if a photograph on an ID document depicts the person presenting the document, or whether two ID photographs are of the same person. This is a common method of establishing identity in security-critical situations, such as at border crossings, despite psychological research showing that people are quite error prone on this unfamiliar face matching task (eg Bruce et al., 1999; Kemp, Towell, & Pike, 1997). In an effort to improve the ability to detect identity fraud, many organisations—including government agencies in the USA, Europe, and Australia—train staff to use strategies that are intended to improve matching performance. However, the methods used in these training courses have not—to our knowledge—been subjected to empirical validation.

One strategy that is a common component of many face comparison training programmes encourages staff to classify the shape of the face shown in the images according to a prespecified set of templates (eg square, oval, or round) (Facial Identification Scientific Working Group, 2011; see also Spaun, 2009). The premise underlying this face shape strategy appears to be that faces with different shapes must be different people, and therefore face shape classification should provide useful information for identity verification decisions. In support of this approach, the psychological literature does suggest that the external features of a face provide information that is diagnostic of identity. For example, unfamiliar face matching accuracy is often better when people match external facial features, rather than internal features (eg Bruce et al., 1999). However, given the lack of direct empirical validation of the face shape strategy, it remains unclear whether this benefit is conferred by sensitivity to changes in face shape per se, or whether participants are responding to differences in other external features such as ears or hairstyle.

Here, we tested two key assumptions underlying the face shape strategy: firstly, that perceived face shape is diagnostic of identity; and, secondly, that classifying faces in this way improves face matching accuracy. We asked participants to classify 100 face photographs of unfamiliar people according to their face shapes (see figure 1a). Unknown to the participants, the stimulus set contained five images of each of twenty identities, and two of these five images were identical photographs (see figure 1b). In addition, we tested participants' face matching ability using the short version of the Glasgow Face Matching Test (GFMT) (see Burton, White, & McNeill, 2010 for details) before and after they were trained to use the face shape strategy.

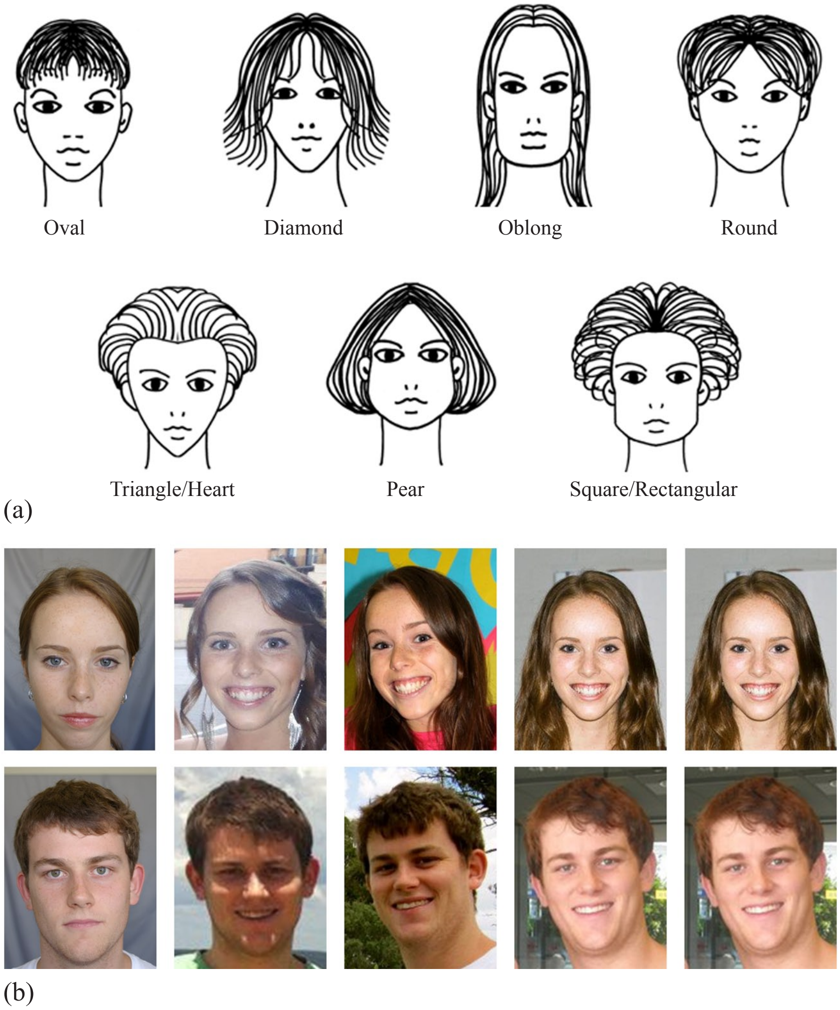

Figure 1. [In colour online, see http://dx.doi.org/10.1068/p7676] (a) Participants were asked to sort 100 face photographs according to these seven face shapes (see text for details). (b) Representative stimuli of two identities used in this experiment.

2 Results

We found that within-rater consistency in shape classification was exceedingly low, with participants judging each identity as having an average of 3 different face shapes ($M = 2.65$; SD = 0.28). Remarkably, in only 7% of cases were all five photos of a given identity judged to have the same face shape. Although this is significantly above chance (0.04%; $t_{27} = 7.54$, $p < 0.001$), such a low level of consistency means that the face shape strategy is unlikely to be useful. Furthermore, participants judged *identical* photos as having the same face shape on only 56% of occasions. Although this rate of consistency is significantly above that predicted by random sorting (14%; $t_{27} = 18.17$, $p < 0.001$), it shows that face shape classification is not stable across repeated presentations of the same image. Further, analysis of interrater reliability using Fleiss's κ (see Fleiss, 1971) also revealed very poor agreement *between* participants ($\kappa = 0.10$). Across raters, a total of 17 of the 20 identities tested were classified as having every one of the possible seven face shape categories on at least one occasion.

One reason for this high level of instability in face shape classification might be that the photographs shown to participants varied slightly with regard to head angle. Many photo ID documents—such as passports—require applicants to supply photos that comply with strict guidelines regarding head angle. To test whether stability of categorisation was greater when head angle did not vary across images, we classified the image set according to passport compliance guidelines for head angle (Australian Passport Office, 2011). A total of 63 of the 100 images met the criteria, leaving each identity with between 1 and 5 passport compliant images. The identical photographs of twelve identities were passport compliant, and so we were able to recalculate within-rater consistency for these pairs. The level of consistency was not greater on this set of identical images compared with the full set of identical images (both 56%, as above). We also calculated interrater reliability on the passport compliant images using Fleiss's κ, and, as before, found very poor agreement between raters on this subset ($\kappa = 0.11$).

Finally, participants' accuracy on the GFMT did not improve after being trained to use the face shape strategy. Mean accuracy was identical at pretest (82%; SD = 13%) and posttest ($M = 82\%$, SD = 14%; $t_{27} = 0.00$, $p = 1$),[1] and there was no change in sensitivity (pretest: $d' = 2.62$; posttest: $d' = 2.77$; $t_{27} = 0.48$, $p = 0.632$) or criterion (pretest: $c = -0.02$; posttest: $c = -0.25$; $t_{27} = 1.26$, $p = 0.219$) from pretest to posttest.

3 Discussion

The key premise underlying the face shape strategy is that perceptual categorisation of face shape is sufficiently stable so that faces perceived to have different shapes are likely to be from different people. In this study we assessed whether or not perceived face shape is diagnostic of identity, and found that it is not. Photographs of the same person were almost never classified as having a single shape—meaning that face shape classification is a very poor predictor of identity. Further, raters frequently judged identical photographs as having different face shapes, indicating that categorical judgments of face shape suffer from low internal consistency. In addition to poor within-rater consistency, we also found poor agreement *between* raters. This is consistent with a previous study where participants were asked to classify facial features according to a fixed taxonomy, and low interrater reliability was also observed (Ritz-Timme et al., 2011).

This inconsistency in face shape judgments is also compatible with previous research showing that people underestimate the range of variation in appearance across multiple images of the same face—by misattributing this within-person variation to changes in identity (Jenkins, White, Van Montfort, & Burton, 2011, experiment 1). Given this instability in identity judgments, it is perhaps unsurprising that we observed high levels of instability in face shape classification. However, it is possible that there are other features which are much more stable across different photos of the same face, and are therefore diagnostic of identity. In future studies it will be important to assess the diagnostic value of individual facial features. Establishing the relative stability of facial features within identity will help inform training strategies by encouraging participants to focus on the set of features that are most diagnostic of identity. On the basis of the results reported here, face shape should be excluded from this feature set.

[1] These data are consistent with normative accuracy scores of 81.3% on this test (SD = 10.4%) (see Burton et al., 2010), suggesting that poor reliability in face shape classification cannot be attributed to the participants having below-average face matching ability.

We also found that face matching performance did not improve following face shape classification. Similar null effects of training have been reported by Woodhead, Baddeley, and Simmonds (1979), who assessed a training programme for unfamiliar face identification that encouraged participants to attend to individual facial features. Matching ability was tested before and after training, and no improvement in performance was observed, suggesting that a featural approach does not improve face matching accuracy (Woodhead et al., 1979, experiments 2 and 3).

We recognise that face shape training is only one component of the training courses for facial image comparison currently being run by various government agencies. It is important to note that other components of these courses, either alone or in combination, may improve matching performance. For instance, these courses often emphasise the variability of appearance between images of one person (eg caused by lighting, image capture settings, or ageing) and encourage staff to distinguish between explainable and unexplainable differences in facial images. This approach is compatible with recent psychological research (see Jenkins et al., 2011), but the effectiveness of this instruction has not been validated. We are also aware of courses that provide feedback on difficult matching decisions, and recent psychological studies have shown that feedback can improve accuracy in matching tasks (Alenezi & Bindemann, 2013; White, Kemp, Jenkins, & Burton, 2014).

In summary, the work reported here suggests that the face shape strategy is not a useful approach for face matching. However, psychological research points to several other training methods which may be more beneficial. Given the security implications of errors in unfamiliar face matching, it is important that the development and validation of occupational training programmes remains a focus for research in this field.

4 Method

4.1 *Participants*

Twenty-eight undergraduate students (mean age = 19.21 years, SD = 1.73 years, eighteen females) participated in return for course credit. This research was approved by the University of New South Wales's Human Research Ethics Advisory Panel (Psychology) in Australia.

4.2 *Materials*

In the shape classification task participants sorted 100 photographs of unfamiliar individuals (members of the USA Olympics team in 2012).[2] These images were sourced from the BBC's 2012 Olympics website and Google, and were not constrained except that the full face had to be visible. The stimulus set consisted of four images and a duplicate image for each of the twenty identities (see figure 1b). Participants were not told that the set contained multiple photographs of the same people; however, it is likely that they noticed the repeated images. The presentation order of photographs was randomised for each participant. Participants were also provided with a 'cheat sheet' containing a description of typical characteristics and two examples of each of the different face shapes. The face shape categories were taken from training courses used by government agencies in the USA, Europe, and Australia. The categories were oval, diamond, oblong, round, triangle/heart, pear, and square/rectangular (see figure 1a). Four versions of this cheat sheet were created, each with the face shapes in a different random order, and one version was randomly allocated to each participant.

To test for training effects, we administered the GFMT (short version) (Burton et al., 2010) before and after the shape classification task. We split this test into two pretests and posttests of equal difficulty using itemised normative data from previous research (Burton et al., 2010). Each test consisted of 20 trials where a pair of faces was presented simultaneously on the screen until the participant responded with a 'same person' or 'different people' decision.

[2] Athletes were selected such that they could not reasonably be expected to be familiar to Australian students. A list of these athletes is available from the authors on request.

4.3 *Procedure*

To assess participants' baseline face matching accuracy, participants completed one half of the GFMT (the version used as pretest was counterbalanced across participants). Participants were then shown a description and examples of each face shape on a computer. These were shown one at a time, and participants could progress to the next face shape in their own time. Afterwards, participants were given the cheat sheet containing the descriptions and examples of each face shape to refer to throughout the face shape classification task and posttest.

The classification task began with one face photograph presented on the screen with buttons for each of the seven face shapes. The face remained on the screen until the participant selected the face shape category they believed that particular face belonged to. The time allowed to make a response was not limited, and the location of the buttons on the screen corresponded with the order on each participant's cheat sheet. Once a selection was made, the next trial began with the presentation of another face. Following the completion of all 100 trials in the shape classification task, participants completed the second half of the GFMT. The entire procedure generally took 15 min to complete.

Acknowledgments. This research was supported by an Australian Postgraduate Award to Alice Towler and an Australian Research Council Linkage Grant to Richard Kemp (LP110100448), in partnership with the Department of Foreign Affairs and Trade (Australian Passport Office).

References

Alenezi, H. M., & Bindemann, M. (2013). The effect of feedback on face-matching accuracy. *Applied Cognitive Psychology*, **27**, 735–753.

Australian Passport Office. (2011). *Photograph Guidelines*. Retrieved from https://www.passports.gov.au/images/photo_guidelines.pdf

BBC. (2012). *Team USA Athletes*. Retrieved from http://www.bbc.com/sport/olympics/2012/countries/united-states/athletes

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, **5**, 339–360.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, **42**, 286–291. doi: 10.3758/BRM.42.1.286

Facial Identification Scientific Working Group. (2011). *Guidelines and Recommendations for Facial Comparison Training to Competency (Version 1.1)*. Retrieved from https://www.fiswg.org/document/viewDocument?id=22

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378–382.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, **121**, 313–323. doi:10.1016/j.cognition.2011.08.001

Kemp, R. I., Towell, N. A., & Pike, G. E. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, **11**, 211–222.

Ritz-Timme, S., Gabriel, P., Obertova, Z., Boguslawski, M., Mayer, F., Drabik, A., … Cattaneo, C. (2011). A new atlas for the evaluation of facial features: Advantages, limits, and applicability. *International Journal of Legal Medicine*, **125**, 301–306.

Spaun, N. A. (2009). Facial comparisons by subject matter experts: Their role in biometrics and their training. In M. Tistarelli & M. S. Nixon (Eds.), *Advances in Biometrics* (pp. 161–168). Heidelberg: Springer-Verlag.

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*, **21**, 100–196. doi: 10.3758/s13423-013-0475-3

Woodhead, M. M., Baddeley, A. D., & Simmonds, D. C. V. (1979). On training people to recognize faces. *Ergonomics*, **22**, 333–343.