Performance of Typical and Superior Face Recognisers on a Novel Interactive Face Matching Procedure

Harriet M. J. Smith[1], Sally Andrews[1], Thom S. Baguley[1], Melissa F. Colloff[2], Josh P. Davis[3], David White[4], & Heather D. Flowe[2]

[1]Department of Psychology, Nottingham Trent University

[2] School of Psychology, University of Birmingham

[3]School of Human Sciences, Institute of Lifecourse Development, University of Greenwich

[4]School of Psychology, UNSW Sydney

Author Note

Harriet M. J. Smith https://orcid.org/0000-0003-2712-5527

Correspondence concerning this article should be addressed to Harriet M. J. Smith, Department of Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham, NG1 4FQ. Telephone number: +44 (0) 115 8484535. Email: harriet.smith02@ntu.ac.uk

Data is available from the authors upon reasonable requests

Abstract

Unfamiliar simultaneous face matching is error prone. Reducing incorrect identification decisions will positively benefit forensic and security contexts. The absence of view-independent information in static images likely contributes to the difficulty of unfamiliar face matching. We tested whether a novel interactive viewing procedure that provides the user with 3D structural information as they rotate a facial image to different orientations would improve face matching accuracy. We tested the performance of 'typical' (Experiment 1) and 'superior' (Experiment 2) face recognisers, comparing their performance using high quality (Experiment 3) and pixelated (Experiment 4) Facebook profile images. In each trial, participants responded whether two images featured the same person with one of these images being either a static face, a video providing orientation information, or an interactive image. Taken together, the results show that fluid orientation information and interactivity prompt shifts in criterion and support matching performance. Because typical and superior face recognisers both benefited from the structural information provided by the novel viewing procedures, our results point to qualitatively similar reliance on pictorial encoding in these groups. This also suggests that interactive viewing tools can be valuable in assisting face matching in high performing practitioner groups.

Keywords: unfamiliar face matching, super-recognisers, individual differences, interactive procedure, face identification

Acknowledgements: We are grateful to Thomas Goodge (Nottingham Trent University) for his help with stimulus development, to Monika Durova, Nikolay Petrov and Daisy Thomas (University of Greenwich) for assisting with data collection, and to Nikolas Pautz (Nottingham Trent University) for creating the figures. We would also like to thank the participants of the University of Greenwich Super Recogniser database for making this research possible.

Performance of Typical and Superior Face Recognisers on a Novel Interactive Face Matching Procedure

Photo-ID is necessary for identity verification in a range of settings, from crossing borders to buying age-restricted goods and accessing services. However, despite the wide use of photo-ID, simultaneous face-matching is alarmingly error-prone. In controlled laboratory and field studies, error rates between 10-30% are typically observed (Bruce et al., 1999; Megreya & Burton, 2006, 2008). False alarms (i.e., incorrectly assigning two people to the same identity) are the most common type of error, with failure rates of 40% to 60% in field tests (Davis & Valentine, 2009; Kemp, Towell, & Pike, 1997). Error rates are high even among passport-issuing officers and do not decrease with additional years of experience or professional training (White, Kemp, Jenkins, Matheson, & Burton, 2014). Investigating ways of reducing errors can have significant benefits to the accuracy of applied tasks such as security screening and police investigations. In this paper we evaluate methods for improving simultaneous one-to-one face-matching that enable viewers to make best use of variations in viewpoint information available to them. We present a novel procedure in which the comparison image can be maneuvered into different orientations at the discretion of the viewer.

Differences in Viewpoint

In face matching, it might be necessary to try and reconcile two images of faces that vary in terms of orientation. If the face is unfamiliar, the viewer will have no knowledge of the person's 3D facial structure. In face memory tasks, differences in orientation between study and test undermine recognition accuracy (Bruce, 1982; Colloff, Seale-Carlisle et al., 2020), but knowledge of 3D facial structure mitigates the effect of viewpoint dependence (Hill, Schyns, & Akamatsu, 1997; Longmore, Liu, & Young, 2008). In face matching, where there is limited memory load, some evidence suggests that performance suffers less across differences in viewpoint than it does in face memory tasks (Estudillo & Bindemann, 2014). However, even relatively minor differences in viewpoint can create problems for unfamiliar face matching (Bruce et al. 1999; Hancock, Bruce, & Burton, 2000), particularly in more difficult matching tasks (Bruce et al., 1999).

Based on these results, it might be expected that providing participants with both frontal and profile facial views would improve matching performance. Surprisingly though, Kramer and Reynolds (2018) found that accuracy did not differ across three conditions in which participants matched two pairs of frontal images, two pairs of profile images, or two pairs of images featuring one frontal and one profile view. The authors explain the lack of benefit in the latter condition by proposing that there may have been no mental integration of the frontal and profile views. Put differently, the participants did not use the two images to build a 3D view-independent representation of the face, perhaps because the orientations were too disparate. However, as people are able to extract information across multiple frontal images of the same face to support the construction of stable representations (Menon, Kemp, & White, 2018; Menon, White, & Kemp, 2015; White et al., 2014), showing a face moving fluidly from side to side may facilitate the building of a view-independent representation.

The Benefit of Movement

The results of various studies attest to the benefit of fluid movement for face perception. Pike, Kemp, Towell and Phillips (1997) found that rigid head rotations improved recognition performance, arguing that such movement provides 3D structural information. However, there are various ways in which a face can move, and much of the literature has focused on non-rigid movement, such as smiling, frowning or speaking as cues to identity (e.g. Knappmeyer, Thornton, & Bülthoff, 2003; Pilz, Thornton, & Bülthoff, 2006; Smith, Dunn, Baguley, & Stacey, 2016). For example, effects of movement observed when recognizing familiar faces (Lander, Christie & Bruce, 1999; Lander & Bruce, 2000) are explained in terms of the ability to access idiosyncratic non-rigid movement stored in memory (Lander & Chuang, 2003). While Thornton and Kourtzi (2002) tested the effect of non-rigid changes in expression on unfamiliar sequential face matching, we are not aware of any studies that have explored the effect of rigid rotation movement in the context of simultaneous face matching. Given the disruptive effect of viewpoint dependence (Bruce et al., 1999), this is an important question.

The Interactive Procedure

Standard face matching tasks in operational contexts involve passive mental comparisons rather than active engagement with images. A procedure in which users can interact with one face in a pair to be matched; maneuvering it fluidly to different viewpoints along a vertical axis, may support matching performance. The education literature is replete with examples of task engagement improving learning outcomes (Freeman et al., 2014). This can be explained by increased attentiveness and depth of encoding (Craik & Lockhart, 1972; Craik, 2002), which are also beneficial to face processing (Bower & Karlin, 1974; Liu, Ward, & Markall, 2007; see also Palermo & Rhodes, 2007). Interactivity should increase the depth of encoding, and rotation should facilitate the building of a 3D view-independent representation, providing structural information that is unavailable in static snapshots. This may enable operators to familiarize themselves with the face, and to gain knowledge of how invariant features range in appearance across different viewpoints of the face (Lander, Christie, & Bruce, 1999; Pike, Kemp, Towell, & Phillips, 1997). Such a procedure will not only confer some of the benefits of familiar face processing but will also enable the operator to maneuver comparison faces into the same viewpoint, reducing within-person variability across images. The procedure has been successfully employed in face memory tasks, with higher discrimination accuracy observed in an interactive lineup compared to a lineup composed of static images of faces, which is commonly used by US police (Colloff, Flowe et al., 2020; Colloff, Seale-Carlisle et al., 2020).

Typical and Super-Recognisers

In recent years, there has been an increasing focus on improving face identification accuracy in applied settings. Given the wide range of individual differences in unfamiliar face perception and recognition ability in both novices (for reviews, see Noyes, Phillips, & O'Toole, 2017; Lander, Bruce, & Bindemann, 2018) and practitioners ([BLINDED]), one of the most promising solutions is to select individuals on the basis of ability. There is a practical need to test the benefits of novel procedures in both typical and "super-recognisers", who are likely to use these solutions in professional settings (e.g., Davis, Lander, Evans, & Jansari, 2016; Davis, Maigut, & Forrest, 2019; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016).

There are also theoretically important reasons to establish whether there are qualitative differences in face processing between typical and super-recognisers. Research focusing on the other-ethnicity bias provides evidence that typical- and super-recogniser performance does not differ in a qualitative way, and both groups are subject to the same influences. For example, recognition memory in both groups was better for own- than other-ethnicity faces (Bate et al., 2019; Robertson, Black, Chamberlain, Megreya, & Davis, 2019). Other studies have observed a heightened inversion effect in super-recognisers (Russell, Duchaine, & Nakayama, 2009), and it has been suggested that they rely more on holistic processing (Bobak, Bennetts, Parris, Jansari, & Bate, 2016). There is however inconsistent support for this conclusion, with some super-recognisers exhibiting enhanced holistic processing and others exhibiting the opposite pattern of performance (Belanova, Davis, & Thompson, 2019). Differences in structural encoding provide an alternative explanation for differences in face processing ability. In Bobak, Hancock and Bate's (2016) one-to-many face matching study, target and array photographs varied according to viewpoint. Super-recognisers were more accurate than controls. One possible explanation provided is that super-recognisers are better at structural encoding strategies that help them to construct a view-independent representation. In contrast, controls may rely more on less helpful pictorial encoding strategies.

The existing literature provides only mixed evidence that typical recognisers and super-recognisers process faces in a qualitatively different way (see Noyes et al., 2017). Testing both types of recogniser using the interactive system speaks directly to this question. As yet, the hypothesis that super-recognisers are better at structural encoding has not been fully tested. However, if the hypothesis is supported, the ease with which super-recognisers extract structural information from static images would likely limit the usefulness of additional orientation information provided by fluid rotation. It might also mean that interactivity does not improve performance for super-recognisers, who do not need to focus on familiarising themselves with the way that faces vary across different viewpoints. In contrast to typical recognisers, super-recognisers may gather structural information automatically, without needing to have their attention focused on it by a procedure.

The Relationship Between Confidence and Accuracy

The relationship between confidence and accuracy has been investigated in face recognition (e.g., Brewer & Wells, 2006; Sauer & Brewer, 2015; Wixted & Wells, 2017), but only one previous face matching study has systematically analysed the relationship between confidence and accuracy (Stephens, Semmler, & Sauer, 2017). Confidence ratings have been recorded in a minority of face-matching studies, with results showing that whilst super-recognisers might be more confident than controls (Bobak, Hancock et al., 2016; Davis et al., 2016), even in typical recognisers, confidence has the potential to be diagnostic of accuracy

(Stephens et al., 2017; White et al., 2014). If confidence predicts accuracy, confidence should be taken into account in applied settings.

The Current Study

To investigate possible methods of improving face matching accuracy, we tested how performance varied according to interactivity and levels of orientation information in both ‘typical’ face recognisers (Experiment 1) and ‘superior’ face recognisers (Experiment 2). Consistent with previous research (e.g. Belanova, Davis, & Thompson, 2018), groups were defined based on scores on the 102-trial standardized Cambridge Face Memory Test: Extended (Russell, Duchaine, & Nakayama, 2009). Superior face recognisers achieved scores of at least 93 out of 102 (91%), expected to be achievable by roughly 2% of the population (Belanova et al., 2018; Bobak, Pampoulov, & Bate, 2016). Typical face recognisers scored below this threshold. Participants compared a static image to either a single static image (frontal condition), a series of static images of the face at different orientations (orientations condition), a video showing the face moving from side to side (moving condition), or an interactive image which could be maneuvered into different orientations using the computer mouse (interactive condition). In Experiments 3 and 4 we directly compared the performance of typical and superior recognisers in the frontal and interactive conditions.

We predicted that typical recognisers would benefit from the availability of orientation information, because it should facilitate the building of a view-invariant representation. We also predicted that typical recognisers would benefit from interactivity because it should direct their attention to the way in which faces vary across viewpoints. If superior recognisers are better at extracting features that are invariant to viewpoint from single images, we would not expect orientation information or interactivity to be as beneficial.

Experiment 1: Typical Face Recognisers

This experiment examined the effect of multiple viewpoints and viewer interaction on face matching. We were also interested in the effects of these stimulus conditions on the confidence-accuracy relationship.

Method

Design

This was a 4 x 2 mixed factorial design. The between-subjects factor was the comparison image type (frontal, orientations, moving, interactive). The within-subjects factor was identity (same or different). The dependent variables were matching accuracy and self-rated confidence.

Participants

Participants who had previously completed the Cambridge Face Memory Test: Extended (CFMT+) on www.superrecognisers.com, and scored 92 or less, were invited to participate via email. All participants had agreed to be contacted about subsequent experiments. A total of 310 participants completed the experiment. In the interactive condition, 26 participants were excluded because they did not move the comparison image in any of the trials. These data were never analysed. The final sample consisted of 284 participants (119 male, 165 female), with an age range of 18-67 years (*M* = 33.7, *SD* = 10.8). Their mean CFMT+ score was 76.3 (*SD* = 9.4). Mean CFMT+ scores in other samples (not excluding extreme scores) tend to vary between around 70 and 75 (Bobak, Pampoulov et al., 2016; Russell et al., 2012). Ethical approval for the experiment was granted by the local Research Ethics Committee.

Apparatus and Materials

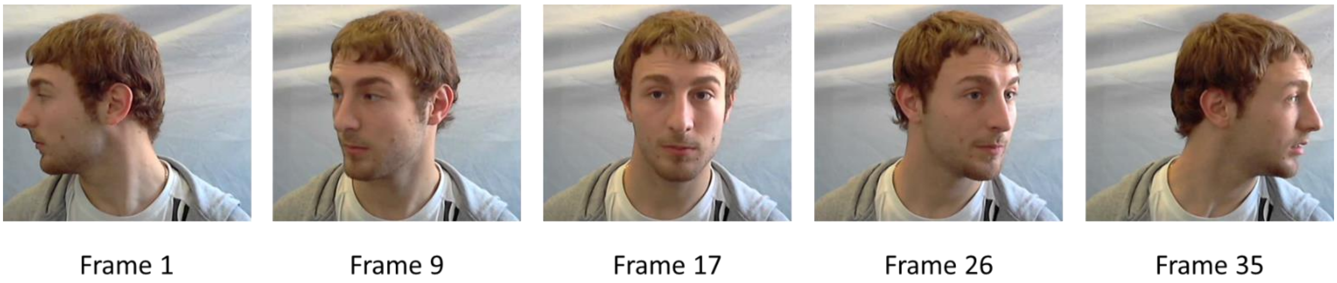
The stimuli were taken from UNSW Unfamiliar Face and Voice Database (White, Burton, & Kemp, 2016). For each of the 233 people in this corpus, there is a high-quality head and shoulders video of their head turning from 90 degrees left to 90 degrees right, as well as a set of Facebook facial images (*M* = 12.03 images, SD = 1.93 images). We selected 94 Caucasian adults (58 female, 36 male) from the database. They had an age range of 17-32 years (*M* = 19.48, *SD* = 2.20). For each person we used the video and two of the Facebook images (Facebook 1 and Facebook 2). The Facebook images were selected according to the following criteria: the images should provide a clear view of the person’s face, be only a head and shoulders shot, and feature no other individuals. The faces showed a variety of different facial expressions, and head orientations. However, the majority were facing towards the camera, with only slight deviations from the frontal orientation.

Each matching trial was constructed using a Facebook image and the video (comparison image). Same identity trials featured two different images of the same person. For different identity trials the foils were selected on the basis of a multidimensional scaling analysis as part of a previous study (White et al., 2016). The experiment consisted of 94 trials. Each identity featured in both a same identity trial and a different identity trial. So that the same image never appeared twice, Facebook 1 was used for same identity trials, and Facebook 2 was used for different identity trials. All images presented in the experiment were the same height (300 pixels) and focal distance.

Each of the four conditions involved presenting different visual information for the comparison image. In the frontal condition, only the front of the face was shown. In the orientations condition, still images of the face were depicted sequentially from five viewpoints (frontal, left and right three quarter, and left and right profile), as shown in Figure 1. Each viewpoint was shown for a total of 500ms. The face appeared to turn from one side to the other and then back again as the five different viewpoints were shown in sequence. In the moving condition, the participant saw a 4 s video clip that showed the face move fluidly from 0-180 degrees (i.e., rotating from 0 degrees through to 180 degrees). In the interactive condition, the user could move the face from 0 to 180 degrees using a computer mouse and pause the face in any angle they wished for any length of time desired. The programme recorded whether participants in the interactive condition moved the faces or not.

Figure 1

A Selection of Frames Extracted from the Studio Video



Frame 1 Frame 9 Frame 17 Frame 26 Frame 35

Note. The images show side, ¾ and frontal orientations on both left and right sides.

The participants completed the experiment online. The website was disabled on mobile phones/tablets. All participants completed the experiment on a desktop/laptop computer.

Procedure

In the invitation email, the participants were provided with a unique ID to use for the experiment. They gave permission that after the deadline for withdrawal had passed, we would then be able to match up their anonymized scores with their CFMT+ scores.

When participants clicked on the link to complete the study, they were randomly allocated to one of the four conditions. Participants completed 94 trials, which were presented in a random order. In each trial the Facebook image was always shown on the left, and the comparison image was shown on the right. Below the images, participants were asked, ‘Are these the same people?’. They clicked either *same* or *different* to register their response. They were also asked, ‘How confident are you in the accuracy of your response from 0% to 100%, with 0% being *not confident at all*, and 100% being *absolutely confident*. They selected from a drop-down menu of 11 possible responses (*0, 10, 20* etc.). No time pressure was imposed. The faces remained visible until participants clicked ‘Next’ to proceed to the next trial. In the orientation and moving condition the right-hand faces continued to move from side to side and back again for the duration of the trial.

Results

The data for the last (94th) trial in the frontal, orientation and moving conditions did not save due to a programming error, so only data from the first 93 trials were analysed. Here we report the results in brief. Supplementary information and analyses are presented in Appendix A.

Accuracy

Data were analysed using multilevel logistic regression with accurate matches scored as 1 and inaccurate matches as 0 in a 4 (image type: frontal, orientations, moving, interactive) x 2 (identity: same or different) factorial design. This analysis treated participants and the two face stimuli sets as fully crossed random factors using the R package lme4 (Bates, Maechler, Bolker, & Walker, 2015; R Core Team, 2018). These results are shown in Table 1.

Table 1

Summary of Likelihood Tests for the 2 x 4 Factorial Analysis, Experiment 1

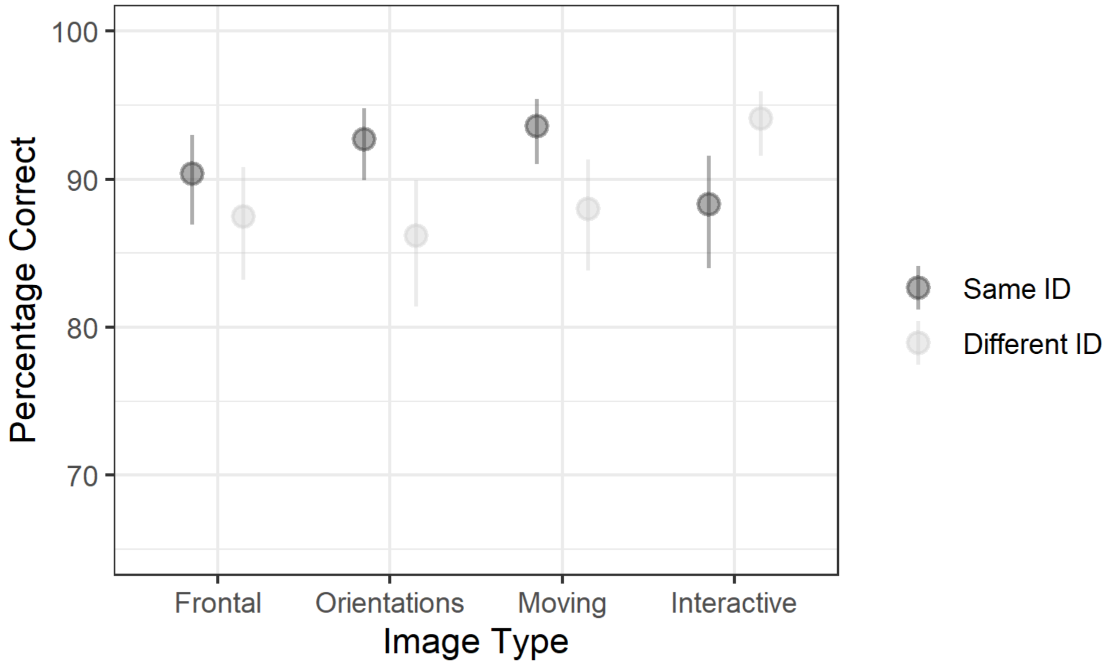
Source	*df*	G^2	*p*
Identity	1	2.39	.121
Image type	3	10.07	.018
Identity x Image type	3	165.1	<.001

The main effect of image type was significant, and there was an interaction between identity and image type. Figure 2 aids interpretation of this main effect and interaction, showing the means and 95% confidence intervals for accuracy in each of the eight conditions.[1]

[1] These means and CIs are back transformed from the log odds estimates in the full interaction model incorporating random effects. These estimates reduce or remove bias from missing responses and exhibit shrinkage. Shrinkage estimators reduce the impact of extreme or unusual units (participants, faces).

Figure 2

Face Matching Accuracy for Frontal, Orientations, Moving and Interactive Conditions, Experiment 1



Note. Error bars show 95% CIs for the condition means.

Overall accuracy in Experiment 1 was 90.1%, 95% CI [87.7, 92.1]. Overall accuracy in the frontal condition was 89.0%, 95% CI [86.1, 91.4], in the orientations condition it was 89.9%, 95% CI [87.1, 92.2], in the moving condition it was 91.2%, 95% CI [88.8, 93.1], and in the interactive condition it was 91.7%, 95% CI [89.1, 93.6]. Pairwise tests with a Hochberg correction (Hochberg, 1988) indicated that the frontal condition had lower average accuracy than the moving and interactive conditions (both $p < .05$), with the other pairwise comparisons non-significant.

Multilevel Signal Detection Analysis

Figure 2 suggests qualitatively different performance for the interactive condition relative to the non-interactive frontal, orientations and moving conditions. Namely, participants appear to be biased towards making ‘same’ responses, but that this bias is reduced or reversed in the interactive condition. To investigate this possibility, we fitted a signal detection theory model as a multilevel probit regression for these data. In this model we treat response (same or different) as the outcome and use identity and image type as predictors using lme4 with participant and the different face sets (i.e. comparison and Facebook images) as random factors (e.g., see Wright, Horry, & Skagerberg, 2009). Table 2 summarizes the criterion and sensitivity *(d′*) estimates for each condition (obtained by transforming the probit regression coefficients). This approach also allowed us to estimate separate random effects for criterion and *d′* (reported in Appendix A, Table A1). These show a clear pattern of differences both in criterion and in *d′*. The *d′* pattern largely follows that observed for accuracy shown in Figure 2, with a slightly more pronounced difference in sensitivity between the frontal and orientations conditions than the moving and interactive conditions. There is also an indication that the criterion shifts between the non-interactive and interactive conditions, with a higher estimate reflecting a more conservative decision standard.

Table 2

Multilevel Signal Detection Analysis: Estimates of Criterion and D Prime (d’), Experiment 1

	Criterion			*d′*		
Condition	*Estimate*	*SE*	95% CI	*Estimate*	*SE*	95% CI
Frontal	1.249	0.119	1.014, 1.479	2.687	0.158	2.377, 2.998
Orientations	1.168	0.129	0.920, 1.423	2.729	0.166	2.403, 3.056
Moving	1.291	0.128	1.044, 1.547	2.928	0.166	2.608, 3.256
Interactive	1.773	0.146	1.448, 2.068	3.029	0.177	2.683, 3.377

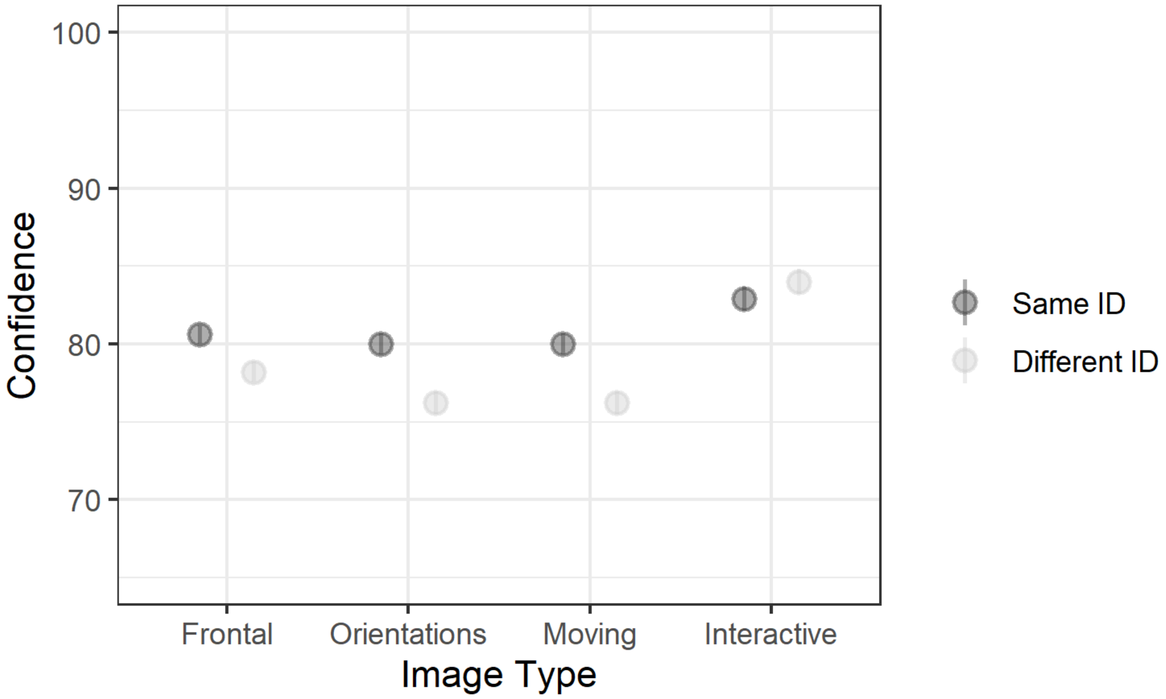
Pairwise tests of the differences in criterion, with *p* adjusted using the Hochberg correction, indicate that the interactive condition had a higher threshold for responding ‘same’ than the other three conditions (all *p* < .001) with no other differences statistically significant ($p > .05$). Thus, typical participants in the interactive condition were more conservative in deciding matches – being more biased towards making ‘different’ responses than in the non-interactive conditions. Additionally, the moving condition tended to have higher d' scores than either the frontal (adjusted one-sided $p = .030$) or orientations conditions ($p = .108$), as did the interactive condition ($p = .024$ and $p = .064$ respectively). A post hoc contrast comparing the average of the moving and interactive conditions with the static conditions also supported this interpretation, $d'_{\text{diff}} = 0.301$, [0.094, 0.511]. The moving and interactive conditions did not differ from each other ($p > .05$).

Confidence

The means and 95% CIs for each of the conditions are shown in Figure 3.

Figure 3

Self-Rated Confidence Following Face Matching Decisions, Experiment 1



Note. Error bars show 95% CIs for the condition means (calculated from the SE).

The Relationship Between Confidence and Accuracy

We ran separate analyses for the four comparison image conditions using the ordinal package in R (Christensen, 2011). Self-rated confidence was the dependent variable, and accuracy (% correct) was the predictor. Two models were compared, one included only intercepts and the other added accuracy as a predictor. Accuracy predicted confidence in all four conditions: frontal (b = 1.1223, SE = 0.062, G^2 = 330.33, p < .001), orientations (b = 1.221, SE = 0.070, G^2 = 301.49, p < .001), moving (b = 1.046, SE = 0.069, G^2 = 230.78, p < .001), and interactive (b = 1.345, SE = 0.094, G^2 = 206.04, p < .001). Descriptively speaking, higher b values indicate a stronger relationship between confidence and accuracy.

Discussion

Overall accuracy was high, exceeding 85% in each condition. There was a main effect of image type, suggesting that typical recognisers benefit from orientation information. This is likely to be because fluid orientation information supports the building of a view-invariant representation, making matching more accurate (Bruce et al., 1999; Hancock et al., 2000; Kramer & Reynolds, 2018). Indeed, performance in both the moving and interactive conditions was more accurate than the frontal condition. The pattern of performance in the frontal, orientations, and moving conditions is consistent with previous face matching literature showing that false alarms are the most common type of error (Davis & Valentine, 2009; Kemp et al., 1997). However, there was an interaction between identity and image type. In the interactive condition accuracy was higher on different identity trials. The multilevel signal detection analysis revealed that typical recognisers were more likely to respond ‘different identity’ in the interactive condition, suggesting that interactivity may increase the salience of differences between facial images.

Confidence predicted accuracy in all the comparison image conditions, supporting previous findings that confidence is diagnostic of accuracy in face matching (Stephens et al., 2017). From an applied point of view, this is reassuring because identifications made with high confidence can have the greatest weight in criminal proceedings (Brewer & Burke, 2002; Cutler, Penrod, & Stuve, 1988; Lindsay, Wells, & Rumpel, 1981).

Experiment 2: Superior Face Recognisers

In Experiment 2 we tested superior face recognisers to investigate whether orientation information and interactivity affect face matching performance. If superior recognisers are particularly good at extracting structural information from static images, we would not expect either orientation information or interactivity to boost performance. As in Experiment 1, we were also interested in the nature of the relationship between confidence and accuracy.

Method

Apart from the following exceptions, the method was identical to Experiment 1.

Participants

Participants who had previously completed the Cambridge Face Memory Test: Extended (CFMT+) on www.superrecognisers.com, and scored 93 or more, were invited to participate via email. A total of 57 participants completed the experiment. In the interactive condition, 9 participants were excluded because they did not move the comparison image in any of the trials. These data were never analysed. The final sample consisted of 48 participants (25 male, 23 female), with an age range of 18-68 years (*M* = 34.7, *SD* = 10). Their mean CFMT+ score was 95.2 (*SD* = 1.7).

Results

As in Experiment 1, the data were not saved for the last trial due to a programming error in the frontal, orientations, and moving conditions. Supplementary information and analyses are available in Appendix B.

Accuracy

Face matching accuracy was analysed using the same method as Experiment 1. Table 3 shows the likelihood chi-square statistic (G^2) and *p*-value associated with comparing individual effects (i.e., comparing a model without the effect to one including all effects of the same order).

Table 3

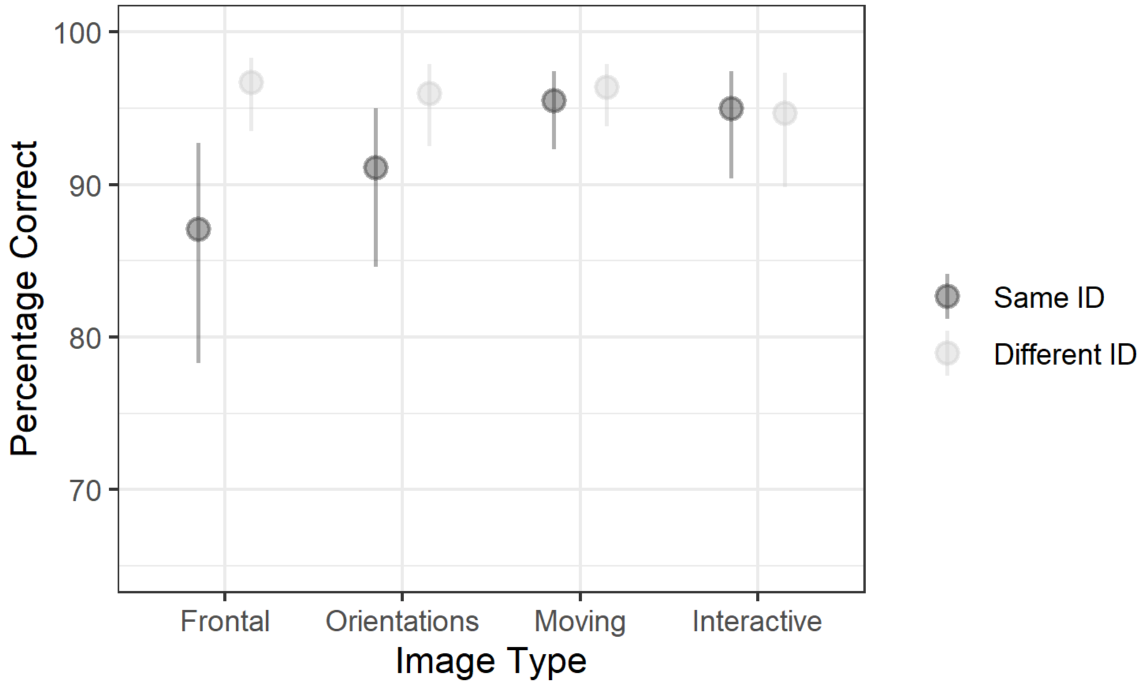
Summary of Likelihood Tests for the 2 x 4 Factorial Analysis, Experiment 2

Source	*df*	G^2	*p*
Identity	1	5.96	.046
Image type	3	4.62	.201
Identity x Image type	3	24.54	<.001

The main effect of identity was significant, and there was an interaction between identity and image type. Figure 5 aids interpretation of the main effect of identity and the interaction between identity and image type, showing the means and 95% confidence intervals for accuracy in each of the eight conditions.

Figure 5

Face Matching Accuracy for Frontal, Orientations, Moving and Interactive Conditions, Experiment 2



Note. Error bars show 95% CIs for the condition means

Overall accuracy in Experiment 2 was 94.7%, 95% CI [92.5, 96.3]. Overall accuracy in the frontal condition was 93.3%, 95% CI [88.6, 96.2], in the orientations condition it was 94.0%, 95% CI [89.9, 96.5], in the moving condition it was 96.0 %, 95% CI [93.6, 97.5] and in the interactive condition it was 94.8%, 95% CI [90.8, 97.1]. Pairwise tests with a Hochberg correction (Hochberg, 1988) indicated that none of the comparisons were significant (*p* > .567). However, there was an advantage for different identity trials 96.0 %, 95% CI [93.8, 97.4] relative to same identity trials, 92.8%, 95% CI [89.3, 95.2].

Multilevel Signal Detection Analysis

Figure 5 suggests qualitatively different performance for the interactive condition and moving conditions relative to the frontal and orientations conditions. As in Experiment 1 this may partly reflect changes in bias and we therefore fitted a signal detection theory model as a multilevel probit regression for these data. We set up the model in the same way as in Experiment 1, treating response (same or different) as the outcome, identity and image type as predictors, and participant and the different face sets as random factors (and again obtained the estimates using brms because of difficulty estimating the Facebook image variance). Table 4 summarizes the criterion and *d*′ estimates for each condition. These show a clear pattern of differences in criterion but less so for *d*′. Despite the appearance of differences in criterion or *d*’ between conditions none of these differences reach statistical significance (all *p* > .05).

Table 4

Multilevel Signal Detection Analysis: Estimates of Criterion and D Prime (d’), Experiment 2

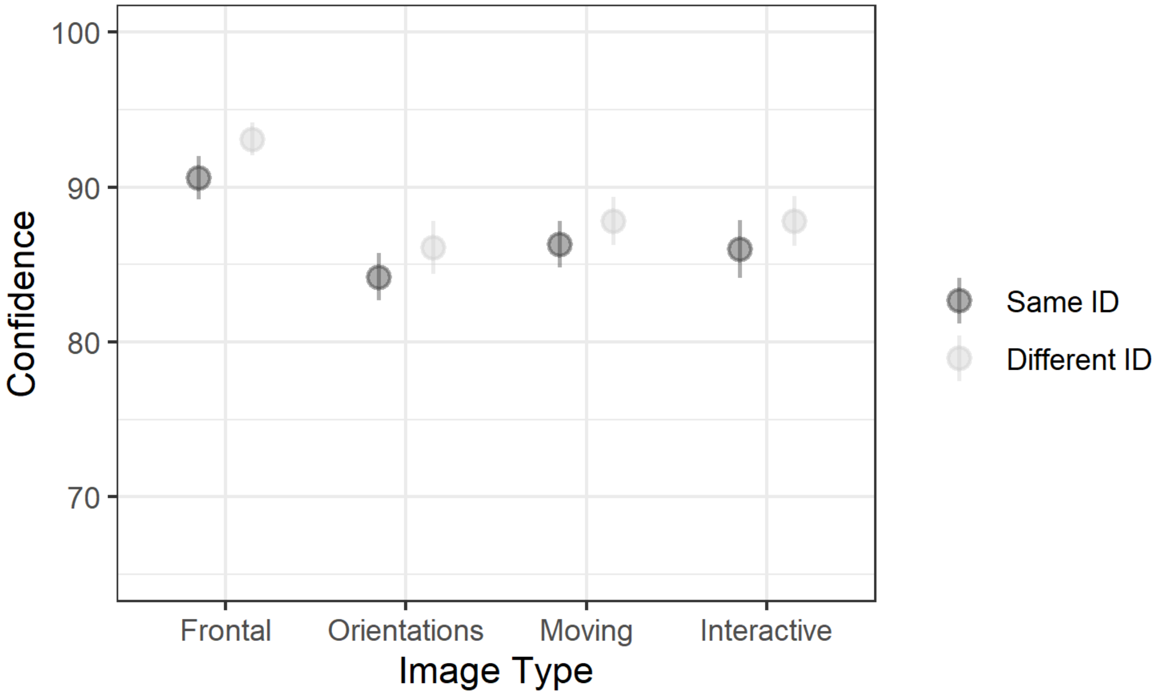
	Criterion			*d*′		
Condition	*Estimate*	*SE*	95% CI	*Estimate*	*SE*	95% CI
Frontal	2.097	0.241	1.639, 2.585	3.598	0.258	3.094, 4.108
Orientations	2.073	0.300	1.497, 2.673	3.626	0.300	3.052, 4.226
Moving	1.994	0.237	1.542, 2.463	3.906	0.258	3.410, 4.427
Interactive	1.850	0.316	1.245, 2.496	3.912	0.325	3.286, 4.567

Confidence

The means and 95% CIs for each of the conditions are shown in Figure 6.

Figure 6

Self-Rated Confidence Following Face Matching Decisions, Experiment 2



Note. Error bars show 95% CIs for the condition means (calculated from the SE)

The Relationship Between Confidence and Accuracy

The relationship between confidence and accuracy was analysed using the same method as Experiment 1. Accuracy predicted confidence in all four conditions: frontal (*b* = 1.794, *SE* = 0.241, G^2 = 54.09, *p* < .001), orientations (*b* = 1.582, *SE* = 0.203, G^2 = 59.61, *p* < .001), moving (*b* = 1.203, *SE* = 0.195, G^2 = 36.91, *p* < .001), interactive (*b* = 2.069, *SE* = 0.245, G^2 = 71.8, *p* < .001).

Discussion

As in Experiment 1, overall accuracy was high (>90%) in each condition. There was a main effect of identity, with participants responding more accurately on different identity trials. This is opposite to the pattern observed for typical face recognisers, who were more accurate on same identity trials in the frontal, orientation, and moving conditions. The results fit with Bobak, Dowsett, & Bate (2016), who found that super-recognisers tended to be more conservative than controls. There was no main effect of image type, which may be because superior recognisers are better at structural encoding, and so unlike typical face recognisers, do not benefit as much from the additional orientation information (Bobak, Hancock et al., 2016). However, there was an interaction between identity and image type: the difference between accuracy on same identity and different identity trials was reduced in the moving and interactive conditions. Whilst the condition means suggest that superior recognisers benefit from fluid movement in the sense that they are less likely to respond conservatively in these conditions, the multilevel signal detection analysis did not reveal any significant differences in criterion across conditions. This could be because overall high performance in Experiment 2 impacts on the ability to detect criterion shifts, or because overall there is less data in comparison to Experiment 1.

Broadly speaking, the confidence-accuracy analyses replicate Experiment 1. There was a relationship between confidence and accuracy in each of the comparison image conditions. The superior recognisers exhibit numerically higher confidence than typical recognisers, which mirrors the pattern of accuracy.

Experiment 3: A Comparison of Typical and Superior Face Recognisers

The results of Experiments 1 and 2 suggest that interactivity has the potential to shift patterns of performance in both typical and superior face recognisers, and may be extremely valuable in settings where it is important to avoid false positive matching decisions (Experiment 1). In Experiments 3 and 4 we compare the novel interactive procedure to the procedure associated with photo-ID, i.e. matching to a frontal image.

The number of superior recognisers tested in Experiment 2 exceeds that of much previous research (Noyes et al., 2017). However, a proportion of participants in the interactive condition were excluded, which risked this condition being underpowered. In Experiment 3 we recruited a greater number of superior recognisers, comparing performance against typical recognisers in order to test whether the two groups process faces in qualitatively different ways, and exhibit different patterns of performance across frontal and interactive conditions. Experiments 3 and 4 were pre-registered (AsPredicted# 30321).

Method

Apart from the following exceptions, the method was identical to Experiment 1 and 2.

Design

This was a 2 x 2 x 2 mixed factorial design. The between-subjects factors were comparison image type (frontal or interactive) and recogniser (typical or superior). The within-subjects factor was identity (same or different). The dependent variables were accuracy, interactivity, and self-rated confidence.

Participants

Participants who had previously completed the CFMT+ on www.superrecognisers.com, were invited to participate via email. We did not send invitations to people who had previously taken part in Experiments 1 or 2. A total of 218 participants completed the experiment. In the interactive condition, 58 participants were excluded because they did not move the comparison image in any of the trials. These data were never analysed. The final sample consisted of 160 participants (104 female, 53 males, 3 prefer not to say) with an age range of 18-73 years, *M* = 43.9, *SD* = 19.6. There were 89 typicals (CFMT+ score: *M* = 71.6, *SD* = 12.6) and 71 superiors (CFMT+ score: *M* = 96.9, *SD* = 2.25).

Results

Supplementary information is presented in Appendix C.

Accuracy

Table 5 shows the likelihood chi-square statistic (G^2) and *p*-value associated with comparing individual effects (i.e. comparing a model without the effect to one including all effects of the same order).

Table 5

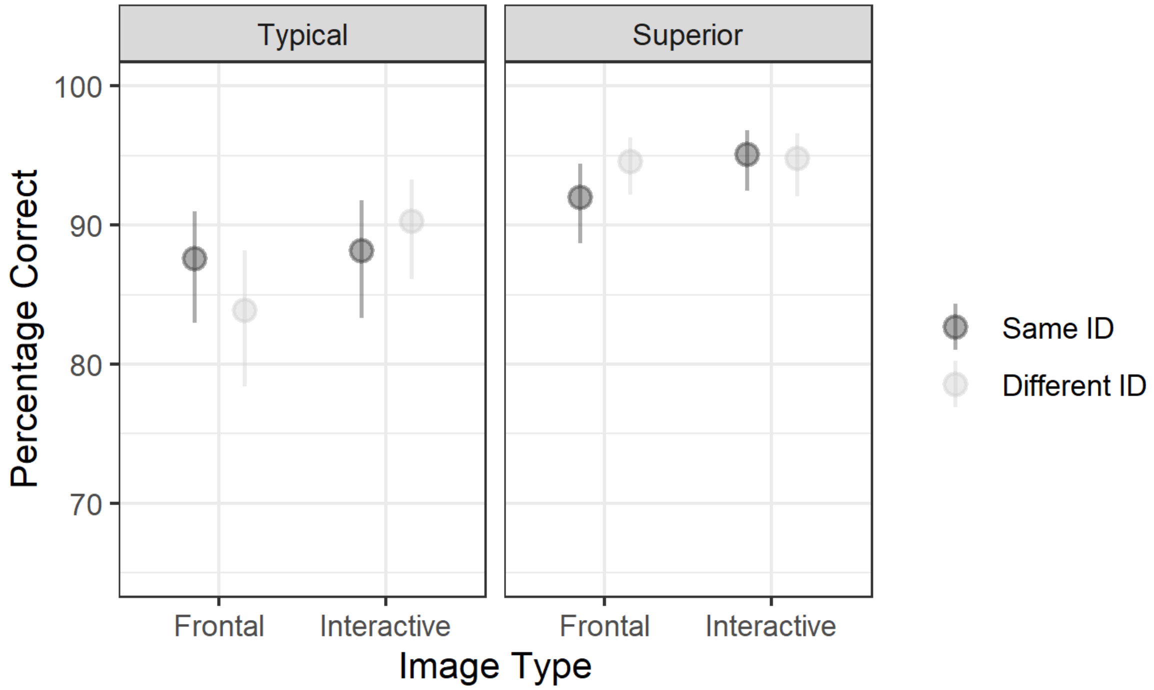
Summary of Likelihood Tests for the 2 x 2 x 2 Factorial Analysis, Experiment 3

Source	*df*	G^2	*p*
Identity	1	.01	.933
Image type	1	9.39	.002
Recogniser	1	62.70	<.001
Identity x Image type	1	2.61	.107
Identity x Recogniser	1	16.37	<.001
Recogniser x Image type	1	<.01	.990
Identity x Image type x Recogniser	1	17.61	<.001

The main effects of image type and recogniser were significant. There was a two-way interaction between identity and recogniser, and a three-way interaction between identity, image type and recogniser. Figure 7 aids interpretation of these effects, showing the means and 95% confidence intervals for accuracy in each of the eight conditions.

Figure 7

Face Matching Accuracy for Typicals and Superiors in the Frontal and Interactive Conditions, Experiment 3



Note. Error bars show 95% CIs for the condition means

Overall accuracy in Experiment 3 was 90.6% [88.1, 92.6]. For typical recognisers in the frontal condition it was 85.8% [82.1, 88.8], and in the interactive condition it was 89.3% [85.8, 92.0]. For superior recognisers in the frontal condition it was 93.4% [91.3, 95.0], and in the interactive condition it was 94.9% [93.0, 96.4].

Interactivity

Following programming improvements, we were able to record and analyse whether or not participants interacted on each trial in the interactive condition[2]. Table 6 shows the likelihood chi-square statistic (G^2) and *p*-value associated with comparing individual effects.

[2] These data did not record reliably for each trial in Experiments 1 and 2.

Table 6

Summary of Likelihood Tests for the 2 x 2 Factorial Analysis of Interactivity, Experiment 3

Source	*df*	G^2	*p*
Identity	1	1.41	.236
Recogniser	1	0.27	.602
Identity x Recogniser	1	4.25	.039

There was a two-way interaction between identity and recogniser, showing that while typical recognisers interacted as much on same identity trials (58.7% [56.04, 61.63]) as different identity trials (58.3% [55.63, 60.97]), superior recognisers interacted more on same identity trials (57.8% [54.85, 60.75] than different identity trials (51.9% [48.92, 54.88]).

We tested whether average accuracy for each trial in the frontal condition (Experiment 3) predicted interactivity in the interactive condition. Average accuracy in the frontal condition provided a metric that was uncontaminated by whether participants interacted. Both typical and superior recognisers were more likely to interact on difficult trials (typical: b = -2.545 , SE = 0.637, G^2 = 14.94, p <.001; superior: b = -3.411, SE = .552, G^2 = 33.73, p < .001).

Multilevel Signal Detection Analysis

As in Experiments 1 and 2 multilevel probit regression was used to fit a signal detection model for responses to same versus different targets. Table 7 shows the estimates of criterion and *d'* for the frontal and interactive conditions by group.

Table 7

Multilevel Signal Detection Analysis: Estimates of Criterion and D Prime (d′), Experiment 3

	Criterion			*d′*		
Condition	*Estimate*	*SE*	95% CI	*Estimate*	*SE*	95% CI
Typical						
Frontal	1.084	0.127	0.834, 1.328	2.385	0.175	2.047, 2.730
Interactive	1.428	0.161	1.115, 1.739	2.765	0.202	2.369, 3.163
Superior						
Frontal	1.782	0.140	1.508, 2.060	3.382	0.185	3.020, 3.748
Interactive	1.885	0.179	1.535, 2.239	3.754	0.219	3.324, 4.184

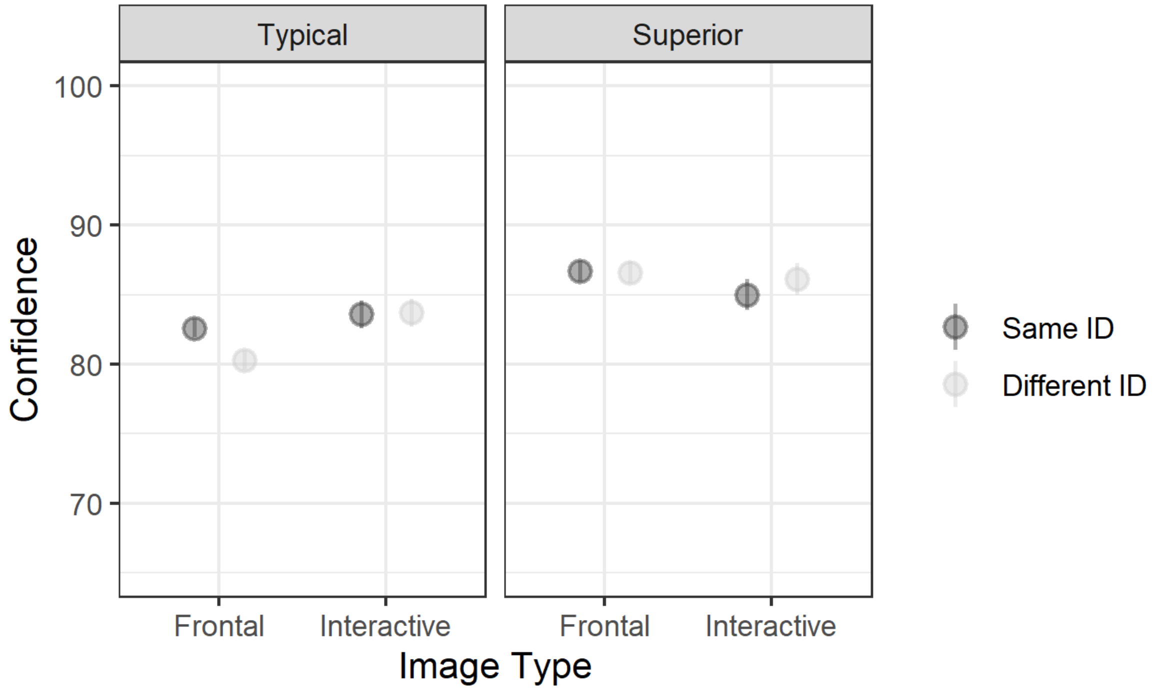
The effects of condition and group and their interaction on criterion and *d′* were tested using contrasts. Criterion was higher on average for superior recognisers than typical recognisers c_{diff} = 0.578 [0.383, 0.777]. There was also a difference in criterion between image conditions, c_{diff} = 0.224 [0.024, 0.426], but no interaction between group and image condition, c_{diff} = 0.242 [-0.698, 0.239]. For *d*’, superior recognisers were better at discriminating matches from non-matches than the typical group d'_{diff} = 0.997 [0.742, 1.257]. The interactive condition also had higher *d′* scores than the frontal condition, d'_{diff} = 0.376 [0.179, 0.576], but there was little indication of a group by condition interaction, d'_{diff} = 0.008 [-0.463, 0.474].

Confidence

The means and 95% CI for each of the conditions are shown in Figure 8.

Figure 8

Self-Rated Confidence Following Face Matching Decisions, Experiment 3



Note. Error bars show 95% CI for the condition means (calculated from the SE).

The Relationship Between Confidence and Accuracy

The relationship between confidence and accuracy was analysed using the same method as Experiments 1 and 2. Accuracy predicted confidence in both of the image conditions for typical recognisers: frontal (b = 1.202, SE = 0.068, G^2 = 314.32, p < .001), and interactive (b = 1.247, SE = 0.110, G^2 = 127.39, p < .001), as well as superior recognisers: frontal (b = 1.491, SE = 0.104, G^2 = 203.57, p < .001), and interactive (b = 1.328, SE = 0.162, G^2 = 66.82, p < .001).

Discussion

In line with the results of Experiments 1 and 2, accuracy was high, exceeding 80% in all conditions. The overall pattern of results replicates and clarifies our previous findings. There was a main effect of recogniser. As expected, the superior recognisers responded more accurately than typical recognisers, and being more conservative, superiors were particularly accurate on different identity trials (Bobak, Dowsett et al., 2016). There was also a main effect of image condition, with higher accuracy for interactive images. The results of Experiment 3 show that superior recognisers do benefit from interactivity. The failure to detect an overall benefit of interactivity in Experiment 2 is therefore unlikely to have been due to ceiling effects as the same stimuli were used across both experiments. Changes in performance across image conditions were not due to changes in criterion; the pattern of *d'* estimates reflects higher sensitivity in the interactive condition. The detection advantage did not vary by group, indicating that the three-way interaction in terms of accuracy is a product of change in bias. As in Experiment 1, typical recognisers respond more conservatively in the interactive condition.

The pattern of responses for typical recognisers was the same as in Experiment 1, despite differences in the mean CFMT+ score. In Experiment 1 the mean CFMT+ score (76.34) was relatively high. In Experiment 3, the mean score of 71.56 sat at the bottom of the range (around 70-75) observed in other studies (Bobak, Pampoulov et al., 2016; Russell et al., 2012).

The confidence-accuracy results replicate Experiments 1 and 2, showing a strong relationship in each condition.

However, having observed high levels of overall accuracy in Experiments 1, 2 and 3, we cannot rule out the possibility that ceiling effects mask the true magnitude of the interactivity effect, particularly as participants tended to interact more on difficult trials. Experiment 4 addressed this issue.

Experiment 4: A Comparison of Typical and Superior Recognisers Matching Pixelated Images

In Experiment 4 the testing conditions were designed to reflect potential challenges encountered in forensic contexts. The police often use low resolution (pixelated) images from

CCTV footage to identify suspects, comparing these against a database of high-quality images. Pixelation reliably reduces accuracy in unfamiliar face matching (Bindemann, Attard, Leach, & Johnston, 2013; Ritchie et al., 2018). In this experiment the Facebook image was degraded by pixelation, so we expected accuracy to be lower than in Experiments 1, 2 and 3. We also expected superiors to outperform typicals, and for performance to be most accurate in the interactive condition.

Method

Apart from the following exceptions, the method was identical to Experiment 3.

Participants

Participants who had previously completed the CFMT+ on www.superrecognisers.com, were invited to participate via email. We did not send invitations to people who had previously taken part in Experiments 1, 2 or 3. A total of 253 participants completed the experiment. In the interactive condition, 56 participants were excluded because they did not move the comparison image in any of the trials. These data were never analysed. The final sample consisted of 197 participants (128 female, 68 male, 1 prefer not to say) with an age range of 21-86 years, *M* = 43.3, *SD* = 16.1. There were 104 superiors (CFMT+ score: *M* = 96.5, *SD* = 2.6) and 93 typicals (CFMT+ score: *M* = 75.8, *SD* = 12.1).

Apparatus and Materials

The Facebook images were pixelated using the Mosaic function in Adobe Photoshop 2020, which converts pixels into weighted averages. Each 6 x 6 pixel square in the image was transformed into a sub-sampled block of equal luminance. Before pixelating, each image had a horizontal resolution of 300 pixels. After pixelating, each image had a horizontal resolution of 50 pixels.

Results

Supplementary information is presented in Appendix D.

Accuracy

Table 8 shows the likelihood chi-square statistic (G^2) and *p* value associated with comparing individual effects (i.e. comparing a model without the effect to one including all effects of the same order).

Table 8

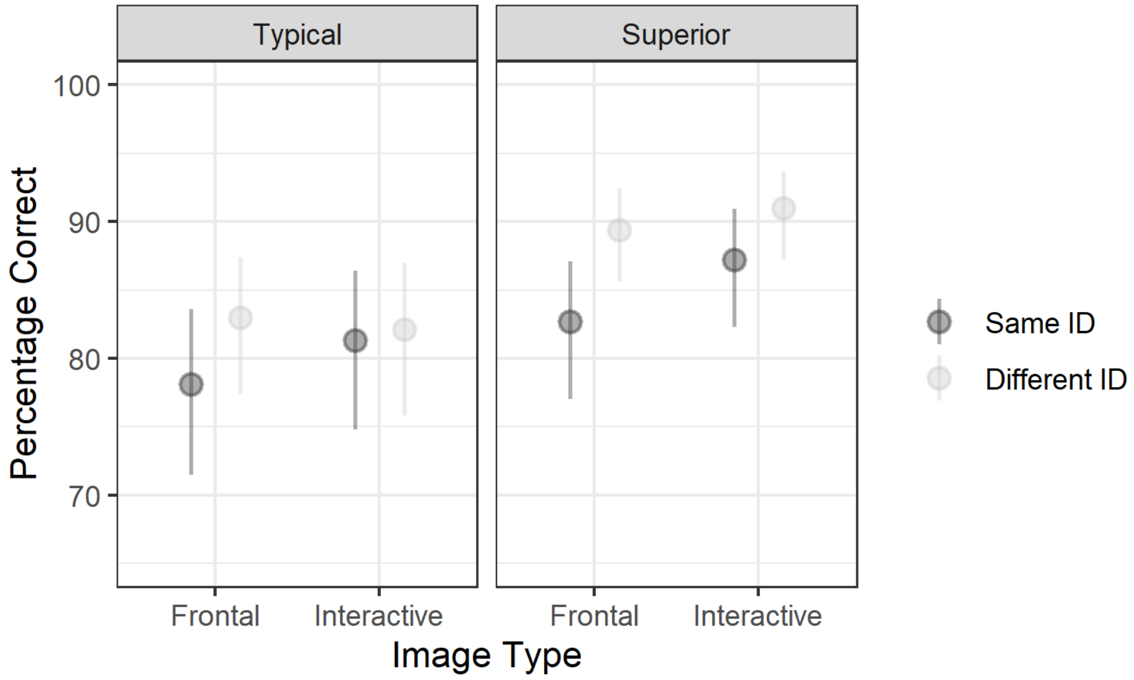
Summary of Likelihood Ratio Tests for the 2 x 2 x 2 Factorial Analysis, Experiment 4

Source	*df*	G^2	*p*
Identity	1	2.71	.100
Image type	1	5.64	.018
Recogniser	1	44.94	<.001
Identity x Image type	1	6.58	.010
Identity x Recogniser	1	12.56	<.001
Recogniser x Image type	1	1.83	.176
Identity x Image type x Recogniser	1	.148	.701

The main effects of image type and recogniser were significant. There was a two-way interaction between identity and image type, and a two-way interaction between identity and recogniser. Figure 9 aids interpretation of these effects, showing the means and 95% confidence intervals for accuracy in each of the eight conditions.

Figure 9

Face Matching Accuracy for Typicals and Superiors in the Frontal and Interactive Conditions, Experiment 4



Note. Error bars show 95% CIs for the condition means.

Overall accuracy in Experiment 4 was 84.6% [80.8, 87.7]. For typical recognisers in the frontal condition it was 80.7% [76.0, 84.6], and in the interactive condition it was 81.7% [76.8, 85.8]. For superior recognisers in the frontal condition it was 86.4% [82.8, 89.3], and in the interactive condition it was 89.2% [85.9, 91.8].

Interactivity

Table 9 shows the likelihood chi-square statistic (G^2) and *p*-value associated with comparing individual effects.

Table 9

Summary of Likelihood Tests for the 2 x 2 Factorial Analysis of Interactivity, Experiment 4

Source	*df*	G^2	*p*
Identity	1	114.29	<.001
Recogniser	1	0.82	.366
Identity x Recogniser	1	26.21	<.001

There was a main effect of identity, with both groups of recognisers interacting more on different identity trials. The difference was greater for superior recognisers (same identity trials: 14.6% [12.82, 16.38]; different identity trials: 62.4% [59.95, 64.85]) than typical recognisers (same identity trials: 17.0% [15.07, 18.93]; different identity trials 53.9% [51.34, 56.46]).

As in Experiment 3, we tested whether average accuracy for each trial in the frontal condition (Experiment 4) predicted interactivity in the interactive condition. Typical recognisers were no more likely to interact on more difficult trials (b = 0.061, SE = 0.796, G^2 = 0.01, p = 0.940), and nor did we detect an effect for superior recognisers (b = -1.443 , SE = 1.112, G^2 = 1.68, p = 0.194).

Multilevel Signal Detection Analysis

As in the earlier experiments a multilevel probit regression was used to obtain estimates of criterion and *d*' for each condition. Table 10 shows the estimates of criterion and *d*' for the frontal and interactive conditions by group.

Table 10

Multilevel Signal Detection Analysis: Estimates of Criterion and D prime (d′), Experiment 4

	Criterion			*d′*		
Condition	*Estimate*	*SE*	95% CI	*Estimate*	*SE*	95% CI
Typical						
Frontal	1.011	0.131	0.760, 1.271	1.807	0.171	1.470, 2.140
Interactive	1.009	0.155	0.702, 1.312	1.912	0.186	1.550, 2.278
Superior						
Frontal	1.397	0.130	1.146, 1.654	2.396	0.172	2.056, 2.738
Interactive	1.469	0.156	1.167, 1.779	2.675	0.189	2.303, 3.041

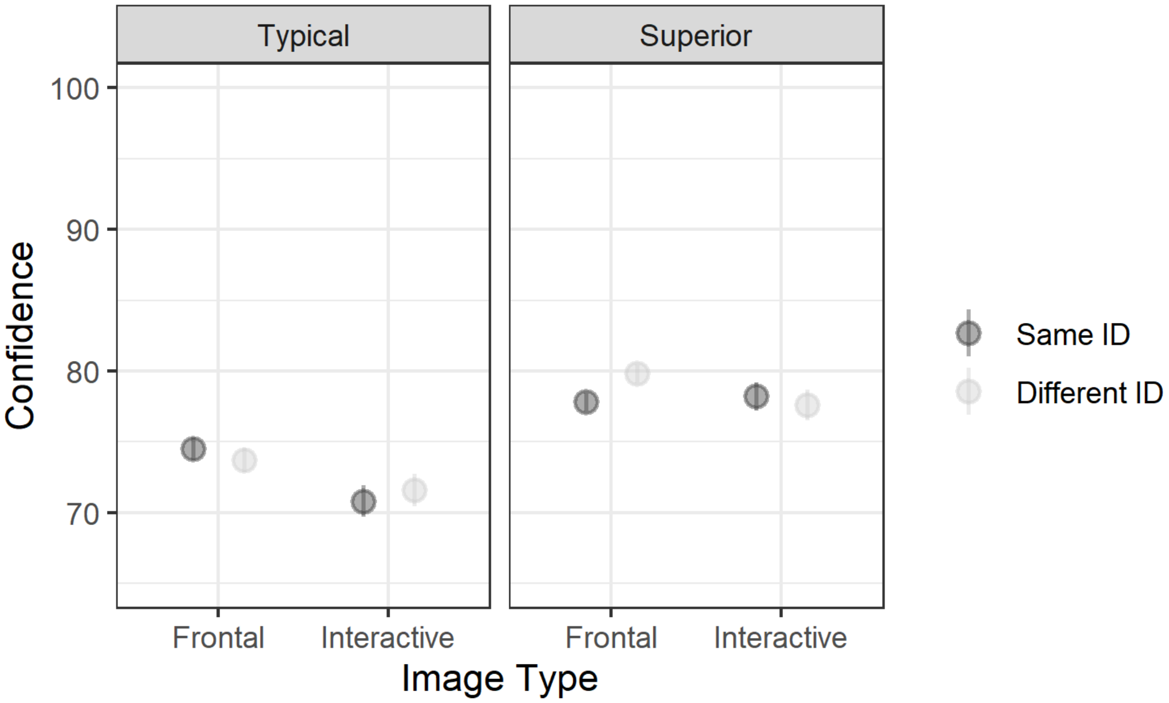
There was a main effect of criterion which, although lower overall than in Experiment 3, remained higher for superior recognisers than typical recognisers, c_{diff} = 0.423 [0.263, 0.580]. However, there was little evidence of a difference in criterion between conditions, c_{diff} = 0.035 [-0.120, 0.194], or of an interaction between group and condition, c_{diff} = 0.074 [-0.297, 0.451]. For *d′* there were main effects of group and condition. Superior recognisers were better at detecting matches versus non-matches than the typical group d'_{diff} = 0.676 [0.499, 0.848]. The interactive condition also had higher *d′* scores than the frontal condition, d'_{diff} = 0.192 [0.021, 0.369], and there was also tentative evidence of a greater effect of the interactive condition for the superior recognisers, d'_{diff} = 0.174 [-0.172, 0.513]. It is worth noting that the advantage for superior recognisers and for the interactive condition shows the same general pattern as Experiment 3, but with smaller effect sizes.

Confidence

The means and 95% CIs for each of the conditions are shown in Figure 10.

Figure 10

Self-Rated Confidence Following Face Matching Decisions, Experiment 4



Note. Error bars show 95% CIs for the condition means (calculated from the SE).

The Relationship Between Confidence and Accuracy

The relationship between confidence and accuracy was analysed using the same method as previous experiments. Accuracy predicted confidence in both of the image conditions for typical recognisers: frontal (b = 0.690, SE = 0.061, G^2 = 127.67, p < .001), and interactive (b = 0.810, SE = 0.086, G^2 = 89.20, p < .001), as well as superior recognisers: frontal (b = 1.037, SE = 0.056, G^2 = 259.01, p < .001), and interactive (b = 1.170, SE = 0.102, G^2 = 130.64, p < .001).

Discussion

In Experiment 4, one image in each pair was pixelated in order to eliminate ceiling effects. Average accuracy was lower than in Experiment 3. For typical recognisers, the error rate was around 20%, sitting in the middle of the range typically observed in lab-based experiments (Bruce et al., 1999; Megreya & Burton, 2006, 2008). As in Experiment 3, superiors were more accurate than typical recognisers, and they responded more conservatively, performing particularly accurately on different identity trials.

Overall accuracy was higher in the interactive condition than the frontal condition. Whilst ceiling effects may have operated in the previous experiments, the results of Experiment 4 provide reassurance that this did not mask the magnitude of the interactivity effect. Based on the pattern of *d'* estimates, it appears that degrading the stimuli and making the task harder reduced the benefit afforded by the interactive condition. The effects are smaller in magnitude, rather than larger.

The two-way interaction between identity and image type reflects an overall advantage for different identity trials, with the advantage less pronounced in the interactive condition. In Experiments 1 and 3, typical recognisers were more likely to respond 'same' when both images were static and high quality. In Experiment 4, pixelating one image likely magnified apparent differences between faces. As a result, same identity trials became more difficult, and accuracy was higher on different identity trials. This pattern is consistent with the results of previous face matching studies. Bindemann et al. (2013) observed a more dramatic drop in performance on same identity trials compared to different identity trials when one image was pixelated. The data we present in Experiment 4 suggests that interactivity may mitigate this effect, supporting performance on same identity trials for both typicals and superiors.

In Experiment 3, participants interacted more on difficult trials. In Experiment 4, both groups of recogniser interacted more on different identity trials despite same identity trials being more difficult. Pixelating one image may have disrupted their assessment of difficulty, preventing optimal use of the system.

As expected, there was a relationship between confidence and accuracy in all conditions. However, the relationship was not as strong as when both images were high quality (Experiments 1, 2 and 3). Indeed, the relationship was weaker for typicals compared to superior recognisers.

General Discussion

Across four experiments we have presented strong evidence that fluid orientation information and interactivity boosts face matching performance. It supports performance across the spectrum of face recognition ability, and across different image qualities. The findings have important security implications, underlining the forensic utility of interactivity for identity verification. Any significant difference, even if the effect sizes are small, has the potential to be meaningful in an applied context. A single fraudulently obtained passport provides the opportunity to open bank accounts, take out loans, or apply for mortgages. Indeed, criminals using fraudulently obtained travel documents, are likely to have convictions for serious crimes (Harper, 2016).

Typical and Super-Recognisers

The findings are important from a theoretical point of view, contributing to the debate about possible differences in the way typical and super-recognisers process faces (e.g., Bate et al., 2019; Bobak, Bennetts et al., 2016; Bobak, Hancock et al., 2016; Bobak, Parris, Gregory, Bennetts, & Bate, 2017; Robertson et al., 2019; Russell et al., 2009). Our data do not fully support the hypothesis put forward by Bobak, Hancock et al. (2016) that super-recognisers are better than typical recognisers at structural encoding and can construct view-independent representations from static images. Both types of recogniser benefitted from additional viewpoint information provided by the interactive image, suggesting at least some reliance on pictorial encoding strategies. Our findings support those of Bate et al. (2019), who argue that super recognisers simply sit at the extreme of the face recognition spectrum.

This does not mean that typical and super-recognisers exhibit identical patterns of performance. When comparing two high-quality images, typical and superior recognisers differ in terms of criterion placement (Experiments 1, 2 and 3) (see also Bobak, Dowsett et al., 2016). By default (i.e., without the benefit of fluid orientation information or interactivity), it would seem that typical recognisers focus on between-image similarities and look for evidence that the two faces depict the same person, whereas superior recognisers focus on between-image differences. This would explain higher accuracy on same identity trials for typical recognisers (Experiments 1 and 3), and higher overall accuracy on different identity trials for superior recognisers (Experiment 2, 3 and 4). With a greater amount of facial information available and the ability to self-select which information to use, interactivity seems to shift the focus for typical recognisers and highlights differences, making them both more accurate, and more conservative.

Crucially though, it cannot be argued that the value of interactivity mainly lies in highlighting differences between images. In Experiment 4 the degraded image quality affected both types of recogniser similarly by increasing the salience of differences and resulting in higher accuracy on different compared to same identity trials. Interactivity mitigated this effect to some extent for both typicals and superiors, driving same identity performance up towards different identity performance.

Using the Interactive Procedure for Identity Verification

Super recognisers are known to outperform typical recognisers on face matching tasks (Belanova et al., 2018; Bobak, Dowsett et al., 2016; Bobak, Hancock et al., 2016), and their skills are sought after in forensic contexts. An innovation that boosts superior recogniser performance when comparing high quality (Experiment 3) and mismatched quality (Experiment 4) images is therefore important. We are not aware of other studies that have provided specific evidence that super recogniser performance can be optimised in such a way. Ritchie et al. (2018) present a method of overcoming the deleterious effect of pixelation by creating an average of several poor-quality images to be compared to a high-quality image, but they do not compare performance across typical and super recognisers. The success of Ritchie et al.'s (2018) method in the field depends on there being several poor-quality images available. Whilst both interactivity and averaging likely work by increasing the amount of visual information available and enabling the operator to reduce the contribution of within-person variability as a source of error (Jenkins, White, Van Montfort, & Burton, 2011; Ritchie et al., 2018), one benefit of interactivity is that it can be used when the police only possess a single poor quality image of the suspect to be compared to the interactive face.

A further benefit of the procedure applies to typical recognisers. In Experiment 1 and 3, interactivity supported the performance of typical recognisers on different identity trials. The utility of interactivity is underlined when we consider that most ID verification tasks involve same identity trials. Accurate performance on different identity trials is therefore crucial for preventing identity fraud.

The Participant Sample

It is important to address points about the samples used in this study. Firstly, the participants were invited to take part via www.superrecognisers.com and are likely to have been highly motivated. They took part in initial studies because of their interest in super recognition, and agreed to be contacted about future studies. We cannot rule out the possibility that typical and superior recognisers differed more in terms of motivation than natural ability (see Noyes et al., 2017). On the other hand, it has been shown that differences in incentive-based motivation between groups do not affect scores (Bobak, Dowsett et al., 2016).

All participants had previously received their results on the CFMT+, Glasgow Face Matching Task, and a short-term face memory test. Whilst they were not explicitly told whether they were super recognisers, they were told whether their scores fell within the top 5, 10, 25 or 50% of participants. We do not believe that this affected the results because in the frontal condition (Experiments 1, 2 and 3) both groups behaved in a way that was consistent with previous literature. Superior recognisers were both more accurate and more conservative than typical recognisers (Bobak, Dowsett et al., 2016), and typical recognisers were more likely to commit errors on different identity trials (Davis & Valentine, 2009; Kemp, Towell, & Pike, 1997).

The superior recognisers in this study were people scoring 93 or more on the CFMT+. Whilst we acknowledge that Bobak, Pampoulov et al. (2016) recommend that a score of 95 should be used as a cut-off for super-recognition, Belanova et al. (2018) have found no difference in outcomes on a series of tests when comparing participants who scored 93/94 to those scoring over 95. We have referred to our participants as ‘superior recognisers’ rather than super-recognisers because the latter term tends to be reserved for people who have undertaken a series of neuropsychological tests. Nevertheless, the mean CFMT+ scores in Experiments 2, 3 and 4, range between 95.19 and 96.50, and so are similar to the means in previous super-recogniser studies (95.7, Bobak, Hancock et al., 2016; 97.7, Bobak, Dowsett et al., 2016).

Related to the above points about the CFMT+ is the potential for measurement error when trying to capture general face recognition and identification ability using existing standardised tests. Such tests do not always predict performance on less standardised tests (Balsdon, Summersby, Kemp, & White, 2018), and there are calls for existing screening protocols for super recognition to be expanded (Bate et al., 2018). However, whilst the CFMT+ is unlikely to enable us to perfectly distinguish typical from super-recognisers, individual differences in ability, test-specific strategies, and within-person differences in attention (e.g., distraction, fatigue) may play a role in explaining at least some of this measurement error. For our purposes, we are confident that the CFMT+ provides a satisfactory way of discriminating between groups of typical and superior face recognisers (Bobak, Pampoulov et al., 2016).

Future Directions

Our results underline the importance of 3D view-independent representations in face matching. As the Facebook images were profile images, the vast majority show people facing towards the camera, only slightly (if at all) offset from centre. The results may have revealed a bigger benefit for the moving/interactive conditions if the Facebook images had varied more in terms of orientation. We would expect performance in the frontal condition to particularly suffer (Bruce et al. 1999; Hancock et al., 2000), but performance in the interactive condition to benefit, boosted by the participants' ability to minimise within-person variability across images and to carefully compare faces at the same orientation.

We cannot be sure how these lab-based findings might translate into specific applied contexts. Whilst the effect may be attenuated in the field, it is equally possible that it might be amplified owing to higher levels of motivation (Moore & Johnston, 2013), and knowledge of incorrect response implications. Future research should test the procedure in the field, and across the full range of image types encountered in forensic and security settings (e.g. greyscale, blurred, or partially occluded faces).

Conclusion

In this paper we tested typical and superior recognisers using a novel interactive face matching procedure. In contrast to standard (i.e. static frontal) one-to-one face matching tasks, the procedure provides fluid orientation information, and the opportunity to interact with the comparison facial image by maneuvering it into different orientations. This easy-to-implement procedure has a range of applied benefits: It optimizes the performance of both typical and superior recognisers, and has the potential to highlight both similarities and differences between facial images. The results support the hypothesis that typical and superior face recognisers process faces in qualitatively similar ways: Reliance on pictorial encoding when viewing static images helps to explain the benefit of the interactive procedure.

References

Abelson, R. P., & Prentice, D. A. (1997). Contrast tests of interaction hypothesis. *Psychological Methods, 2*(4), 315-328. http://dx.doi.org/10.1037/1082-989X.2.4.315

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioural sciences.* Palgrave Macmillan.

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*. *3*(1), 25. http://dx.doi.org/10.1186/s41235-018-0114-7

Bate, S., Bennetts, R., Hasshim, N., Portch, E., Murray, E., Burns, E., & Dudfield, G. (2019). The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills. *Journal of Experimental Psychology: Human Perception and Performance, 45*(3), 363-377. https://doi.org/10.1037/xhp0000607

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications, 3*(1), 22. https://dx.doi.org/10.1186/s41235-018-0116-5

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(*1), 1-48. http://dx.doi.org/10.18637/jss.v067.i01

Belanova, E., Davis, J. P., & Thompson, T. (2018). Cognitive and neural markers of super-recognisers' face processing superiority and enhanced cross-age effect. *Cortex, 98,* 91-101. http://dx.doi.org/10.1016/j.cortex.2018.07.008

Belanova, E., Davis, J. P., & Thompson, T. (2019) *Holistic face processing in face recognition super-recognisers and typical-range-ability controls.* PsyArXiv. https://doi.org/10.31234/osf.io/mkhnd

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar-face matching. *Applied Cognitive Psychology*, *27*(6), 707-717. https://doi.org/10.1002/acp.2970

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex*, *82,* 48-62. http://dx.doi.org/10.1016/j.cortex.2016.05.003

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PLOS One, 11*(2), e0148148. https://doi.org/10.1371/journal.pone.0148148

Bobak, A. K., Hancock, P. J., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology*, *30*(1), 81-91. http://dx.doi.org/10.1002/acp.3170

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology*, *7,* 1378. http://dx.doi.org/10.3389/fpsyg.2016.01378

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental prosopagnosia and “super” face recognition. *The Quarterly Journal of Experimental Psychology*, *70*(2), 201-217. http://dx.doi.org/10.1080/17470218.2016.1161059

Bower, G. H., & Karlin, M. B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology*, *103*(4), 751-757. http://dx.doi.org/10.1037/h0037190

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353-364. http://10.1023%2FA%3A1015380522722

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*(1), 11-30. http://dx.doi.org/10.1037/1076-898X.12.1.11

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*(1), 105-116. http://dx.doi.org/10.1111/j.2044-8295.1982.tb01795.x

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339-360. http://dx.doi.org/10.1037/1076-898X.5.4.339

Bürkner, P. C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal, 10*(1), 395-411. http://dx.doi.org/10.32614/RJ-2018-017

Christensen, R. H. B. (2011). Analysis of ordinal data with cumulative link models—estimation with the R-package 'ordinal'. Available at http://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pdf.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 335-359. http://dx.doi.org/10.1016/S0022-5371(73)80014-3

Colloff, M., Seale-Carlisle, T., Karoğlu, N., Rockey, J., Smith, H. M., Smith, L., ... & Flowe, H. D. (2020). *Enabling witnesses to reinstate perpetrator pose during a lineup test increases accuracy.* PsyArXiv. https://10.31234/osf.io/2rwgh

Colloff, M.F., Flowe, H.D., Smith, H.M.J., Seale-Carlisle, T.M., Meissner, C.A., Rockey, J. C., Pande, B., Kujur, P., Parveen, N., Chandel, P., Singh, M.M., Pradhan, S., & Parganiha, A. (2020). *Active exploration of faces in police lineups increases discrimination accuracy for own- and other-race faces.* PsyArXiv. https://doi.org/10.31234/osf.io/tvga4

Craik, F. I. (2002). Levels of processing: Past, present... and future? *Memory*, *10*(5-6), 305-318. http://dx.doi.org/10.1080/09658210244000135

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671-684. http://dx.doi.org/10.1016/S0022-5371(72)80001-X

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12*(1), 41-55. http://dx.doi.org/10.1007/BF01064273

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. http://dx.doi.org/10.1002/acp.1490

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology*, *30*(6), 827-840. http://dx.doi.org/10.1002/acp.3260

Davis, J. P., Maigut, A., & Forrest, C. L. D. (2019). The wisdom of the crowd: A case of post- to ante-mortem face matching by police super-recognisers. *Forensic Science International*, 109910. https://doi.org/10.1016/j.forsciint.2019.109910

Estudillo, A. J., & Bindemann, M. (2014). Generalization across view in face memory and face matching. *i-Perception*, *5*(7), 589-601. http://dx.doi.org/10.1068/i0669

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410-8415. http://dx.doi.org/10.1073/pnas.1319030111

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*(9), 330-337. http://dx.doi.org/10.1016/S1364-6613(00)01519-9

Harper, T. (2016, October 23). Home Office ‘walked away from huge problem of face passports’. *The Times.* Retrieved from https://www.thetimes.co.uk/article/home-office-walked-away-from-huge-problem-of-fake-passports-9sjd0s9kk

Hill, H., Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition, 62*(2), 201-222. http://dx.doi.org/10.1016/S0010-0277(96)00785-8

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800-803. http://dx.doi.org/10.1093/biomet/75.4.800

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323. https://doi.org/10.1016/j.cognition.2011.08.001

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54-69. http://dx.doi.org/10. 1037/ a0028347

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning,* *Memory, and Cognition, 22*(5), 1304-1316. http://dx.doi.org/10.1037/0278-7393.22.5.1304

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*(3), 211-222. http://dx.doi.org/10.1002/(SICI)1099-0720(199706)11:3<211::AID-ACP430>3.0.CO;2-O

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research*, *43*(18), 1921-1936. https://doi.org/10.1016/S0042-6989(03)00236-0

Kramer, R. S., & Reynolds, M. G. (2018). Unfamiliar face matching with frontal and profile views. *Perception, 47*(4), 414-431. http://dx.doi.org/10.1177/0301006618756809

Lander, K., & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology*, *12*(4), 259-272. https://doi.org/10.1207/S15326969ECO1204_01

Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition*, *12*(3), 429-442. https://doi.org/10.1080/13506280444000382

Lander, K., Bruce, V., & Bindemann, M. (2018). Use-inspired basic research on individual differences in face identification: Implications for criminal investigation and security. *Cognitive Research: Principles and Implications*, *3*(1), 26. http://dx.doi.org/10.1186/s41235-018-0115-6

Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition*, *27*(6), 974-985. 10.3758/bf03201228

Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition*, *27*(6), 974-985. http://dx.doi.org/10.3758/BF03201228

Lindsay, R. C., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*(1), 79-89. http://dx.doi.org/10.1037/0021-9010.66.1.79

Liu, C. H., Ward, J., & Markall, H. (2007). The role of active exploration of 3D face stimuli on recognition memory of facial information. *Journal of Experimental Psychology: Human Perception and Performance, 33*(4), 895-904. doi:10.1037/0096-1523.33.4.895

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 77-100. http://dx.doi.org/10.1037/0096-1523.34.1.77

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865-876. http://dx.doi.org/10.3758/BF03193433

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*(4), 364-372. http://dx.doi.org/10.1037/a0013464

Menon, N., Kemp, R. I., & White, D. (2018). More than a sum of parts: robust face recognition by integrating variation. *Royal Society Open Science, 5*(5), 172381. https://doi.org/10.1098/rsos.172381

Menon, N., White, D., & Kemp, R. I. (2015). Variation in photos of the same face drives improvements in identity verification. *Perception*, *44*(11), 1332-1341. https://doi.org/10.1177/0301006615599902

Moore, R. M., & Johnston, R. A. (2013). Motivational incentives improve unfamiliar face matching accuracy. *Applied Cognitive Psychology, 27*(6), 754-760. https://doi.org/10.1002/acp.2964

Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a super-recogniser? In M. Bindemann & A. M. Megreya (Eds.), *Face Processing: Systems, Disorders, and Cultural Differences* (pp. 173-201). *Nova*.

Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia*, *45*(1), 75-92. http://dx.doi.org/10.1016/j.neuropsychologia.2006.04.025

Pike, G. E., Kemp, R. I., Towell, N. A., & Phillips, K. C. (1997). Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition, 4*(4), 409-438. https://doi.org/10.1080/713756769

Pike, G. E., Kemp, R. I., Towell, N. A., & Phillips, K. C. (1997). Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition*, *4*(4), 409-438. http://dx.doi.org/10.1080/713756769

Pilz, K. S., Thornton, I. M., & Bülthoff, H. H. (2006). A search advantage for faces learned in motion. *Experimental Brain Research*, *171*(4), 436-447. https://doi.org/10.1007/s00221-005-0283-8

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ritchie, K. L., White, D., Kramer, R. S., Noyes, E., Jenkins, R., & Burton, A. M. (2018). Enhancing CCTV: Averages improve face identification from poor-quality images. *Applied Cognitive Psychology*, *32*(6), 671-680. https://doi.org/10.1002/acp.3449

Robertson, D. J., Black, J., Chamberlain, B., Megreya, A., & Davis, J. P. (2019). Super-recognisers show an advantage for other race face identification. *Applied Cognitive Psychology*, *34*(1), 205-216. https://doi.org/10.1002/acp.3608

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A.M. (2016). Face recognition by Metropolitan Police super-recognisers. *PLOS, One*, *11*(2): e0150036. https://doi.org/10.1371/journal.pone.0150036

Rosnow, R. L., & Rosenthal, R. (1996). Contrasts and interactions redux: Five easy pieces. *Psychological Science*, *7,* 253-257. http://dx.doi.org/10.1111/j.1467-9280.1996.tb00369.x

Russell, R., Chatterjee, G., & Nakayama, K. (2012). Developmental prosopagnosia and super-recognition: No special role for surface reflectance processing. *Neuropsychologia*, *50*(2), 334-340. http://dx.doi.org/10.1016/j.neuropsychologia.2011.12.004

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252-257. http://dx.doi.org/10.3758/PBR.16.2.252

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. In T. Valentine & J. P. Davis (Eds.), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (pp. 185-208). Wiley-Blackwell.

Smith, H. M. J., Baguley, T. S., Robson, J., Dunn, A. K., & Stacey, P. C. (2019). Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance. *Applied Cognitive Psychology*, *33*(2), 272-287. https://doi.org/10.1002/acp.3478

Smith, H. M., Baguley, T. S., Robson, J., Dunn, A. K., & Stacey, P. C. (2019). Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance. *Applied Cognitive Psychology*, *33*(2), 272-287. https://doi.org/10.3758/s13414-015-1045-8

Stephens, R. G., Semmler, C., & Sauer, J. D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence-accuracy relationship in unfamiliar face matching. *Journal of Experimental Psychology: Applied*, *23*(3), 336-353. http://dx.doi.org/10.1037/xap0000130

Thornton, I. M., & Kourtzi, Z. (2002). A matching advantage for dynamic human faces. *Perception*, *31*(1), 113-132. https://doi.org/10.1068/p3300

Weber, N., Woodard, L., & Williamson, P. (2013). Decision strategies and the confidence–accuracy relationship in face recognition. *Journal of Behavioral Decision Making*, *26*(2), 152-163. http://dx.doi.org/10.1002/bdm.1750

White, D., Burton, A. L., & Kemp, R. I. (2016). Not looking yourself: The cost of self-selecting photographs for identity verification. *British Journal of Psychology, 107*(2), 359-373. http://dx.doi.org/10.1111/bjop.12141

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS one*, *9*(8), e103510. http://dx.doi.org/10.1371/journal.pone.0103510

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10-65. http://dx.doi.org/10.1177/1529100616686966

Wright, D.B., Horry, R. & Skagerberg, E.M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods, 41*(2), 257-267. http://dx.doi.org/10.3758/BRM.41.2.2

Appendix A: Experiment 1

Accuracy

Effects of individual predictors were compared using likelihood ratio tests (G^2) comparing a model without the effect to one including all effects of the same order (e.g., all main effects or all two-way interactions).[3] We first fitted an intercept only model (with no predictors but with random effects for participants and faces), the estimate of the *SD* of the Facebook image random effect was 0.33, for the comparison image random effect it was 1.07, and for the participant effect it was 0.47 (changing to 0.37, 1.04 and 0.46 for a full 2 x 4 factorial model). This indicates that most of the variation at level 2 of the model is attributable to the variability in the comparison faces (77%) rather than the Facebook faces (7%) and participants (15%). It also indicates that conventional analyses ignoring variability among stimuli would greatly inflate Type I error rates (e.g., see Baguley, 2012; Judd, Westfall, & Kenny, 2012).

The identity by image condition interaction is shown in Figure 2 and suggests that the interaction is largely explained by the advantage for same ID over different ID found in the non-interactive conditions being reversed for the interactive condition. The adequacy of this explanation can be illustrated by computing $r_{alerting}$ – the correlation between coefficients for an interaction contrast capturing this pattern and the cell means (Rosnow & Rosenthal, 1996; Abelson & Prentice, 1997; Baguley, 2012). The correlation is .845, indicating that the reversal of the same-different advantage for the interactive condition accounts for 71.5% of the variation in the cells means for the interaction.[4]

[3] This is analogous to tests of effects in ANOVA using Type II sums of squares (and equivalent to Type III sums of squares in a balanced design).

[4] For a logistic regression we could compute this correlation for the log odds, odds or predicted probabilities. The log odds scale (used here) is arguably the most appropriate, but in practice the choice makes little difference. The correlation on the odds scale is slightly weaker and that on the probability scale slightly stronger.

Multilevel signal detection analysis

Initial analyses produced warnings that the estimate of the Facebook image variance was close to zero. For this reason, we refitted the model using the R Bayesian regression brms (Bürkner, 2018) using default priors (which should produce estimates similar to those for frequentist estimation in lme4). Although estimates for fixed effects were similar between the two packages, we adopt the cautious approach of reporting estimates and inferences from the posterior distributions obtained from brms.

The random effects for the reported analysis are presented in Table A1. There is a reliable positive correlation between criterion and *d'* at the participant level, suggesting that participants who are better at discriminating matches from non-matches, also adopt higher thresholds.

Table A1

Multilevel Signal Detection Analysis: Random effects for Criterion and D Prime (d'), Experiment 1.

	Criterion			*d′*		
Condition	*Estimate*	*SE*	95% CI	*Estimate*	*SE*	95% CI
FB face						
SD	0.274	0.209	0.023, 0.209	0.455	0.284	0.049, 0.944
Correlation	.417	0.529	-.703, .959			
Comp face						
SD	0.575	0.176	0.166, 0.772	0.434	0.332	0.032, 1.065
Correlation	.362	.504	-.663, .948			
Participants						
SD	0.611	0.033	0.561, 0.666	0.522	0.035	0.467, 0.580
Correlation	0.563	0.059	.465, .652			

The relationship between confidence and accuracy

As recommended in the face recognition literature (e.g., Brewer & Wells, 2006; Juslin, Olsson, & Winman, 1996; Weber, Woodard, & Williamson, 2013), Stephens et al. (2017) used calibration analysis, which involves calculating the proportion of correct responses for each level of confidence. However, in face matching experiments, participants complete a series of trials, and therefore encounter multiple stimuli. A criticism of using calibration to analyse data from multiple trial experiments such as face matching is that it involves aggregating data, and therefore ignoring variability at the stimulus level (see Smith, Baguley, Robson, Dunn, & Stacey, 2019). This source of variability may however be relevant to the results (Clark, 2012; Judd, Westfall, & Kenny, 2012). If some people look more similar across images than other people do, failing to take this into account will reduce generalisability.

Appendix B: Experiment 2

Accuracy

In the intercept only model (with no predictors, but with random effects for participants and faces), the estimate of the *SD* of the Facebook image random effect was 0.45, for the comparison image random effect it was 1.11, and for the participant effect it was 0.77 (changing to 0.55, 1.02 and 0.71 for a full 2 x 4 factorial model). The pattern of *SD*s for the random effects was similar to Experiment 1. Again, most of the variation at level 2 of the model (61%) is attributable to the variability in the comparison faces.

The interaction between identity and image type appears to reflect a decrease in the difference in accuracy between same and different trials from the static frontal and orientation conditions compared to the non-static moving and interactive conditions. As in Experiment 1 we computed r_{alerting} for a contrast reflecting this pattern of differences (large for the first two conditions and negligible for the other two conditions) and mean accuracy in each condition. Here r_{alerting} was .87 indicating that a pattern of large differences in the static conditions and negligible differences in the moving and interactive conditions accounted for 75% of variation in accuracy between conditions.

Multilevel signal detection analysis

The random effects are presented in Table B1. As in Experiment 1, there is a reliable positive correlation between criterion and *d’* at the participant level.

Table B1

Multilevel Signal Detection Analysis: Random effects for Criterion and D Prime (d'), Experiment 2.

	Criterion			*d′*		
Condition	*Estimate*	*SE*	95% CI	*Estimate*	*SE*	95% CI
FB face						
SD	0.300	0.226	0.023, 0.716	0.597	0.343	0.071, 1.158
Correlation	.420	0.517	-.668, .955			
Comp face						
SD	0.596	0.204	0.154, 0.849	0.459	0.349	0.037, 1.113
Correlation	.222	.546	-.783, .932			
Participants						
SD	0.778	0.110	0.619, 0.969	0.558	0.118	0.381, 0.752
Correlation	0.314	0.200	-.025, .613			

Appendix C: Experiment 3

Accuracy

In the intercept only model (with no predictors but with random effects for participants and faces), the estimate of the *SD* of the Facebook image random effect was 0.20, for the comparison image random effect it was 1.11, and for the participant effect it was 0.65 (changing to 0.19, 1.11 and 0.49 for a full 2 x 2 x 2 factorial model). The pattern of *SD*s for the random effects was similar to Experiment 1 and 2. Once again, the majority of variation at level 3 of the model is attributable to the variability in the comparison faces.

Multilevel signal detection analysis

The random effects are presented in Table C1. As in Experiment 1 and 2, there is a reliable positive correlation between criterion and *d'* at the participant level.

Table C1

Multilevel Signal Detection Analysis: Random effects for Criterion and D Prime (d'), Experiment 3.

	Criterion			*d′*		
Condition	*Estimate*	*SE*	95% CI	*Estimate*	*SE*	95% CI
FB face						
SD	0.239	0.180	0.020, 0.581	0.491	0.276	0.063, 0.942
Correlation	.260	0.532	-.755, .914			
Comp face						
SD	0.543	0.148	0.232, 0.720	0.454	0.330	0.041, 1.080
Correlation	.130	.549	-.819, .906			
Participants						
SD	0.626	0.048	0.552, 0.706	0.548	0.051	0.470, 0.632
Correlation	0.350	0.101	.182, .503			

Appendix D: Experiment 4

Accuracy

In the intercept only model (with no predictors but with random effects for participants and faces), the estimate of the *SD* of the Facebook image random effect was 0.46, for the comparison image random effect it was 1.06, and for the participant effect it was 0.45 (changing to 0.50, 1.03 and 0.37 for a full 2 x 2 x 2 factorial model). As in Experiments 1, 2 and 3, the majority of variation at level 3 of the model is attributable to the variability in the comparison faces.

Multilevel signal detection analysis

The random effects are presented in Table D1. As in Experiment 1, 2 and 3 there is a reliable positive correlation between criterion and *d'* at the participant level.

Table D1

Multilevel Signal Detection Analysis: Random effects for Criterion and D Prime (d'), Experiment 4.

	Criterion			*d′*		
Condition	*Estimate*	*SE*	95% CI	*Estimate*	*SE*	95% CI
FB face						
SD	0.332	0.233	0.011, 0.784	0.571	0.318	0.032, 1.168
Correlation	.534	0.479	-.776, .986			
Comp face						
SD	0.581	0.200	0.055, 0.854	0.450	0.336	0.017, 1.220
Correlation	.373	.509	-.851, .978			
Participants						
SD	0.573	0.038	0.503, 0.651	0.451	0.038	0.379, 0.529
Correlation	0.552	0.073	.395, .681			