Redesigning photo-ID to improve unfamiliar face matching performance

David White[1], A. Mike Burton[2], Rob Jenkins[3] & Richard I. Kemp[1]

[1] School of Psychology, The University of New South Wales, Australia

[2] School of Psychology, University of Aberdeen, UK

[3] Department of Psychology, University of York, UK

ACKNOWLEDGEMENTS:

This research was supported by an ARC grant to Kemp (LP110100448), a bilaterally funded grant to Kemp (ARC: LX0083067), Burton and Jenkins (ESRC: RES-000-22-2519) and an ESRC Professorial Fellowship to Burton (ES/J022950/1). We thank Mark Howard for his help collecting data for Experiment 1, and also Graham Nisbett, Filippo Caranti and the Eva Renaldi for making the photographs in Figure 1 available for publication under Creative Commons licenses (CC BY 2.0).

(71 words)

CORRESPONDING AUTHOR:

Dr. David White

School of Psychology

The University of New South Wales

Kensington

Sydney

NSW 2052

david.white@unsw.edu.au

(+61) 2 9385 3254

RUNNING HEAD: Redesigning photo-ID

WORD COUNT (excluding abstract and figure captions): 5160 words

ABSTRACT

Viewers find it difficult to match photos of unfamiliar faces for identity. Despite this, the use of photographic ID is widespread. In this study we ask whether it is possible to improve face matching performance by replacing single photographs on ID documents with multiple photos or an average image of the bearer. In three experiments we compare photo-to-photo matching with photo-to-average matching (where the average is formed from multiple photos of the same person) and photo-to-array matching (where the array comprises separate photos of the same person). We consistently find an accuracy advantage for average images and photo arrays over single photos, and show that this improvement is driven by performance in match trials. In the final experiment, we find a benefit of four-image arrays relative to average images for unfamiliar faces, but not for familiar faces. We propose that conventional photo-ID format can be improved upon, and discuss this finding in the context of face recognition more generally.

(159 words)

KEYWORDS

Face Recognition; Unfamiliar Face Matching; Identity Verification; Facial Image Comparison; Image Averaging.

INTRODUCTION

Photo ID documents are frequently used as proof of identity. Despite recent advances in biometric technology and storage capacity of identity documents (e.g. passports), facial appearance remains the most common means of checking identity at borders. In addition, photo-ID is often required in everyday settings, for example when purchasing age-restricted goods such as alcohol or tobacco. However, the widespread use of photo-ID is at odds with psychological research, which consistently finds that viewers perform poorly when matching unfamiliar people to their photos.

Estimates of human face matching performance vary depending on specifics of the task. However even under optimal conditions people are surprisingly inaccurate at identity verification from photographs. In an early study, Kemp, Towell and Pike (1997) found that supermarket cashiers made over 30% errors when verifying the identity of shoppers from Photo-ID cards, despite knowing they were taking part in a trial. When an attempt was made to match foils to similar photos (same gender, ethnicity, similar age and hairstyle), false acceptance rates rose to over 60%. Laboratory-based matching studies have tended to use photo-to-photo matching tasks, but also show high error rates (e.g. Bruce et al, 1999, 2001; Clutterbuck & Johnston, 2002; Megreya & Burton, 2006). However, when these studies are extended to include live matches, equivalently poor performance is seen. For example, Megreya & Burton (2008) reported an error rate of over 15% in a task requiring viewers to match a person to a recent high quality photo, even though no time limit for decisions was imposed.

Importantly, face-matching performance is transformed by familiarity. Across a wide range of identification and matching tasks, viewers are consistently excellent at recognizing familiar faces, even under very poor viewing conditions (Burton et al, 1999; Jenkins, White, Van Monfort & Burton, 2011; Hole, George, Eaves & Rasek, 2002). Indeed, performance on matching tasks has been shown to be a good index of familiarity (Clutterbuck & Johnston, 2002, 2004, 2005). Furthermore, superficial image changes (e.g. viewing angle, expression) severely impair identification of unfamiliar faces, but do not impair identification of familiar faces (e.g., Bruce, Valentine & Baddeley, 1987; Hancock, Bruce & Burton, 2000; Hill & Bruce, 1996).

Poor levels of performance lead one to ask how the problem might be addressed in practical settings. Perhaps it is possible to improve photo-ID by replacing the photograph with something that viewers find easier to match. One alternative might be to use video rather than photos on ID cards, as many cards now contain chips with sufficient storage for this. As it turns out, matching a person to a simultaneously presented high-quality video does not solve the problem (Davis & Valentine, 2009; Experiment 3). In Davis & Valentine's (2009) study, both hits and false alarms were unacceptably high - with error rates of 26% in match, and over 40% in mismatch trials - for video clips that were captured just one week earlier. Here we take a different approach, asking whether aspects of *familiar* face recognition, which is known to be highly accurate, can be built into the unfamiliar matching task.

Burton et al (2005) proposed a model of familiarity-based performance based on averaging together multiple images of the same face (Jenkins & Burton, 2011). According to this model, a stored representation is incrementally refined with each

encounter. The effect of adding more images to the average is thus to eliminate superficial differences, while preserving aspects of the images that are common across photos. By this process, the representation comes to emphasize unchanging features of the face that are diagnostic of the particular identity. An average image has been shown to be a useful representation for automatic computer-based face recognition systems, in the sense that matching new photos to an average gives much better performance than matching new photos to an existing photo (Jenkins & Burton, 2008a).

In this paper we ask whether a similar advantage for average images is seen in human performance. Using an unfamiliar face matching task, we tested whether viewers perform better when matching a photo to an average image than when matching two photos. For comparison, we also tested whether matching a photo to a photo array confers any advantage. Importantly, photo arrays preserve information about within-person variability in appearance (see Jenkins et al. 2011), whereas average images emphasize central tendency. Variance information could potentially boost performance by indicating the range of possible images that each face can project.

EXPERIMENT 1

In this experiment we test whether it is easier to match a face photo to another photo or to an average image. We compare performance for familiar and unfamiliar faces by testing participants in two locations (UK and Australia), and presenting images of national celebrities who are famous in only one of these locations. In this way, the

same stimuli can be used as both familiar and unfamiliar faces, eliminating any potential confound between stimulus set and familiarity.

Method

Participants

Participants were 44 volunteers from The University of New South Wales, Australia (28 females, mean age 19.5) and 44 volunteers from University of Glasgow, UK (27 female, mean age 23.7).

Stimuli and Materials

We constructed a stimulus set based on 40 UK national celebrities and 40 Australian national celebrities. These celebrities were chosen to be known by participants in one country, but not the other (for example, national TV presenters, sports personalities, politicians). For each of the 80 celebrities, we collected 13 images using Google Image search. The images thus sampled natural variability in facial, environmental, and image-level parameters (Jenkins et al., 2011). We constrained image selection by accepting only those that were of sufficient resolution (minimum 80 pixels between the eyes), and where head-angle was no more than twenty degrees from full face. For each celebrity, 12 photos were randomly selected to form the average image, and the remaining photo was set aside for use as the target photograph in the matching task. To construct the average image, we co-registered the twelve photos of each face by aligning landmark anatomical features to a standard face template using in-house image morphing software. This allowed us to calculate the average RGB values of each pixel in a linear space. These ‘shape-free’ average textures were then morphed

back to the average shape of the twelve images to produce the final average (for details see Burton et al., 2005).

For each celebrity, four stimulus-pairs were created: photo-photo and photo-average pairs, in both match (same identity) and mismatch (different identity) combinations. As our average face images are automatically cropped to remove extraneous background, we cropped the comparison (i.e. non-target) image in the photo-photo pairings in the same way (Jenkins & Burton, 2008b). All images were presented on a computer monitor at a resolution of 200 by 300 pixels (see Figure 1 for example stimuli).

---------- FIGURE 1 ----------

Design and Procedure

Participants completed a 160-trial face-matching test (one match and one mismatch trial per celebrity). Each trial comprised a (target) photo of a celebrity on the left side of the screen and either a second (comparison) photo or an average image on the right. Comparison photos were selected at random from the same set that had been used to create the average images. For match trials, the target photo was of the same celebrity, and for mismatch trials the target photo depicted a different unfamiliar face that matched the same basic verbal description as the target face (e.g. young adult male with dark hair). Participants indicated same identity or different identity judgments via keypress. The task was self paced, and stimuli remained on screen until a response was made. Trial order was randomised throughout.

The familiarity manipulation was then checked by showing participants the names of all the celebrities that were presented in the experiment, and asking whether they were familiar with each person's face. As expected, familiarity with home celebrities was high (UK, M = 34.4 SD = 5.5; Australia, M = 26.9, SD = 10.6), and familiarity with overseas celebrities was low (UK, M = 2.0 SD = 2.8; Australia, M = 1.2, SD = 1.6). Home celebrities that turned out to be unfamiliar, and overseas celebrities that turned out to be familiar, were excluded from analysis for each subject.

Results

For all experiments in this paper we present accuracy separately for match and mismatch trials. Previous research has shown that face matching accuracy on match trials is not predictive of accuracy on mismatch trials (e.g. Megreya & Burton, 2007). For this reason, we chose not to rely on statistics that combine these measure of performance. However, for the interested reader, we also provide analysis of non-parametric Signal Detection Theory statistics in Supplementary Materials (A' and B"; see Stanislaw & Todorov, 1999).

---------- FIGURE 2 ----------

Accuracy data for Experiment 1 is shown in Figure 2. For match trials, two-way ANOVA with the within-subjects factors of Familiarity (Familiar, Unfamiliar) and Image Type (Photo, Average) revealed significant main effects of both Familiarity, [$F(1,87) = 219$, $p<0.01$] and image type, [$F(1,87) = 37.5$, $p<0.01$], as well as a significant interaction between these factors [$F(1,87) = 4.36$, $p<0.05$]. Simple main effects showed an advantage for average images over photos in both the Unfamiliar

condition [F (1,87) = 27.5, p < 0.01, Cohen's d = 1.124], and the Familiar condition [F (1,87) = 10.7, p<0.01, d = 0.703], with the interaction being driven by a larger effect for unfamiliar faces. For mismatch trials there was a significant effect of Familiarity, [F (1,87) = 18.4, p<0.01], but no effect of Image Type and no interaction (Fs < 1).

Discussion

Consistent with all previous research (e.g. Clutterbuck & Johnston, 2005; Jenkins et al 2011), matching was more accurate for familiar than unfamiliar faces. More importantly for the current study, we found better face matching performance for average images than for single photos. Rather surprisingly, this was true for familiar faces as well as unfamiliar faces, despite high overall accuracy in the familiar condition. One possible interpretation of this finding is that familiarity was not asymptotic for these national (as opposed to global) celebrities.

Overall, the results demonstrate a performance boost for average images that may be of practical benefit. The averaging technique eliminates some of the transient characteristics of a photograph that profoundly affect appearance, but are irrelevant to identity (e.g. effects of lighting direction). Current technology would allow such digital images to be stored on photo-ID cards, potentially improving identification accuracy by human operators as well as for machines (Jenkins & Burton, 2008a). In the next experiment, we ask whether matching to a photo array might also yield performance benefits.

EXPERIMENT 2

In this experiment we tested whether photo-ID might be improved if it contained more than one photograph of the bearer. Identifying unfamiliar people from photographs is difficult, because a person's appearance varies from one snapshot to the next (Jenkins et al., 2011). We reasoned that incorporating such variation into a photographic representation might make the task easier. Because photo arrays are likely to require more elaborate processing than single images, we also included a study duration manipulation, to test whether any benefit of photo arrays requires extended study time.

Method

Participants

Seventy-two undergraduates from University of New South Wales participated in the study (36 female; mean age 19.7 years, SD = 2.8). None had participated in previous experiments in our lab.

Stimuli

In this experiment we used photographs of 80 people who were unfamiliar to our Australian participants (as verified at the end of the experiment). Thirty of these were UK celebrities used in the previous experiment, and the remaining fifty were consenting undergraduate psychology students who volunteered photos of themselves from their Facebook accounts. From this set we selected six photos of each face at random for use in the experiment. One of these was chosen at random to be the target

photo and the remaining five were used as array photos. For each identity we then selected a similar looking person from existing databases to use as foils in mismatch trials. All images were presented in full colour, cropped to a 2:3 aspect ratio and scaled to 200 x 300 pixels.

---------- FIGURE 3 ----------

Design and Procedure

Trial Type (match, mismatch) and Array Size (1, 2, 3, 4 photos) were manipulated within-subjects, and Study Time (3 sec, 6 sec, 9 sec) was manipulated between-subjects. Participants were allocated to one of the three Study Time groups at random. Participants completed a 160 trial face-matching test (one match and one mismatch trial per identity). Each trial consisted of a target image on the left side of the screen and a photo array on the right. Array photos were selected at random from the five available photos on a trial-by-trial basis, and were presented in a random order in a predefined display configuration (see Figure 3 for an example display). For match trials, the photo array was presented alongside the target photo. For mismatch trials, the array was paired with the foil photograph.

On each trial, the participants' task was to decide whether the person on the left side of the display was the same as the person on the right. We specifically instructed participants that photos appearing on the right side (i.e. the array) would always show the same person. As in Experiment 1, participants indicated same person or different person decisions via keypress. The task was self-paced, and stimuli remained on screen until response. After each decision, participants rated their confidence on a

scale from 1 to 100, so that we could relate objective performance to decisional confidence. Trials were presented in a random order. Counterbalancing was achieved by rotating stimulus identities through Array Size conditions across participants, so that each identity was presented in each condition an equal number of times[1].

Participants responded by clicking on onscreen response buttons, and Study Time was manipulated by delaying presentation of these buttons. Participants were instructed that the delay should be used to study the faces, and were asked to respond quickly and accurately once the response buttons appeared.

Results

---------- FIGURE 4 ----------

Accuracy data for Experiment 2 are shown in Figure 4. A three-way mixed ANOVA with the within-subjects factors of Trial Type (match, mismatch) and Array Size (1, 2, 3, 4), and the between-subjects factor of Study Time (3 sec, 6 sec, 9 sec) revealed significant main effects of Trial Type [$F(1, 69) = 14.4$, $p<0.05$] and Array Size [$F(3,69) = 10.6$; $p<0.05$], but no main effect of Study Time [$F(2,69) = 1.43$, $p<0.05$]. The three-way interaction between these factors was not significant ($F < 1$), and neither were the two-way interactions between Study Time and Trial Type, and between Study Time and Array Size ($Fs < 1$). Thus Study Time did not affect performance in this situation.

[1] We did not counterbalance target image through array image positions, because it was not clear how to achieve this for mismatch trials. However, we note that this method of counterbalancing would provide a better model for the use of photo-ID in real world situations, where the appearance of 'targets' would vary across encounters.

There was a significant interaction between Trial Type and Array Size [$F(3,207) = 21.3$, $p < 0.05$]. Simple Main Effects revealed a significant effect of Array Size for match trials [$F(3,213) = 38.4$, $p < 0.05$], but not for mismatch trials, [$F(3,213) = 1.75$, $p > 0.05$]. The effect of Trial Type was non-significant for single-photo arrays ($F < 1$, $d = 0.063$), but significant for array of two photos [$F(1,71) = 12.6$, $p < 0.05$, $d = 0.840$], three photos [$F(1,33) = 29.3$, $p < 0.05$, $d = 1.284$], and four photos [$F(1,33) = 27.5$, $p < 0.05$, $d = 1.244$].

We also carried out planned comparison t-tests to break down the main effect of Array Size. Because there were no significant main effects or interactions involving study time, we collapsed across this factor before preceding. Overall accuracy was 79.8% (SD = 8.1) for one-photo arrays, 83.0% (SD = 11.3) for two-photo arrays, 82.6% (SD = 13.2) for three-photo arrays, and 85.4% (SD = 9.2) for four-photo arrays. Planned comparisons revealed a significant difference between one-photo and two-photo arrays [$t(71) = 2.49$, $p < 0.05$, Cohen's $d = 0.325$], but no differences between two-photo and three-photo arrays ($t < 1$, $d = 0.033$) or between three-photo and four-photo arrays [$t(71) = 1.59$, $p > 0.05$, $d = 0.246$].[2]

Response times and confidence ratings were also collected in this experiment. These measures both corroborated the accuracy measure, showing that participants were more confident in their correct decisions when matching multiple-photo arrays, compared with single-photo arrays. As with accuracy data, this effect was found for match trials only, and saturated at array size two. Response time data confirmed that

[2] This pattern was also obtained in a separate experiment that excluded the Study Time factor (see Supplementary Materials, page 7).

the performance improvement was not due to a speed-accuracy tradeoff, as responses in match trials were faster for multiple photo arrays than for single photo arrays (see Supplementary Materials for full details of this analysis).

Analysis of Similarity Ratings

Our findings show that face matching performance can be improved by presenting multiple comparison photos. We have previously argued that a single photographic sample may not contain sufficient data for purposes of identification (Jenkins & Burton, 2011). Evidently, additional samples go some way to solving that problem.

Although it is beyond the scope of this paper to specify the cognitive mechanism of the observed performance enhancement, we note that there are at least two broad processes that could account for a multiple-photo advantage. One possibility is that the identity decision is dominated by the array photo that is most similar to the target photo. Alternatively, viewing multiple images may lead the participant to construct a more abstract representation of the face against which to match the target.

We attempted to distinguish between these accounts by collecting similarity ratings for all target and array photos presented in Experiment 2. Our aim was to establish whether trial performance was better predicted by the similarity between the target photo and the best (most similar) array photo, or by the average similarity between the target photo and the array photos. To this end, 28 participants (17 Female; Mean Age = 19.3; SD = 2.1) each rated half of 800 comparisons. As it turned out, *best item similarity* and *average similarity* were themselves very highly correlated (pooled

Spearman's rho = 0.91), so it was not possible to distinguish between the two accounts using this method (see Supplementary Materials for methods and analysis).

Discussion

Our results show that multiple-photo arrays can improve unfamiliar face matching performance. As with Experiment 1, this advantage was observed for match but not mismatch trials, so that the overall improvement was driven by increased accuracy in detecting true matches. Requiring participants to spend more time on their decisions did not improve performance. This might suggest that the critical information can be extracted from multiple photographs rather quickly (i.e. within three seconds).

Alternatively, performance may be limited by the cognitive demands of processing information from multiple face images, rather than by the information in the images. Previous research has shown that in some circumstances, face identity processing can be subject to strict capacity limits (see Bindemann et al. 2005, 2007). Thus the information advantage of multiple photos may be partly offset by the increased processing demands that they impose. If so, it is possible that a single average image might be preferable to an array of separate photographs. We test this possibility in the final experiment.

EXPERIMENT 3

In the final experiment, we directly compared unfamiliar face matching performance for two different types of face representation - average images and photo arrays. In the previous experiment we found that two-photo arrays improved performance over single comparison photographs, and that increasing array size further yielded no

additional (statistically significant) benefit. However, since overall accuracy was numerically highest for four-photo arrays (85.4%), we used four photos for the arrays in this experiment.

Method

Participants were 28 volunteers from University of New South Wales, Australia (13 females; mean age = 20.7) and 28 volunteers from University of Glasgow, UK (17 females; mean age = 24.2).

The procedure was the same as Experiment 1, except that we replaced the photo condition with a four-photo array condition. Participants completed a 160-trial face-matching test (one match and one mismatch trial per celebrity). Each trial comprised a photo of a celebrity on the left side of the display and either an average image or a photo array on the right. As in Experiment 1, average images were constructed from 12 photos of the person. Four-photo arrays were generated on a trial-by-trial basis, by selecting four of these twelve photographs at random.

For match trials, the target and comparison images showed the same person. For mismatch trials, the comparison image was of a different unfamiliar face that matched the same basic description as the target. Participants were asked to indicate whether the face on the left (target) was the same as the face on the right (average image or photo array). As in the previous experiment, it was made clear to participants that the four photos in any array always showed the same person.

Results

---------- FIGURE 5 ----------

Accuracy data from Experiment 3 is shown in Figure 5. The overall pattern is similar to that observed in Experiment 1, except that here we found an advantage for photo-arrays over average images. For match trials, a two-way within subject ANOVA with factors Familiarity (Familiar, Unfamiliar) and Image Type (Average Image, Photo Array) revealed significant main effects of both Familiarity [$F(1,55) = 106$, $p < 0.01$] and Image Type, [$F(1,55) = 6.97$, $p < 0.01$], as well as a significant interaction between these two factors [$F(1,55) = 5.91$, $p < 0.05$]. Simple Main Effects confirmed that the performance benefit for photo arrays was significant for unfamiliar faces [$F(1,55) = 9.48$, $p < 0.01$, Cohen's $d = 0.415$] but not for familiar faces ($F < 1$, $d < 0.01$).

For mismatch trials there was a significant main effect of Familiarity only [$F(1,55) = 5.61$, $p < 0.05$], with no significant effect of Image Type ($F < 1$) and no interaction ($F < 1$). Thus, matching performance using photo arrays exceeded performance using average images, but this benefit was specific to same-person trials and unfamiliar faces.

GENERAL DISCUSSION

In all three experiments, we found that alternatives to single-photograph representations of faces can improve face matching performance. In Experiment 1, matching a photograph to an average image was more accurate than matching two photographs. In Experiment 2, matching a photograph to a multi-photo array was better than matching two photographs. Finally, in Experiment 3, matching a

photograph to a multi-photo array was better than matching a photograph to an average image. Our findings have important implications for face matching in occupational settings. Foremost, they demonstrate that single-photo representations of faces are suboptimal, and could be superseded by representations that incorporate within-person variability. Either stabilizing the variability (by image averaging) or increasing the number of samples (by presenting multiple photos) improves matters.

The specific pattern of improvement for average images and photo arrays was also consistent across experiments. In each experiment, improvement was observed only in trials where target and comparison images showed the same person (match trials). Apparently, providing more information about a person's appearance allows a viewer more accurately to identify that person in a true match. Importantly, this benefit was not accompanied by a general response bias to make 'same person' responses, because accuracy on mismatch trials was always unaffected by extra visual information. This is an important point, as it shows that the performance benefit observed in match trials does not come at the cost of a performance decrement in mismatch trials. Instead, we find a net gain in accuracy. In particular, it appears that presenting multiple photographs of a face allows participants to be more accommodating of within-person variance in appearance. Future research may discover complementary methods for improving mismatch performance without impairing match performance, as required for detection of identity fraud.

It should be noted that this pattern of results is somewhat discrepant with our original motivation. We sought to map aspects of familiar face processing onto unfamiliar face processing to improve performance in the latter domain. Previous research

demonstrating familiarity-based improvement on matching tasks has found that familiarity improves performance on both match *and* mismatch trials (e.g. Megreya & Burton, 2006, 2007; Clutterbuck and Johnston, 2002, 2004, 2005). Although we also found enhanced performance for familiar faces, the effect here was more pronounced for match trials than for mismatch trials, perhaps due to the broader heterogeneity of our stimulus images. Nevertheless, we observed the advantage for averages and photo-arrays *only* in match trials, suggesting that these formats confer partial benefits of familiarity. Previous studies typically report small effects of image-based familiarization procedures on matching performance (e.g. Clutterbuck & Johnston, 2005; Osborne & Stevenage, 2008), or find that it does not improve accuracy at all (Clutterbuck & Johnston, 2005). One direction for future research might be to develop methods that accelerate the process of familiarization, and enhance the improvement in matching performance seen here.

Another fruitful direction for future studies would be to manipulate within-person image homogeneity as a variable in its own right. Doing so should help to establish whether it is the similarity of the closest matching photograph, or the similarity of the entire array that drives improved performance in photo array conditions. In previous work (Burton et al, 2005; Jenkins & Burton, 2011), we have proposed that familiar face recognition is highly accurate *precisely because* it is based on representations that summarise within-person variability in appearance. The resulting representations are robust, in the sense that they can be matched to novel images of the same person, provided that these vary in ways that are consistent with previous perceptual experience. The idea behind the image formats tested here is to build variability into

the representation. If unfamiliar viewers are exposed to some variability in the target person, they may benefit from some of the advantage of face familiarity.

Finally, we should note that all the experiments here use photo-to-image matching. In real photo-ID settings, people are usually asked to make a match to a live person. In fact, the relatively small literature comparing photo-to-image and live-to-live matching has found surprisingly little difference in performance between the two (Kemp et al, 1997; Davis & Valentine, 2009; Megreya & Burton, 2008). For this reason, we expect that the performance benefits seen here would generalise to live face-matching settings. Nevertheless, it is important to test this, and such experiments will form the basis of future work. In that work it will be important to establish not only whether the basic improvements in photo-ID format are observed, but also how they might interact with characteristics of the observer (e.g. face recognition aptitude: White, Kemp, Jenkins & Burton, 2013), and with realistic environmental factors such as time constraints and cognitive load.

In summary, we have shown that traditional forms of photo-ID could be improved by replacing individual photographs with representations derived from multiple photos of the same face. Based on our current findings, we expect that this would have a beneficial effect on identity verification procedures in occupational settings. Future research should determine the optimal range of within-person variability, and how best to summarize it. For now, it is clear that a single photograph is not the best way to represent facial appearance.

REFERENCES

Bindemann, M., Burton, A. M., Hooge, I. T. C., Jenkins, R. & De Haan, E. H. (2005). Faces retain attention. *Psychonomic Bulletin & Review*, *12*, 1048-1053.

Bindemann, M., Jenkins, R., & Burton, A. M. (2007). A Bottleneck in Face Identification. *Experimental Psychology*, *54*(3), 192–201.

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, 7*, 207-218.

Bruce, V., Henderson, Z., Newman, C. & Burton, A.M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7,* 207-218.

Bruce, V., Valentine, T., & Baddeley, A. (1987). The Basis of the 3/4 View Advantage in Face Recognition. *Applied Cognitive Psychology*, *1*, 109–120.

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

Burton, A. M., Jenkins, R., Hancock, P. J. & White, D. (2005). The power of averages: Robust Representations for Face Recognition. *Cognitive Psychology*, *51*, 256- 284.

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity using an indirect face-matching measure. *Perception, 31*(8), 985-994.

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition, 11*, 857-869.

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology, 17*, 97-116.

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, *23*(4), 482–505.

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

Hill, H. & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 986–1004.

Hole, G., George, P., Eaves, K. & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception, 31,* 1221-1240.

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

Jenkins, R., & Burton, A. M. (2008a). 100% Accuracy in Automatic Face Recognition. *Science*, *319*, 435–435.

Jenkins, R., & Burton, A.M. (2008b). Response to comment on "100% accuracy in automatic face recognition". *Science, 321*, 912d.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1671–1683.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121,* 313-323.

Megreya, A, M. & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34*, 865-876.

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14*, 364–372.

Osborne, C. D., & Stevenage, S. V. (2008). Internal feature saliency as a marker of

familiarity and configural processing. *Visual Cognition*, *16*(1), 23-43.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, *31*(1), 137-149.

White, D., Kemp, R. I., Jenkins, R. & Burton, A. M. (2013). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*. Advance online publication. doi: 10.3758/s13423-013-0475-3

Figure 1. Example image pairs used in Experiment 1. In each trial a comparison image (left) was paired with either an average image (top row) or a single photograph (bottom row). Image pairs in the left column show the same person (match). Image pairs in the right column show different people (mismatch).

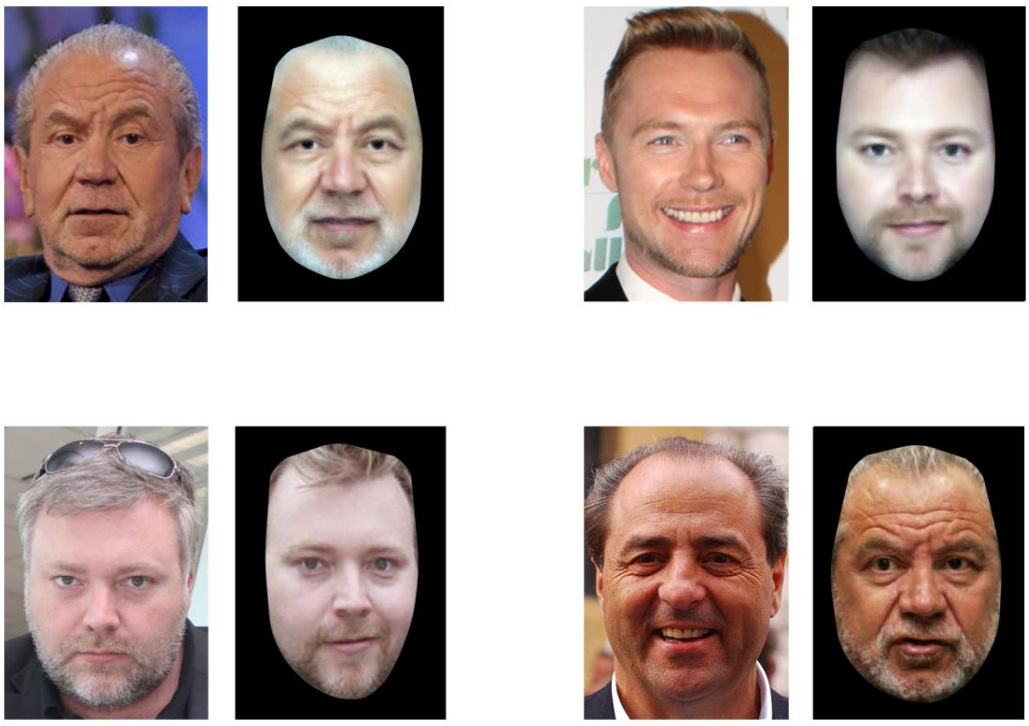


Figure 2. Mean accuracy (percent correct) for the face matching task in Experiment 1 (± standard error).

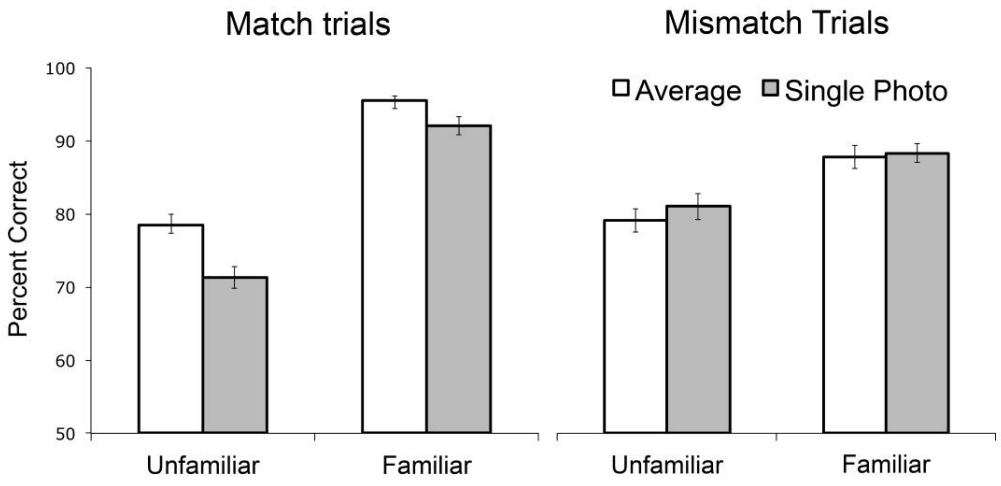


Figure 3. Example stimulus displays from Experiment 2, showing each of four Array Size conditions. Displays in the left column (one-photo and three-photo arrays) show match trials, and arrays in the right column (two-photo and four-photo arrays) show mismatch trials.

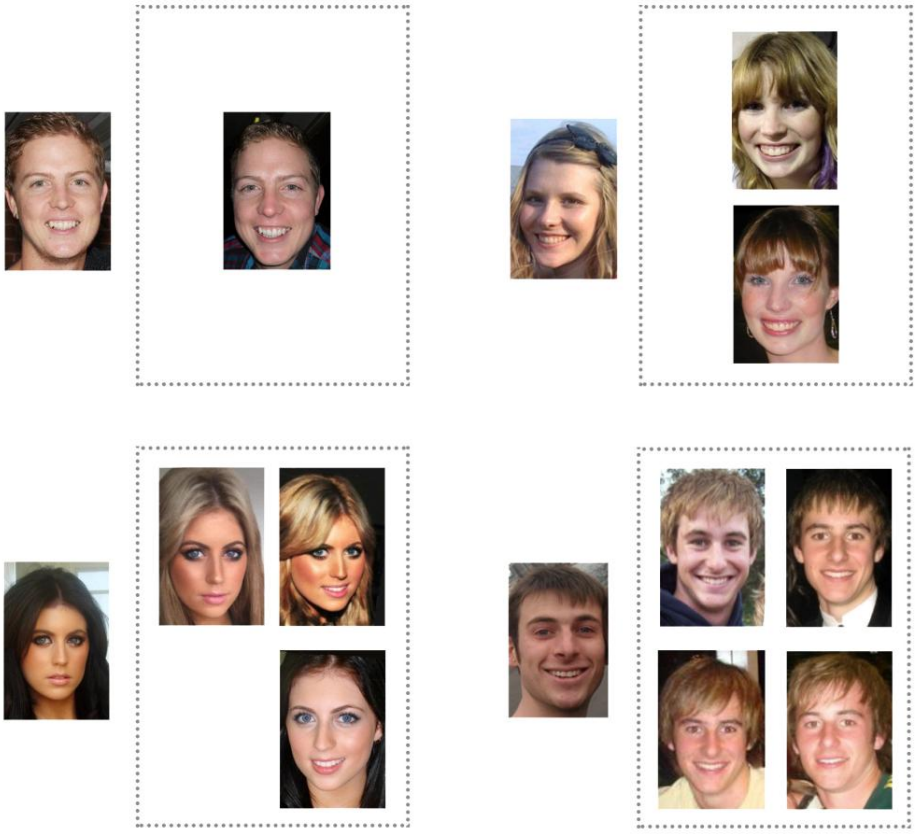


Figure 4. Mean accuracy (percent correct) for the face matching task in Experiment 2 (± standard error).

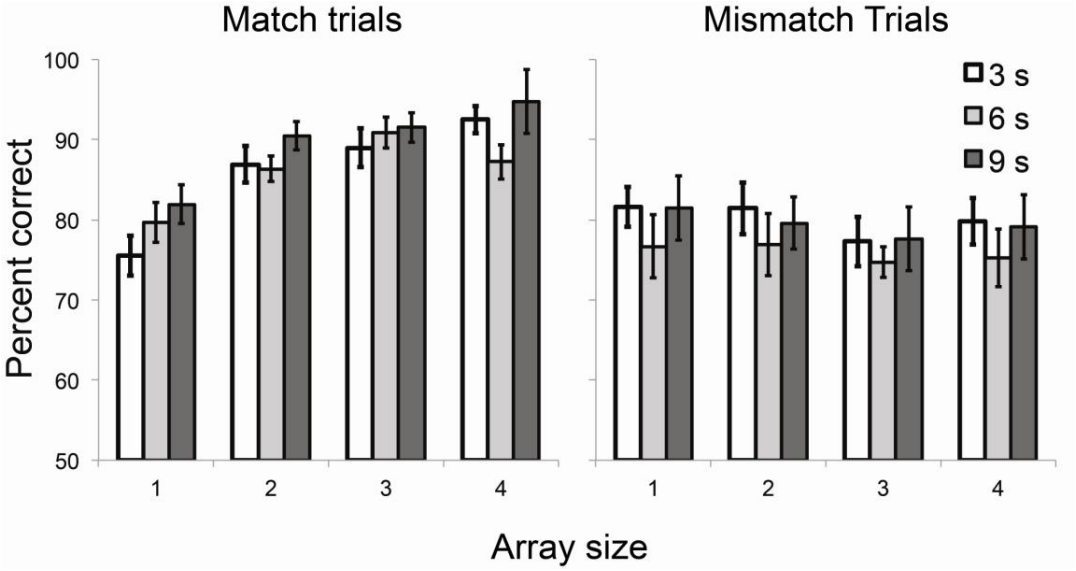
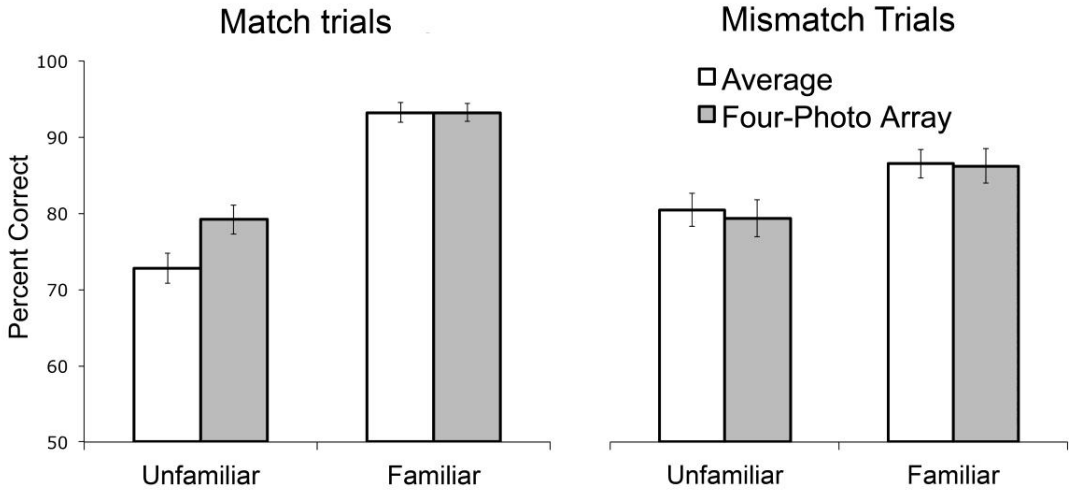


Figure 5. Mean accuracy (percent correct) for the face matching task in Experiment 3 (± standard error).



Supplemental Materials
Click here to download Supplemental Materials: XAP_white_1032_supp.pdf