Applied Cognitive Psychology, *Appl. Cognit. Psychol.* **27**: 769–777 (2013)
Published online 11 November 2013 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.2971

Crowd Effects in Unfamiliar Face Matching

DAVID WHITE[1]*, A. MIKE BURTON[2], RICHARD I. KEMP[1] and ROB JENKINS[3]

[1]*School of Psychology, The University of New South Wales, Sydney, Australia*
[2]*School of Psychology, University of Aberdeen, Aberdeen, UK*
[3]*Department of Psychology, University of York, York, UK*

Summary: Psychological research shows that humans can not reliably match unfamiliar faces. This presents a practical problem, because identity verification processes in a variety of occupational settings depend on people to perform these tasks reliably. In this context, it is surprising that very few studies have attempted to improve human performance. Here, we investigate whether distributing face matching tasks across groups of individuals might help to solve this problem. Across four studies, we measure the accuracy of the 'crowd' on a standard test of face matching ability and find that aggregating individual responses produces substantial gains in matching accuracy. We discuss the practical implications of this result and also suggest ways in which this approach might be used to improve our understanding of face perception more generally. Copyright © 2013 John Wiley & Sons, Ltd.

INTRODUCTION

Face photographs are often used to verify personal identity. For example, individuals may be required to produce photo-identification (ID) documents when crossing borders or when making financial transactions. In other settings, comparison of facial images from closed-circuit television footage can play an important role in identifying the perpetrators of crimes. Given the pivotal role of face matching decisions in person ID, it is of practical significance that viewers are surprisingly poor at matching unfamiliar faces.

Poor levels of performance have been found in a variety of face matching tasks. In a pioneering experiment, Kemp, Towell, and Pike (1997) found that retail assistants accepted fraudulent photo-ID on over 50% of trials, despite knowing their performance was being monitored. In a laboratory task, Bruce et al. (1999) reported an error-rate of 30% when participants had to pick out a target face from an array of 10 high-quality mug shots. Error rates also remain unacceptably high (typically $> 10\%$) for pairwise comparisons, in which viewers decide whether two photos show the same person or different people (Burton, White, & McNeill, 2010; Megreya & Burton, 2006).

In sum, humans can not reliably match unfamiliar faces (Burton & Jenkins, 2011). One possible response to this fact is to replace human viewers with automatic face recognition systems, as has begun to happen in some applied settings, notably border control. However, although there have been significant improvements in the accuracy of automatic face recognition in recent years (O'Toole, Phillips, et al., 2007; Phillips et al., 2011), it is also clear that these technologies do not work perfectly and are especially error-prone in unconstrained environments (Phillips et al., 2012). As a result, human operators are often required to compare candidate images that are suggested by the computer system. Indeed, because computers are able to search very large databases for matching faces, their introduction has actually *increased* human workload for face matching tasks in some circumstances (Graves et al., 2011; White, Tan, & Kemp, 2013).

*Correspondence to: Dr David White, School of Psychology, The University of New South Wales, Kensington, Sydney NSW 2052, Australia.
E-mail: david.white@unsw.edu.au

Given the intensive and sustained effort devoted to improving automatic face recognition, it is perhaps surprising that so little effort has been made to improve human performance on this task. With the exception of a handful of attempts to improve accuracy though training (e.g., White, Kemp, Jenkins, & Burton, 2013; Woodhead, Baddeley, & Simmonds, 1979), familiarization (e.g., Clutterbuck & Johnston, 2005), and changes in image format (White, Burton, Jenkins, & Kemp, in press), few psychological studies have attempted to raise accuracy on face matching tasks above baseline performance.

In this paper, we investigate whether unfamiliar face matching can be improved by aggregating the responses of groups of individuals. This approach is inspired by research showing that average estimates made by groups of people are often remarkably close to veridical (Galton, 1907). In difficult tasks that attract large variation in estimates across individuals—such as guessing the number of beans in a jar—the average group estimate is highly accurate (Krause, James, Faria, Ruxton, & Krause, 2011). Indeed, it is often the case for cognitive tasks that the group estimate is more accurate than the best individual estimate (Kerr & Tindale, 2004), which has led to this grouping effect being termed the *wisdom-of-crowds* phenomenon (Surowiecki, 2004).

Importantly, the wisdom-of-crowds effect does not occur for all tasks equally, and in some tasks requiring expert knowledge, the crowd's wisdom is particularly unreliable (Krause et al., 2011). However, when average probability estimates made by small groups of judges are used to predict the veracity of *general* knowledge statements—such as 'which country has a larger area, New Zealand or the UK?'—these are typically more accurate than estimates made by each individual judge (Ariely et al., 2000).

There are a number of reasons to expect that unfamiliar face matching tasks might benefit from grouping of response data. For example, diversity in response patterns is necessary for crowd effects to occur, because this increases the likelihood that individual errors are uncorrelated (Hong & Page, 2004). We now know that there are large and stable individual differences on unfamiliar face matching tasks (Burton et al., 2010; Megreya & Burton, 2006). In addition, recent evidence suggests that unfamiliar face matching is a particularly noisy decision process, with poor levels of intrapersonal consistency (Bindemann, Avetisyan, & Rakow, 2012). These conditions

of between-subject diversity, coupled with the low baseline performance, might make face matching tasks fertile ground for crowd effects. On the other hand, it would appear that some individuals are particularly good at recognizing faces (Russell, Duchaine, & Nakayama, 2009), and recent research has identified individuals who are accurate specifically at matching unfamiliar faces (Burton et al., 2010; Megreya & Burton, 2006; Russell et al., 2009). Thus, an alternative solution would be to recruit expert populations for unfamiliar face matching tasks.

Here, we address this unresolved question by applying crowd analysis to a standard test of human face matching ability, the Glasgow Face Matching Test (GFMT; Burton et al., 2010). Our general method for measuring crowd effects is the same throughout this paper. First, we collect response data on the GFMT. Afterwards, we use a resampling technique to generate large numbers of groups for each level of crowd size, allowing us to estimate the accuracy of group populations and compare these to populations of individuals.

In Study 1, we resample existing normative data on the GFMT to investigate whether combining individual same/different responses across participants by majority vote rule improves face matching performance. Then, in Study 2, we replace binary decisions with similarity rating responses on the GFMT, enabling us to combine response data across subjects by averaging. In Study 3, we then test the success of this method under more challenging conditions, by measuring crowd performance on the short version of the GFMT using a web-based data collection procedure. Finally, in Study 4, we directly compare alternative response scales to determine the optimal method for matching faces by response aggregation.

STUDY 1

In this study, we test whether combining individual responses by majority vote improves accuracy on the GFMT. We also test whether accuracy of low-performing individuals is improved by response aggregation. If benefits are also observed in this group then we can conclude that improvement is not driven by the inclusion of high-performing participants.

Method

Participants

The data set used in this study consisted of itemized response data from 300 participants (180 female). The mean age of the group was 30.8 years, with a standard deviation (*SD*) of 14 (see Burton et al., 2010 for details).

Materials and procedure

The GFMT is a psychometric test designed to evaluate an individual's ability to match images of unfamiliar faces. It comprises 84 match and 84 mismatch image pairs, where match pairs show two images of the same person taken under similar lighting conditions, on the same day, but using different digital cameras. For mismatch pairs, one of these images is paired with a similar looking person from the database so that each identity appears once in a match pair and once in a mismatch pair. Examples are shown in Figure 1. To establish normative data on this test, all 168 pairs were presented in a different random order for each participant. On each trial, a face pair was presented centrally on the screen, and participants were required to indicate whether the two images were the same person or not, using a two-alternative forced-choice (2AFC) procedure. The task was self-paced, and on average, participants completed it in 15 minutes.

Figure 1. Example test pairs from the Glasgow Face Matching Test. Images on the top row are both of the same person, and images on the bottom row are different people

Results

Individual analysis

Mean accuracy on the GFMT is 89.9% ($SD = 7.3$), with performance ranging from 62% to 100% correct. Performance on match trials is slightly higher (92%) than on mismatch trials (88%). For detailed analysis of this data, see Burton et al. (2010).

Crowd analysis

To test for crowd effects in GFMT response data, we randomly generated 300 groups for each level of crowd size —2, 4, 8, 16, 32, and 64 subjects. For each group, we then calculated the proportion of 'same' responses made to each item, and then applied a majority vote decision rule which registered a 'same' crowd response when 50% or more subjects made 'same' responses, and a 'different' crowd response when more than 50% of subjects made 'different' responses (note that on 50–50 split decisions an arbitrary 'same' crowd response was recorded). Crowd accuracy was then calculated as the proportion of correct responses, separately for match and mismatch trials. The results of this analysis are shown in Figure 2.

To test the significance of crowd effects, we carried out *t*-tests between each successive increment of crowd size for match and mismatch trials separately (Bonferroni-corrected $p = 0.004$). Accuracy was significantly lower for groups of two participants than for individual participants for both match, $t\ (596) = 8.99$, $p < 0.05$, and mismatch trials, $t\ (596) = 8.98$, $p < 0.05$. However, after combining the responses of four individuals crowd performance exceeded individual performance for both match, $t\ (598) = 4.62$, $p < 0.05$,

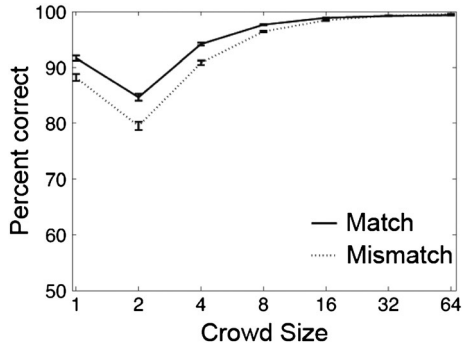

Figure 2. Accuracy for match and mismatch trials as a function of crowd size in the Study 1. Each data point represents average crowd performance for 300 randomly sampled groups. Error bars denote standard error of the mean

and mismatch trials, t (598) = 3.57, $p < 0.05$. Cumulative improvements in matching accuracy for match and mismatch trials were found between all remaining increments in group size ($p < 0.05$), and overall accuracy reached 99.5% for groups of 64. This is a large improvement over individual performance on this test: Of the 300 participants tested in the normative data collection, only two outperformed the crowd (i.e., scoring 100%).

Next, we tested for crowd effects in groups of individuals who performed poorly on the task. First, we selected individuals who scored lower than one *SD* below the mean in overall accuracy (<83%). This resulted in a subset of 47 subjects, from which we repeated the sampling procedure, randomly creating 300 groups per level of crowd size. Crowd performance from low-performing individuals is shown in Figure 3.

Individual accuracy in our low performance group was superior to crowd accuracy for crowd size of two both in match, t (345) = 6.47, $p < 0.05$, and mismatch trials, t (345) = 11.3, $p < 0.05$. Individual accuracy was also better than crowd accuracy for crowd size of four in mismatch trials, t (345) = 3.43, $p < 0.05$, and for match trials, there was no difference in performance, t (345) = 1.22, $p > 0.05$. For crowd sizes of eight participants, we found that crowd performance surpassed individual performance in both match, t (345) = 11.7, $p < 0.05$, and mismatch trials, t (345) = 5.89, $p < 0.05$. Cumulative improvements in performance were observed for each subsequent doubling of group size ($p < 0.05$ for each comparison after Bonferroni-correction). Moreover, average crowd

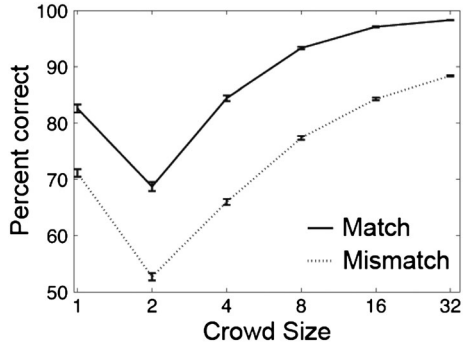

Figure 3. Accuracy data in match and mismatch trials for crowds sampled from poor performers on the Glasgow Face Matching Test (<1 standard deviation below mean accuracy). Error bars denote standard error

performance in groups of eight or more (M = 85.4; SD = 2.46) was better than the best performing individual in this group of 47 subjects (max = 82.1%).

Discussion

Our initial results show a large benefit of combining face matching responses across individuals. Mean crowd performance exceeded individual accuracy for groups of four individuals and higher, suggesting that face matching is improved by aggregating the responses of small groups of people. Remarkably, for individuals who performed poorly on the task, we found that mean crowd performance exceeded the best performing individual for groups of eight and larger. Furthermore, when groups were sampled from the entire performance range, near-perfect performance (99.2%) was achieved by aggregating the responses of 32 participants.

Although encouraging, the results of Study 1 also raise an important limitation of the chosen aggregation method. The majority rule decision criteria caused pairs of participants to perform worse than individual participants. This is presumably because combining binary responses often produces an ambiguous 50–50 vote in group sizes with an even number of members. In this case, an arbitrary 'same' crowd response was registered, explaining why there was a larger bias toward 'same' responses in crowd, relative to individual responses (Figures 2 and 3). In order to avoid this situation, in the next study, we replaced same/different decisions with a rating scale, which enabled individual responses to be combined by averaging.

STUDY 2

Rather than making explicit identity judgments, participants in this study instead rated the similarity of face pairs. This approach has previously been used to compare human and computer performance on face matching tasks (e.g., O'Toole, An, Dunlop, & Natu, 2012; O'Toole, Phillips, et al., 2007), when automatic face recognition software returns a continuous match score variable. Although previous face matching studies have typically asked subjects to rate the likelihood that two images are of the same person (e.g., O'Toole, Phillips, et al., 2007), we simply ask the participants here to rate image similarity, without encouraging explicit identity decisions.

Collecting similarity judgments in this way allows representational and decisional components to be separated, which may be beneficial in certain situations. By using human viewers to generate similarity scores, this enables a system administrator to then control decisional criterion by manipulating gain according to the risk associated with specific types of error. For example, in situations where it is particularly important to avoid 'miss' decisions, the threshold for 'same' responses could be set lower than if the priority is to avoid 'false alarms' (for a detailed discussion of the separation between representational and decisional components of perceptual processes see Macmillan & Creelman, 2004).

Because rating data were used in the current study, we calculated both individual and group performance by first calculating hit and false alarm rates for each similarity threshold and plotting these to produce the receiver

operating characteristic (ROC). We then calculated the area under ROC curve (AUC), which was used as our dependent variable. This measure is widely used to assess performance of classification rules (Krzanowski & Hand, 2009) and has a variety of applications. For example, it is often used to test the diagnostic value of symptoms in predicting the presence or absence of medical conditions (e.g., Pepe, 2003). Here, we use AUC as a measure of the extent to which ratings of similarity discriminate between match and mismatch test items of the GFMT. We predict that aggregating similarity ratings across groups of participants will strengthen the crowd effects observed in the previous study.

Method

Subjects

Thirty students from the University of Glasgow (23 female) participated in the experiment and received either course credit or cash payment. Participants were aged between 17 and 29 years ($M = 20.4$; $SD = 3.1$).

Materials and procedure

As with Study 1, participants were tested on the long version of the GFMT (Burton et al., 2010). However, instead of asking participants to make a 2AFC same/different decision, we instead asked participants to rate the similarity of the two images using a Likert scale (from 1 to 7). Image pairs were presented sequentially on a computer monitor and participants made ratings while both images remained on-screen. The task was self-paced and took an average of 15 minutes to complete.

Results

Individual data

First, we calculated the individual classification performance by generating ROC curves for each of the 30 participants. ROC curves for both individuals and 'crowds' were calculated using hit and false alarm rates at each point in our rating scale (1 through 7). Mean classification accuracy, as measured by the area under these ROC curves (AUC), was 0.907 ($SD = 0.06$; max = 0.986; min = 0.781), which is comparable with normative same-different accuracy data on this test ($M = 89.9\%$; Burton et al., 2010).

Crowd analysis

To provide an accurate estimate of the rate at which performance improves as a function of group size, for each group size, we randomly generated 435 groups (this is the number of unique permutations of pairs in 30 instances). For each group, we then calculated the mean similarity score for each test item and repeated this across all possible permutations of group. This method produced an array of column vectors for each level of group, and we used these to generate a ROC curve individually for each group, by measuring hit and false alarm data at the same seven thresholds used to calculate individual ROCs.

Histograms of group AUC scores for each level of crowd size are shown in Figure 4. It is apparent from these data that averaging similarity ratings across small groups of participants had a large effect on classification accuracy. In the following, all tests of significance are adjusted for multiple comparisons using Bonferroni-correction ($p = 0.012$). Crowd performance exceeded individual performance for crowd size of two ($M = 0.949$; $SD = 0.032$), $t\,(463) = 8.19$, $p < 0.05$. Further crowd improvements were observed between crowd sizes of two and four, $t\,(868) = 15.3$, $p < 0.05$, four and eight, $t\,(868) = 15.5$, $p < 0.05$, and between groups of eight and 16, $t\,(868) = 15.5$, $p < 0.05$. Moreover, for group sizes of seven and above, mean group accuracy ($M = 0.988$; $SD = 0.006$) exceeded that of the best performing individual ($M = 0.986$; $SD = 0.06$). This accuracy score also represents a large improvement in performance relative to normative GFMT data ($M = 89.9\%$).

Discussion

In Study 2, we observed large benefits of averaging similarity ratings of small groups of independent raters. The rate of improvement observed is quite surprising, with near-perfect performance being observed after combining responses of just eight subjects. Indeed, substantial improvements equating to around 5% in overall accuracy are observed after aggregating the responses of just two people. This represents a considerable improvement over Study 1, where aggregating 2AFC responses of two participants produced *poorer* accuracy than individual responses. It is also apparent that improvements in mean accuracy were accompanied by a reduction of variance. Therefore, response averaging not

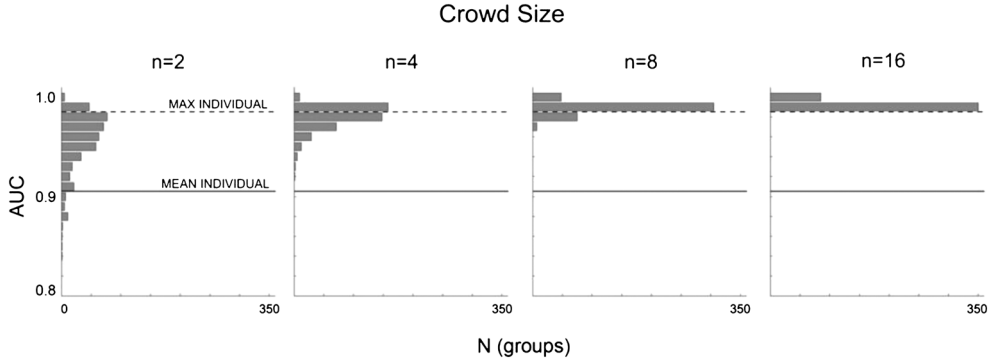

Figure 4. Frequency histograms showing the distribution of area under ROC curve (AUC) values for small group sizes. Group AUC values were calculated separately for all possible combinations of participants in Study 2 at four levels of group size. Mean (solid line) and maximum (dashed line) individual AUC scores are shown for comparison

only provides gains in absolute performance—it also provides stability, which safeguards face matching processes from fluctuations in accuracy caused by individual variation.

We note however that performance gains for larger groups were somewhat muted in comparison with the large improvement that was seen in smaller groups. There are two possible accounts for this leveling-out of group performance. One possibility is that average ratings become stable after averaging together small arrays of match score data, thereby attenuating the contribution of subsequent recruits. On the other hand, the upper limit might be ceiling effects imposed by the fact that group performance is already close to the maximum score attainable in the test. In the next experiment, we investigated this by setting participants a more difficult task.

STUDY 3

In this study, we aim to define the upper limit of crowd effects in unfamiliar face matching tasks, and so, we set participants the short version of the GFMT, which comprises the most difficult items from the longer version of the test (mean normative performance on this test is 81.3% with an *SD* of 9.7, Burton et al., 2010). In addition, we 'crowdsource' similarity ratings via on-line data collection, to simulate one possible method for improving matching accuracy in applied settings. We also measure crowd effects in conditions where face processing is known to be compromised—by presenting test items upright, inverted (i.e., upside down), and negated (i.e., in photographic negative). This allows us to establish the extent to which response aggregation also benefits matching tasks with stimuli that people are not experienced at identifying, allowing us to determine whether the benefit of response averaging might generalize to other pattern matching tasks.

Method

Subjects

Eighty-seven US residents volunteered for this study via the crowdsourcing website Amazon Mechanical Turk (mturk.com). Because of our web-based data collection method, inclusion criteria were set to ensure participants were paying sufficient attention to the task. We tested participants' focus by including three 'catch' trials displaying identical images: if participants' average similarity rating to these catch trials was below 7 (on a scale from 1 to 7), then they were excluded prior to analysis. Twenty-five participants failed to meet the inclusion criteria, leaving a final sample of 62 participants (41 female) with an average age of 36 years ($SD = 12.9$).

Materials and procedure

In this study, participants completed the short version of the GFMT. This version comprises the 20 most difficult match and 20 most difficult mismatch trials from the long version of the test. Each participant was presented with each face pair under three different stimulus conditions (Figure 5), giving a total of 120 trials. The order of presentation was fully randomized, and on each trial, participants were asked to rate the similarity of the images on a Likert scale (1 to 7). On average, participants took 28 minutes to complete the task ($SD = 15$ minutes).

Results

Individual analysis

Mean AUC score for upright image pairs was substantially lower than published normative data for the GFMT (70% compared with 81%). Average individual performance scores were superior for upright ($M = 0.696$; $SD = 0.113$) relative to both inverted ($M = 0.636$; $SD = 0.094$), $t\ (61) = 4.42$, $p < 0.05$, and negated ($M = 0.594$; $SD = 0.098$) image pairs, $t\ (61) = 6.64$, $p < 0.05$. Inverted images produced better performance than did negated images, $t\ (61) = 3.26$, $p < 0.05$ (Bonferroni-corrected $p = 0.017$).

In addition, mean by-item ratings in the three conditions were strongly correlated (upright and inverted, *Pearson's* $r = 0.875$; upright and negated, $r = 0.814$; inverted and negated, $r = 0.788$), and subject AUC data were also correlated across conditions (upright and inverted, $r = 0.475$; upright and negated, $r = 0.348$; inverted and negated, $r = 0.439$). This result is consistent with previous research showing that performance on upright and inverted unfamiliar face matching is highly correlated and supports the contention that similar pattern matching processes are involved when processing unfamiliar faces in each of these conditions (Megreya & Burton, 2006).

Crowd analysis

As in Study 2, we calculated average similarity score column vectors for multiple permutations of group at each level of group size. Given the large sample, we limited the computation to the first 1000 randomly sampled group permutations. From each set of 1000 group ratings, we then calculated individual AUC scores using the same method as in Studies 1 and 2. Average AUC scores for each level of crowd size are shown in Figure 6. As can be seen from this figure, crowd benefits occurred in all stimulus conditions. However, the largest crowd effect observed in the upright image condition, where performance improved by around 20% when comparing individual discrimination accuracy with groups of 20 or larger. This also represents a 10% improvement on GFMT normative data. Crowd effects were not as pronounced for

Figure 5. Example stimuli from Study 3. The Glasgow Face Matching Test was presented upright (a) and also under more challenging conditions where the images were inverted (b) or negated (c)

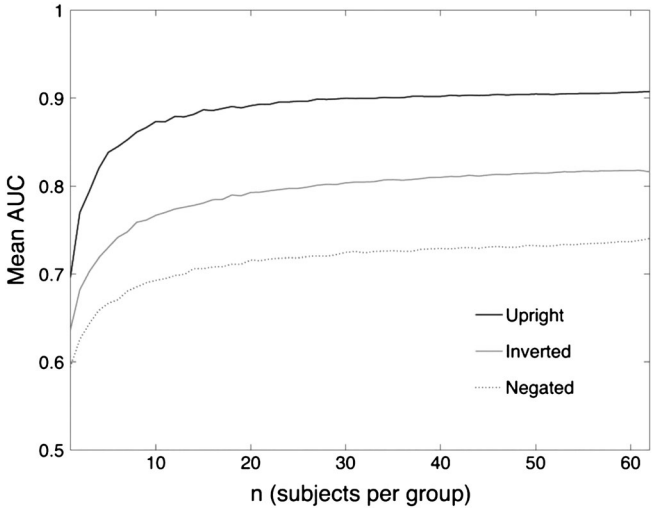

Figure 6. Results of the crowd analysis in Study 3

either negated (around 10%) or for inverted stimuli (around 15%), suggesting that crowd effects are constrained by the difficulty of the comparison task.

For upright faces, crowds of two people ($M = 0.769$; $SD = 0.093$) produced superior performance to individuals ($M = 0.696$; $SD = 0.113$), $t\ (1060) = 6.21$, $p < 0.05$. Large improvements in performance were also observed between crowd sizes of two ($M = 0.696$; $SD = 0.113$) and four ($M = 0.821$; $SD = 0.071$), $t\ (1998) = 14.3$, $p < 0.05$, and between four and eight, ($M = 0.860$; $SD = 0.071$), $t\ (1998) = 14.26$, $p < 0.05$. All significance tests were adjusted for multiple comparisons (Bonferroni-corrected $p = 0.0036$). Although performance did continue to increase beyond groups of 10, it is apparent from Figure 6 that there was a leveling-out in cumulative improvement, such that the benefit of additional group members diminished as crowd size increased. For inverted and negated stimuli, we observed similar crowd effects in both conditions.

Discussion

The results of this web-based study again show a large crowd advantage for unfamiliar face matching. By combining the responses of just two people, we were able to improve face matching performance by 7% relative to individual AUC scores and for crowd sizes of four and above performance surpassed GFMT normative accuracy scores. Crowd effects were also observed for inverted and negated stimuli; however, these improvements were not as large as improvements on upright matching. This result shows that the crowd effects we have observed here do not occur equally for all types of image comparison and may therefore be limited by the expertise that people have with the stimuli that they are comparing.

These data also show that the benefit of response aggregation does not continue to increase monotonically with respect to group size. Here, an upper limit of crowd accuracy was established after aggregating the responses of around 10 raters, and this pattern appeared to be consistent across experimental conditions. Because the level of performance was substantially lower than 100%, we can conclude that this limit is not imposed by the upper limit of our performance scale. Instead, this limit is likely to be caused by average ratings stabilizing in groups of 10 or more. Therefore, it appears that while performance on unfamiliar face matching tasks is improved by combining responses of nonexpert populations, there is little benefit to combining responses of larger groups.

Having established an upper limit, one might ask why perfect crowd performance is not observed. We note that the upper limit observed here is established because, on average, some mismatch pairs are consistently perceived as being more similar than some match pairs, across subjects. We submit that this situation is likely to occur on any unfamiliar face matching task and with any set of participants, because variability caused by changes in photographs can outweigh variability caused by changes in the *person* (Jenkins & Burton, 2011; Jenkins, White, Van Montfort, & Burton, 2011). Nevertheless, averaging responses across participants has provided some improvement on normative data, despite lower individual accuracy rates. Clearly, however, error rates of 10% represent an unacceptably high risk in many applied settings. Therefore, in the next study, we ask whether a more suitable method for collecting similarity ratings might produce more robust crowd performance.

STUDY 4

In Study 3, we found that averaging similarity ratings provided an improvement on the short version of the GFMT but that individual accuracy on the task was lower than published normative data. One possibility is that ratings of image similarity do not adequately discriminate between same and different identities. For example, while the two images in Figure 1 (top row) might vary in a range of superficial ways (e.g., lighting conditions), it might nevertheless be clear to the majority of viewers that these images are of the same person. Therefore, in this study, we asked whether the performance reported in the previous experiment might be improved by replacing the similarity rating procedure (how similar are these images?) with a response that requires explicit identity judgments (how likely are these images to be of the same person?).

Method

Subjects

Sixty-nine US residents volunteered for this study via the crowdsourcing website Amazon Mechanical Turk (mturk.com). Inclusion criteria were set as in the previous study, and four participants were excluded on this basis. In addition, we excluded five participants whose performance did not differ significantly from chance (0.5). The final sample consisted of 60 participants (25 female) with an average age of 34 years ($SD = 12.9$).

Materials and procedure

The materials and method were identical to the previous experiment; however, we manipulated here the type of scale used to rate face pairs. For the similarity rating group ($n = 30$), the procedure was identical to the previous experiment (1 = very disimilar; 7 = very similar). For the identity

rating group ($n = 30$), we replaced similarity ratings with a rating of the 'likelihood that the two images are of the same person' (1 = sure different; 7 = sure same).

Results

Individual analysis

For the similarity rating group, planned comparisons between individual AUC scores on the different stimulus conditions show a similar pattern to the previous study, with performance on upright pairs being superior to performance on both inverted, $t\ (58) = 6.19$, $p < 0.05$, and negated pairs, $t\ (58) = 5.59$, $p < 0.05$. However, there was no difference between performance on inverted and negated conditions ($t < 1$). Likewise, for the identity rating group, upright was superior to inverted, $t\ (58) = 9.71$, $p < 0.05$, and negated, $t\ (58) = 10.9$, $p < 0.05$, with no significant difference between upright and negated pairs ($t < 1$).

More importantly for the current study, for upright pairs, participants in the identity rating condition ($M = 0.809$; $SD = 0.101$) produced superior performance to the similarity rating condition ($M = 0.724$; $SD = 0.105$), $t\ (58) = 3.17$, $p < 0.05$. However, the differences between identity and similarity groups were not significant for either inverted, $t\ (58) = 1.24$, $p > 0.05$, nor negated stimuli, $t\ (58) = 1.78$, $p > 0.05$, suggesting that information which was diagnostic of identity became less accessible in these conditions.

Crowd analysis

We treated response data in the same manner as previous studies. For each experimental condition, we randomly sampled 435 permutations of group and calculated individual AUC scores using the same method as before. The average AUC scores for both experimental conditions, at each level of crowd size, are shown in Figure 7. It is clear from these plots that replacing similarity ratings with identity-based ratings provided a substantial boost in performance and that strong crowd effects were observed in both conditions.

For the similarity rating group, crowds of two raters ($M = 0.778$; $SD = 0.087$) produced superior performance to individuals ($M = 0.724$; $SD = 0.105$), $t\ (463) = 3.22$, $p < 0.05$. Large improvements in performance were also observed between crowd sizes of two and four ($M = 0.843$; $SD = 0.063$), $t\ (436) = 12.5$, $p < 0.05$, and between four and eight ($M = 0.892$; $SD = 0.04$), $t\ (463) = 13.7$, $p < 0.05$. For the identity rating group, crowds of two raters ($M = 0.881$; $SD = 0.074$) produced superior performance to individuals ($M = 0.809$; $SD = 0.101$), $t\ (463) = 4.98$, $p < 0.05$. Large improvements in performance were also observed between crowd sizes of two and four ($M = 0.941$; $SD = 0.04$), $t\ (436) = 14.7$, $p < 0.05$, and between four and eight ($M = 0.971$; $SD = 0.020$), $t\ (463) = 14.2$, $p < 0.05$. For inverted and negated stimuli, we observed similar crowd effects in both conditions.

Given the differences between individual performance in similarity and identity rating groups, it is perhaps unsurprising that the identity group also outperformed the similarity group at crowd sizes of two, $t\ (463) = 18.6$, $p < 0.05$, four, $t\ (463) = 27.1$, $p < 0.05$, and eight, $t\ (463) = 37.0$, $p < 0.05$. Peak mean AUC scores for both conditions were found at crowd sizes of 29, for which the identity rating condition ($M = 0.985$; $SD = 0.002$) was significantly higher than the similarity rating condition ($M = 0.917$; $SD = 0.005$), $t\ (463) = 276$, $p < 0.05$.

Discussion

The results of this study demonstrate a substantial enhancement of both individual performance and crowd performance by replacing an image similarity rating-scale with a scale that explicitly requires participants to make an identity judgment. Moreover, crowd performance on the short version of the GFMT approached perfect levels of accuracy, representing a 20% improvement on normative individual data (Burton et al., 2010).

GENERAL DISCUSSION

Overall, our results replicate previous reports of poor face matching performance in individual participants. Average AUC scores in the GFMT long version was similar to published normative data of 89.9% correct (Study 2;

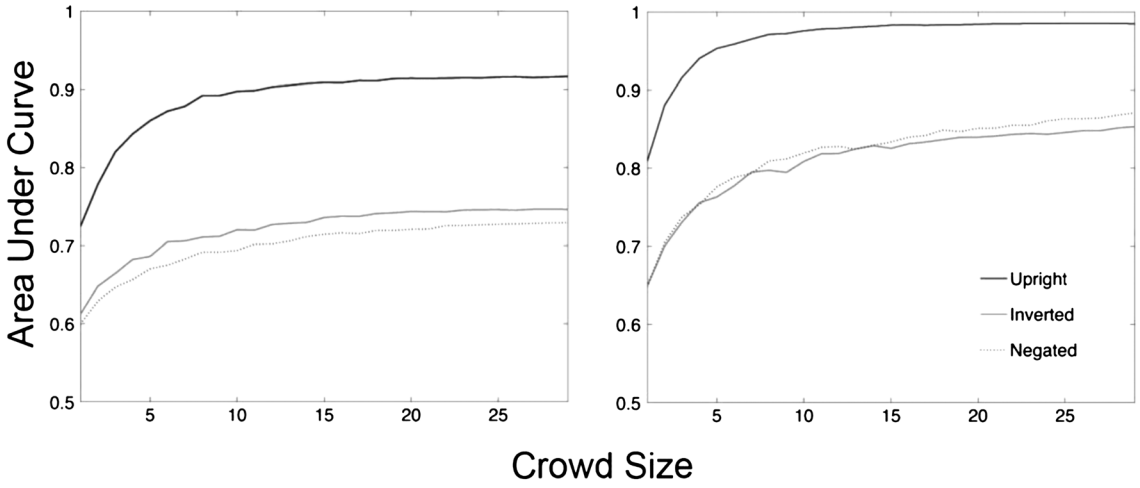

Figure 7. Crowd effects in Study 4. Participants were either instructed to rate the similarity of images (left panel) or to rate the likelihood that the two images were of the same person (right panel)

AUC = 0.90). Although AUC scores on the short form version of this test were below the level of established norms of 81% correct (Study 3; AUC = 0.70), these were equivalent to normative scores when identity judgments were made using a rating scale response (Study 4; AUC = 0.81).

More importantly, our results show that poor individual performance can be improved substantially by aggregating response data across small groups of people. In Study 2, we found that averaging similarity ratings made to pairs of face images produced a large benefit to same-different item discrimination. Furthermore, this method produced ceiling performance remarkably quickly, and crowd performance exceeded that of the best performing individual in groups of eight and larger. The results of Study 3 also revealed an effective upper limit of performance that was substantially below 100%, which was established in small group sizes, with only slight improvements observed for groups of 10 and above.

In Study 4, we demonstrated that this upper limit was caused partly by the rating-scale used to collect similarity scores. When we asked participants to make identity judgments using a rating scale (i.e., how likely are these images to be of the same person?) we found that averaging these responses produced near-perfect crowd performance. The results of Study 4 appear to show a marked distinction between processes involved in rating image similarity and rating similarity of identities, with identity ratings producing superior individual and group performance. This result is practically significant because it suggests that task instructions have a large impact on matching accuracy when using a rating scale procedure. In addition, we propose that this may tell us something important about the task of unfamiliar face matching more generally. Although unfamiliar face matching might have more in common with simple image-matching than face recognition processes, our expertise with faces does appear to support unfamiliar face matching performance, because presenting comparison images upside down impedes matching accuracy (Megreya & Burton, 2006). The data reported here are consistent with this account, showing that participants are able to focus ratings of similarity toward features of the face that are diagnostic of identity.

Overall, we believe the crowd effects reported here are of considerable practical significance, suggesting that the accuracy of real world face matching tasks could be improved by averaging the responses of small groups of people. For example, when difficult unfamiliar face matching decisions are encountered, these might be 'crowdsourced' using a similar technique to that described in Studies 3 and 4. Further, in occupational settings where face matching decisions are made on computers by multiple users (e.g., Graves et al., 2011; White, Tan, et al., 2013), network systems that aggregate responses across individuals are a realistic goal (see also Dror, Wertheim, Fraser-Mackenzie, & Walajtys, 2012).

We have previously suggested that the GFMT might be used as a recruitment tool for identifying people that are particularly good at unfamiliar face matching (Burton et al., 2010; see also Russell et al., 2009). However, the results of this series of studies suggest that a more valuable approach might be to combine identity judgments across populations, rather than seeking the expert opinion of high performers. In future research, it will be important to clarify the relationship between individual differences of group members and crowd accuracy. We suspect that wisdom-of-crowds is the greatest when individuals making up the set use a diverse range of cognitive strategies to perform a task (Hong & Page, 2004; O'Toole, Abdi, Jiang, & Phillips, 2007). Accordingly, it might be of interest to investigate which group combinations produce the largest crowd benefits and to devise methods for encouraging diversity in strategy on face matching tasks. In addition to the apparent practical benefits, such research might help to specify cognitive strategies used to identify faces from photographs.

For similar reasons, crowd effects might also benefit other areas of forensic image comparison. One recent paper demonstrates a very large degree of both intra-expert and inter-expert variability in fingerprint analysis (Dror et al., 2011). The authors found that judgments of expert fingerprint analysts were highly varied, with counts of minutiae on a latent fingerprint ranging from 9 to 30. One way to tackle such inconsistencies might be to combine expert judgments by averaging, in order to enable more reliable decisions. Further, recent research has shown that the accuracy of police identity line-ups can be improved by collecting confidence judgments from witnesses (Brewer, Weber, Wootton, & Lindsay, 2012). Our results suggest that in cases where more than one person witnessed a crime, identification accuracy might be improved by computing the average of witnesses' confidence scores.

In summary, we have shown that averaging responses from small groups of untrained and nonexpert individuals drastically improves performance on face matching tasks. We hope that this result might help to improve the reliability of identity verification processes in a variety of occupational settings. For example, the results reported here might be of interest to legal practitioners, in courtroom settings where defendants are identified using closed-circuit television evidence. In such cases, where there is sufficient detail in images to make identity judgments, the veracity of identity verification procedures might be improved by pooling jurors' identity estimates, as opposed to relying on estimates provided by lone experts.

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 Jana Bördgen for collecting data for Study 2.

REFERENCES

Ariely, D., Wing-Tung, A., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., … Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*, 130–147.

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*(3), 277–291.

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science*, *23*(10), 1208–1214.

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.

Burton, A. M., & Jenkins, R. (2011). Unfamiliar face perception. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *The Handbook of Face Perception*. Oxford, UK: Oxford University Press.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291.

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology*, *17*(1), 97–116.

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison. *Forensic Science International*, *208*(1–3), 10–17.

Dror, I. E., Wertheim, K., Fraser-Mackenzie, P., & Walajtys, J. (2012). The impact of human-technology cooperation and distributed cognition in forensic science: Biasing effects of AFIS contextual information on human experts. *Journal of Forensic Sciences*, *57*(2), 343–352.

Galton, F. (1907). Vox Populi. *Nature*, *75*, 450–451.

Graves, I., Butavicius, M., MacLeod, V., Heyer, R., Parsons, K., Kuester, N., … Johnson, R. (2011). The role of the human operator in image-based airport security technologies. *Studies in Computational Intelligence*, *338*, 147–181.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385–16389.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Proceedings of the Royal Society B*, *366*, 1671–1683.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323.

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

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*(1), 623–655.

Krause, S., James, R., Faria, J. J., Ruxton, G. D., & Krause, J. (2011). Swarm intelligence in humans: Diversity can trump ability. *Animal Behaviour*, *81*(5), 941–948.

Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. London: Chapman & Hall.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. New York, US: Psychology Press.

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

O'Toole, A. J., An, X., Dunlop, J., & Natu, V. (2012). Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception (TAP)*, *9*(4), 16.

O'Toole, A., Abdi, H., Jiang, F., & Phillips, P. J. (2007). Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics*, *37*(5), 1149–1155.

O'Toole, A. J., Phillips, P. J., Jiang, F., Ayyad, J., Penard, N., & Abdi, H. (2007). Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(9), 1642–1646.

Pepe, M. S. (2003). *The statistical evaluation of medical tests of classification and prediction*. Oxford: University Press.

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D., … Weimer, S. (2012). The good, the bad, and the ugly face challenge problem. *Image and Vision Computing*, *30*(3), 177–185.

Phillips, P. J., Flynn, P. J., Bowyer, K. W., Bruegge, R. W. V., Grother, P. J., Quinn, G. W., & Pruitt, M. (2011). Distinguishing identical twins by face recognition. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 185–192.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin and Review*, *16*(2), 252–257.

Surowiecki, J. (2004). *The wisdom of crowds*. US: Random House.

White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (in press). Redesigning photo-ID to improve unfamiliar face matching. *Journal of Experimental Psychology: Applied.*

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

White, D., Tan, M., & Kemp, R. I. (2013). Verifying passport photographs using computer-assisted face recognition systems. Manuscript submitted for publication.

Woodhead, M. M., Baddeley, A. D., & Simmonds, D. C. V. (1979). On training people to recognize faces. *Ergonomics*, *22*(3), 333–343.