Article:

Ritchie, Kay L., White, David, Kramer, Robin S.S. et al. (3 more authors) (2018) Enhancing CCTV:Averages improve face identification from poor-quality images. Applied Cognitive Psychology. pp. 671-680. ISSN: 0888-4080

https://doi.org/10.1002/acp.3449

Enhancing CCTV: Averages improve face identification from poor quality images

Journal:	*Applied Cognitive Psychology*
Manuscript ID	ACP-17-0177.R2
Wiley - Manuscript type:	Research Article
Keywords:	face identification, averages, pixelated images, CCTV

Abstract

Low quality images are problematic for face identification, for example when police identify faces from CCTV images. Here we test whether face averages, comprising multiple poor quality images, can improve both human and computer recognition. We created averages from multiple pixelated or non-pixelated images, and compared accuracy using these images and exemplars. To provide a broad assessment of the potential benefits of this method, we tested human observers (n = 88; Experiment 1), and also computer recognition, using a smartphone application (Experiment 2) and a commercial one-to-many face recognition system used in forensic settings (Experiment 3). The third experiment used large image databases of 900 ambient images and 7980 passport images. In all three experiments, we found a substantial increase in performance by averaging multiple pixelated images of a person's face. These results have implications for forensic settings in which faces are identified from poor quality images, such as CCTV.

Key words: Face identification, averages, pixelated images, CCTV.

Introduction

Police forces use CCTV images for suspect identification, and this process can utilise both human operators and computer face recognition systems. It is important, therefore, to understand the effect of poor quality images on both human and computer performance. Our goal here is to test a quick and easy method of image enhancement, namely averaging, to establish whether this can improve face recognition from poor quality images for both human observes and computer systems.

Although human observers are accurate in identifying familiar people from poor quality CCTV footage (Burton, Wilson, Cowan & Bruce, 1999), studies have shown that accuracy in identifying unfamiliar people from CCTV is poor (Bruce et al., 1999; Davies & Thasen, 2000; Davis & Valentine, 2009; Walker & Tough, 2015). Pixelation also harms the ability to identify familiar people from both static and moving images (Lander, Bruce & Hill, 2001), and can completely extinguish this ability at very high levels of pixelation (Demanet, Dhont, Notebaert, Pattyn & Vandierendonck, 2007). As the quality of the CCTV is reduced due to image compression, the ability to make face identifications from the videos decreases (Keval & Sasse, 2008). Recently, however, it has been shown that experts such as forensic facial examiners are able to overcome this problem to some extent (White, Phillips, Hahn, Hill & O'Toole, 2015), but their expertise is most advantageous when working with high quality images (Norell et al., 2015; White, Norell, Phillips & O'Toole, 2017).

A recent study examined performance on a face matching task in which participants were required to indicate whether two simultaneously presented images showed the same person or two different people. When one image in the face pair was pixelated, face matching performance was surprisingly robust, only dropping below chance level with images presented at a resolution of 8 pixels in width (Bindemann, Attard, Leach & Johnston, 2013). At a level of pixelation which reduced performance, but not as low as chance, performance was significantly improved by reducing the size of the pixelated image, thus reducing the perceptual effect of the large-scale edge information in the image.

Computer recognition of faces as assessed with standard evaluation measures such as the FERET (Phillips, Moon, Rizvi & Rauss, 2000) and the FRVT (Blackburn, Bone & Phillips, 2001) typically outperforms human unfamiliar face recognition (O'Toole et al., 2007) but does not perform perfectly (O'Toole et al., 2007; Phillips, Flynn, Scruggs, Bowyer & Worek, 2006; Zhao, Chellappa, Phillips & Rosenfeld, 2003). Direct comparisons of humans and face recognition algorithms have shown that, although algorithms outperform humans on frontal face images (Phillips & O'Toole, 2014), for images showing extreme illumination and pose, humans win out against computer algorithms (Phillips, Hill, Swindle & O'Toole, 2015).

Recent work in the field of computer science has utilised a variety of techniques such as noise suppression and super-resolution, in an attempt to overcome the harmful effects of poor image quality on computer face recognition, achieving various degrees of success (Buciu & Gacsadi, 2011; Rudrani & Das, 2011). To date, these techniques have only been applied to images in such a way as to test for improvements in machine recognition. Other techniques seek to assess image quality and improve face recognition performance by simply rejecting images which fall below a given threshold, but this is problematic because there is no agreement on a reliable indicator of quality (Luo, 2004; Fronthaler, Kollreider & Bigun, 2006; Beveridge et al., 2011). Moreover, in some situations poor quality images may be all that is available, for example when poor quality CCTV footage is the only evidence linking a suspect to a crime scene.

Here we address this problem by examining whether combining information across multiple poor quality images can benefit human and computer matching accuracy. In applied settings, multiple images of a person are often available, for example multiple screenshots from CCTV footage. We focus on one promising approach that has been shown to improve both human and computer matching - averaging together multiple images of a single identity, as shown in Fig 1 (Burton, Jenkins, Hancock & White, 2005; Jenkins & Burton, 2008; White, Burton, Jenkins & Kemp, 2014). In a prior study, images of celebrities were uploaded to an online implementation of an industry standard face recognition system (FaceVACS). Accuracy of identification of exemplars was only 54%, climbing to 100% for average images (Jenkins & Burton, 2008). A subsequent study showed that the automatic face recognition algorithm used in Android smartphone devices' "face unlock" system was improved from 45% for single images to 68% for averages (Robertson, Kramer & Burton, 2015). One study has also shown that average images also improve human accuracy for face matching tasks (White et al., 2014). Accuracy for matching an average of 12 images of an individual to one exemplar image was higher than accuracy for matching two exemplars.

Figure 1 here

Averaging together multiple pixelated images from CCTV footage, for example, ought to reduce the noise introduced by the pixelation, and lead to a clearer representation of the identity. Simply by taking multiple low resolution images whose noise is uncorrelated, and averaging them together in a high resolution space, one increases the amount of information present by comparison to a single image. Here, we apply the technique of face averaging to the problem of face identification from poor quality images. We present three experiments investigating the effect of averaging multiple degraded images in order to produce a better representation of the person pictured. The first experiment tests human face matching, the second experiment uses a smartphone app, available to the general public, and the final experiment tests a commercial face recognition application, currently used in the security industry. The final experiment also uses a large number of images in two different databases – an ambient image database of 900 images from the *labelled faces in the wild* set (Huang, Ramesh, Berg & Learned-Miller, 2007), and images taken from an existing database of 7980 real passport images.

Experiment 1. Human face matching

This experiment investigates the effect of pixelation and averaging on human face matching performance. In a face matching task, participants are shown two images simultaneously and asked to decide whether or not they show the same person. A recent study found that pixelating one of the two images in a matching task reduces performance (Bindemann et al., 2013). Here, we averaged together multiple pixelated images to establish whether averages would give rise to higher accuracy than single pixelated images. We hypothesised that unfamiliar face matching accuracy will be poorer for pixelated than unpixelated images, and that averages of pixelated images would produce an increase in accuracy compared to pixelated exemplars.

Method

Participants

Eighty-eight participants took part in this experiment (16 males; mean age: 24 years, range: 18-65 years). All were members of the University of York, UK, or the University of Lincoln, UK, and took part voluntarily or in exchange for course credit. This study was approved by the Ethics Committee of the Department of Psychology, University of York and the School of Psychology Research Ethics Committee at the University of Lincoln. All participants gave written informed consent.

Stimuli and Procedure

Eleven images of 96 different unfamiliar identities (50% women) were downloaded from the Internet using Google Image searches for celebrities from different countries, and were selected in order to be unfamiliar to our UK-based participants. Familiarity checks on a different group of participants (not tested in the current studies) confirmed the IDs were unfamiliar to UK viewers. Images were broadly full-facing, but sampled natural variability in facial and environmental parameters, akin to those used in previous face matching research (Ritchie et al., 2015). In addition, for each identity, one 'foil' image was collected. This was an image of another unfamiliar identity (not appearing in the original 96) matching the verbal description of the target identity. The images were high quality, and cropped to 380x570 pixels. Each of these images was also downsampled to size 30x45 pixels and then resized back to their original dimensions. This method provided pixelated and unpixelated versions of the image set.

We created average images by initially deriving the shape of each image using a semi-automatic landmarking system designed to register 82 points on the face aligned to anatomical features. Each average was created by warping the 10 images of an identity to the average shape of those 10 images, and then calculating the mean RGB colour values for each pixel. The unpixelated images were landmarked using our semi-automatic system (where only five locations are selected manually – for details, see Kramer, Young, Day & Burton, 2017). After pixelation, the images were again landmarked using the system. Therefore, landmarking of the pixelated images was inherently less precise, given that our system (and the human user selecting the five locations) had far less photographic detail to work with.

Ten images of each identity, unpixelated and pixelated, were used to form averages, with the one excluded image used as the 'match' image. Note that 'pixelated averages' are therefore averages of pixelated images, not averages created and then themselves pixelated. The 'mismatch' image was the foil collected previously for that identity. Due to the procedure used for creating the averages, all background information was removed from the average images. Therefore, to ensure that reference exemplar images were consistent with the averages, all background information was also removed from reference exemplars. Match and foil images were presented naturally with background information intact (see Fig 2).

Figure 2 here

Each trial consisted of the reference image (unpixelated exemplar, pixelated exemplar, average of unpixelated images, average of pixelated images) presented on the left of the screen, and the test image (match or foil) presented on the right. Each participant saw each ID once in the experiment, with each ID counterbalanced by condition across participants. There were 12 trials per condition (always 50% women).

Results and Discussion

Fig 3 shows mean accuracy for the human face matching task. Following previous research (White et al., 2014), we analysed the data for match and mismatch trials separately, using a 2 (image type: exemplar, average) x 2 (pixelation: unpixelated, pixelated) ANOVA.

For match trials, there was a significant main effect of image type ($F(1,87) = 35.00$, $p < .001$, $\eta_p^2 = .29$), a significant main effect of pixelation ($F(1,87) = 38.84$, $p < .001$, $\eta_p^2 = .31$), and a significant interaction between image type and pixelation ($F(1,87) = 4.11$, $p = .046$, $\eta_p^2 = .05$). We therefore considered the simple main effects of pixelation at each level of image type. These simple main effects were significant for both exemplars ($F(1,174) = 38.25$, $p < .001$, $\eta_p^2 = .18$) and averages ($F(1,174) = 13.90$, $p < .001$, $\eta_p^2 = .07$), meaning that unpixelated exemplars and averages were more easily matched to the test image than pixelated exemplars and averages. We also considered the simple main effects of image type at each level of pixelation. These simple main effects were significant for both pixelated ($F(1,174) = 32.63$, $p < .001$, $\eta_p^2 = .16$) and unpixelated images ($F(1,174) = 8.71$, $p < .005$, $\eta_p^2 = .05$), meaning that averages outperformed exemplars for both image types. The effect size for the average advantage was much greater for pixelated than for unpixelated images, suggesting that image averaging is especially beneficial where image quality is low.

A 2 (image type: exemplar, average) x 2 (pixelation: unpixelated, pixelated) ANOVA on mismatch trials found a significant main effect of pixelation ($F(1,87) = 70.41$, $p < .001$, $\eta_p^2 = .45$), a non-significant main effect of image type ($F(1,87) = .26$, $p = .611$, $\eta_p^2 < .001$), and a non-significant interaction between image type and pixelation ($F(1,87) = .68$, $p = .412$, $\eta_p^2 = .01$). For mismatch trials, pixelated images gave rise to poorer performance than unpixelated images, but there was no effect of averaging. The result is in-line with the previous work on this topic (White et al., 2014), where averaging improved performance on match but not non-match trials.

Figure 3 here

Analysis of accuracy scores on match trials show that averages improve performance for both pixelated and non-pixelated images, with a greater effect of averaging for pixelated images. However, because this interaction was not observed in non-match trials, it may reflect a response bias. In order to clarify whether the interaction was driven by improvements in perceptual sensitivity, we analysed the results using a signal detection theory model. In this analysis, hits correspond to correct match trials and false alarms correspond to incorrect mismatch trials. Paired samples *t*-tests on d-prime (*d'*) values showed a significant difference between accuracy for pixelated exemplars ($M = .43$) and pixelated averages ($M = .80$), $t(87) = 3.797$, $p < .001$, $d = 0.41$, but a non-significant difference between accuracy for unpixelated exemplars ($M = 1.32$) and unpixelated averages ($M = 1.44$), $t(87) = 1.431$, $p = .156$, $d = 0.15$. Therefore, averaging improved sensitivity only for pixelated images and not for unpixelated images.

Paired samples *t*-tests on criterion (c) values showed a significant difference between the bias for unpixelated exemplars ($M = -.12$) and unpixelated averages ($M = .01$), $t(87) = 3.275$, $p = .002$, $d = 0.35$, and between the bias for pixelated exemplars ($M = -.10$) and pixelated averages ($M = .05$), $t(87) = 2.724$, $p = .008$, $d = 0.29$.Taken together, these results show that face averages comprising high quality images increased participants' bias to respond that two images show the same person, without increasing overall sensitivity.

Overall, the results of Experiment 1 show that accuracy on a face matching task is reduced when one image in the pair is pixelated. Averaging together several pixelated images, however, reduces this cost to performance. Further, the interaction between pixelation and averaging suggests that averaging is especially beneficial to human performance when image quality is poor. Creating face averages is computationally inexpensive and easy to achieve with various freely available softwares such as Psychomorph (Tidemann, Burt & Perrett,

2001) or InterFace (Kramer, Jenkins & Burton, 2017). We therefore suggest that this technique could be used in a variety of settings to improve human face matching.

While Experiment 1 addressed the effect of pixelation and averages on human face matching, we were also interested in establishing whether averaging can overcome difficulties associated with poor quality imagery in computer face recognition systems. In the following experiments, we turned our attention to testing the effect of image averaging with commercial face recognition software.

Experiment 2. Face recognition using a publicly available smartphone app

In this experiment, we tested a smartphone face recognition app with our pixelated images and averages. The use of automatic face recognition systems has rapidly increased in recent years to the point where these are commonly used in consumer electronics, for example as a security feature or as a means of organising personal photo albums. The developers of these systems typically do not publish the algorithms on which they operate as these are commercially sensitive. However, recognition accuracy is typically high, without being perfect, though performance is somewhat dependent on the quality of images. We therefore decided to test a contemporary, publicly available smartphone app. We expected the app to show reduced performance with pixelated photos – and we aimed to establish whether accuracy with these degraded images could be improved by averaging them.

We used the smartphone application *FaceDouble* version 1.0 (TeamSOA, Inc.) which is designed to return a celebrity lookalike for an image uploaded by the user. Following the procedure of a previous study (Jenkins & Burton, 2008) which used a similar face recognition app, we uploaded one celebrity face image at a time, to test whether the app would return an image of that same celebrity as the best 'lookalike'. This gives us the opportunity to use ambient, naturally-occurring images to test automatic recognition: A face is 'recognised' if the app returns an image of the same person as presented to it.

Method

We used 30 probe images of each of 10 Hollywood celebrities (5 female) selected from Google Images, used in previous research (Burton, Kramer, Ritchie & Jenkins, 2016). Images showed head and shoulders, and sampled natural variability. As in Experiment 1, the 30 original images of each identity were also pixelated from the original size of 380x570 pixels to 32x48 pixels (and then re-enlarged). This again gave us the same set of 30 unpixelated and pixelated images for each celebrity. We created 30 averages for each identity by randomly selecting 30 sets of 10 images to be averaged together (allowing overlap between sets/averages), repeating this process for unpixelated and pixelated image sets. Averages in each set were correspondent such that the first average of each set comprised the same 10 images (pixelated and unpixelated) and so on (see Fig 4 for example stimuli).

Each image was uploaded individually into the FaceDouble application on an Apple iPhone5 handset. When the returned identity matched that of the uploaded image, we recorded a 'hit'. Otherwise, we recorded a 'miss'. The app responds with a celebrity 'lookalike'. When the app returns the lookalike, it shows the celebrity's profile, as opposed to the closest matching image of that celebrity. Therefore it is not possible to eliminate identical picture returns as has been done previously (Jenkins & Burton, 2008). The image that the app uses in its profile of each celebrity was not included in our original sets of 30 images per celebrity.

Figure 4 here

Results and Discussion

Fig 5 shows the mean percent of correct identity responses from the smartphone app. A 2 (image type: exemplar, average) x 2 (pixelation: unpixelated, pixelated) ANOVA revealed a significant main effect of image type ($F(1,9) = 93.20$, $p < .001$, $\eta_p^2 = .91$), a main effect of pixelation ($F(1,9) = 77.36$, $p < .001$, $\eta_p^2 = .90$), and a significant interaction between image type and pixelation ($F(1,9) = 47.25$, $p < .001$, $\eta_p^2 = .84$). Simple main effects showed an effect of image type at both the unpixelated ($F(1,18) = 7.91$, $p < .01$, $\eta_p^2 = .31$) and the pixelated level ($F(1,18) = 139.22$, $p < .001$, $\eta_p^2 = .89$), meaning that averages outperformed exemplars both when the exemplars and the images comprising the average were unpixelated, and when they were pixelated. Simple main effects also showed an effect of pixelation for both exemplars ($F(1,18) = 123.77$, $p < .001$, $\eta_p^2 = .87$) and averages ($F(1,18) = 12.77$, $p < .005$, $\eta_p^2 = .42$), meaning that unpixelated exemplars and averages comprising unpixelated images led to higher accuracy in identity recognition than pixelated exemplars and averages comprising pixelated images.

Figure 5 here

These results show a number of interesting effects. First, the overall level of performance of the automatic recognition system is rather good. The system recognised 86% of celebrities' images in their raw (unpixelated) form. This is rather impressive performance, given the unconstrained nature of the images used – simply collected from internet search. Second, there is a considerable advantage to recognition of averages – as with previous research (Jenkins & Burton, 2008), the system recognised 100% of all averages of the celebrities tested.

As predicted, pixelation severely damaged the recognition rates of the automatic system, with performance dropping to a quarter of that of the original images (22% accuracy). However, this drop in performance was almost entirely overcome by averaging the pixelated images together. In this case, we see performance of standard images (at 86% in Fig 5) being almost equalled by the simple graphical manipulation on very severely degraded pixelated images (79% in Fig 5). This is a very impressive performance boost for the automated recognition system.

The results of this experiment are promising, in that it appears a simple averaging procedure can enhance automatic recognition of poor quality images. However, from this single experiment, we cannot judge whether the result will generalise to other automated systems. Furthermore, we had no control over the database of images used for matching, and so we do not know whether the results are dependent on the type of images available for internet searches on celebrities. In the next experiment, we tested a rather different face recognition system, designed for forensic and security purposes rather than for consumer electronics. This allowed us to control the composition of the image database and extract more detailed performance measures, as described below.

Experiment 3. Commercial face recognition system and large image databases

Here, we test the benefit of image averaging using a commercially available face recognition system. We had the opportunity to test the effectiveness of our averaging technique using *FaceVACS-DBScan 5.1.2.0* running Cognitec's B10 algorithm (Cognitec, 2017) which compares a face image to a large image database. We created two large image databases: an ambient image database comprising 900 celebrity images from the ‘labelled faces in the wild’ set (Huang, Ramesh, Berg & Learned-Miller, 2007); and a passport image database comprising 7980 passport images of Australian citizens. The *ambient image database* comprised images captured in unconstrained environmental conditions, typically taken by photojournalists. Here, we use this database to simulate the type of imagery commonly found in forensic casework. The *passport image database* simulates the type of imagery stored in databases of secure identity documents, which may be accessed in the course of forensic casework (Grother & Ngan, 2014; Garvie, Bedoya & Frankle, 2016).

We added ten ambient images of each of our target celebrities to the ambient image database, and two passport-compliant images of each of the target celebrities to the passport image database. We used these databases to test our averaging technique by entering our experimental stimuli (i.e., unpixelated exemplars, unpixelated averages, pixelated exemplars, and pixelated averages) as probe images, and recorded hits when the system returned the same identity from the database.

Method

We evaluated the effectiveness of the averaging technique using two large test databases. The *ambient image database* consisted of 1000 images, one image each of 900 identities (450 female), taken from the ‘labelled faces in the wild’ set that has been used in recent benchmark tests of automatic face recognition software (Huang et al., 2007). We ensured that the images of the 900 non-matching identities in this dataset did not duplicate any of the target celebrities. We added 100 images of the target celebrities (10 images of each) to the database. So as to keep these images consistent with the other images in the database, we sourced them from the internet using the same collection method as described in the paper accompanying the original database (Huang et al., 2007), and cropped them to 250 x 250 pixels to be the same size as the database images (Fig 6A). The database images of our target celebrities were not included in our original image set for each identity, ensuring that there could not be identical image matches, and the database images did not contribute to any of our averages.

The *passport image database* comprised 8000 images. Non-matching images in this database were one passport photograph each of 7980 Australian citizens selected to be of a similar age to the target celebrities (i.e., between ages of 30 and 60). We added two images of each of the 10 target celebrities. So as to keep these images as consistent as possible with the database images, we selected these to be compliant with passport photo guidelines (front-facing, background removed; see Fig 6B). We divided the test database into 3990 male and 3990 female identities and conducted tests of male and female probe images separately.

Figure 6 here

The probe images used to search the databases in Experiment 3 were 10 images of each of the 10 celebrities in each image type (unpixelated exemplar, unpixelated average, pixelated exemplar, pixelated average). This resulted in a total of 400 probe images. These were a subset of the images used in Experiment 2.

Results and Discussion

We compared matching accuracy for the four probe image types using the following procedure. First, we counted how many times out of 100 probe images a target image of the correct identity was returned by the algorithm as the top ranking match. For the *ambient image database,* 99/100 unpixelated exemplars resulted in matches at rank 1, 100/100 unpixelated averages, 76/100 pixelated exemplars, and 96/100 pixelated averages. For the *passport image database*, the total of 98/100 unpixelated exemplar probe images, 100/100 unpixelated averages, 68/100 pixelated exemplars and 97/100 pixelated averages returned an image of the correct identity at rank 1.

The rank 1 position results show a pattern consistent with previous experiments. Face identification for unpixelated images was very high, but pixelating these images reduced performance by around a quarter. Averaging improved performance to 100% in the unpixelated condition, but more markedly in the pixelated condition, averaging poor quality images together produced performance equivalent to unpixelated single images.

Next, we counted how many of the 10 target images of the correct identity appeared in the top N ranked images returned by the system, the ‘candidate list’, for each of the 100 probe images in each condition. We repeated this analysis for 5 levels of candidate list size (10, 20, 40, 80, 160). This test protocol reflects the operation of algorithms configured for 1:n database search. In operational scenarios, the top N ranked match images are shown to a human reviewer who must inspect the images and decide if the target identity appears in this image gallery (White, Dunn, Schmid & Kemp, 2015; Grother & Ngan, 2014). Therefore here, the number of correct images of the target identity returned to the gallery represents the performance of the system across different levels of algorithm threshold. For the *ambient image database*, the maximum number of hits per probe was 10 and for the *passport image database*, the maximum number of hits was 2.

Figure 7 here

Fig 7 shows the mean number of hits for each probe image type as a function of gallery size for both the Ambient Image and Passport Image test sets. It is clear that results replicate the pattern found in previous experiments. Averaging improved performance of the recognition software for both pixelated and original images, and this benefit was largest for pixelated images.

For consistency with analysis of previous experiments, we conducted 2 (image type) x 2 (pixelation) ANOVAs separately for ambient image and passport image database tests. A single ANOVA was conducted for each test, collapsing over levels of gallery size. For both tests, there was a significant main effect of image type (ambient: $F(1, 99) = 179.20$, $p < .001$, $\eta_p^2 = .64$; passport: $F(1, 99) = 20.52$, $p < .001$; $\eta_p^2 = .17$), pixelation (ambient: $F(1, 99) = 477.30$, $p < .001$, $\eta_p^2 = .83$; passport: $F(1, 99) = 31.78$, $p < .001$, $\eta_p^2 = .24$) and a significant interaction between factors (ambient: $F(1, 99) = 104.71$, $p < .001$, $\eta_p^2 = .51$; passport: $F(1, 99) = 16.58$, $p < .001$, $\eta_p^2 = .14$). Analysis of simple main effects showed that averaging benefited accuracy for both unpixelated and pixelated images with the ambient image database (unpixelated: $F(1, 198) = 7.64$, $p < .01$, $\eta_p^2 = .04$, pixelated: $F(1, 198) = 281.04$, $p < .001$, $\eta_p^2 = .59$). For the passport image database, averaging benefited accuracy for pixelated ($F(1, 198) = 37.09$, $p < .001$, $\eta_p^2 = .16$) but not unpixelated probe images ($F(1, 198) = 0.47$, $p = .494$, $\eta_p^2 < .001$). Simple main effects also showed a significant detrimental effect of pixelation for both exemplars and averages for the ambient image database (exemplars: $F(1, 198) = 532.21$, $p < .001$, $\eta_p^2 = .73$, averages: $F(1, 198) = 87.39$, $p < .001$, $\eta_p^2 = .31$). Finally, simple main effects showed a significant detrimental effect of pixelation for both exemplars and averages for the passport image database (exemplars: $F(1, 198) = 48.10$, $p < .001$, $\eta_p^2 = .20$, averages: $F(1, 198) = 7.04$, $p < .01$, $\eta_p^2 = .03$).

Thus, results of Experiment 3 replicate the findings of the previous experiments; showing that averaging improves face matching performance, especially when averaging low resolution, pixelated images. The fact that averaging did not benefit performance for unpixelated probe images in the passport image database appears to be due to the ceiling level accuracy on this portion of the test.

The databases used in this experiment were intended to simulate those used in real forensic face identification settings. The results produced in the experiments here were conducted by the researchers, and should therefore not be construed as a maximum-effort full-capacity result. In practice, it is unlikely that a database would include more images of the target identity than non-matching identities as our databases did here. Nonetheless, this experiment goes some way to simulating the real-world problem of identifying a suspect from low quality CCTV images when provided with a database of high quality previously-collected images. The results show that averaging together multiple independent, poor quality images may provide a better representation of the suspect for use in automatic face recognition systems. In practice, many of the systems used in real-world settings have a front-end where investigators can manipulate images. Based on our current results, we would suggest that averaging could be built into these systems at this initial stage in order to improve accuracy for pixelated images.

General Discussion

In all three experiments, recognition of pixelated images was worse than unpixelated originals. Pixelation, at the resolutions tested here, is clearly detrimental to recognition. Further, we have presented a method for overcoming this by averaging together multiple pixelated images. In all three experiments, averages of pixelated images outperformed pixelated exemplars. The first experiment tested unfamiliar human observers, the second used a publicly available smartphone app, and the third investigated a commercially available face recognition system. These three methods mimic the real world settings of automatic and human face recognition from poor quality images such as face recognition algorithms used by police, and suspect identification from poor quality images.

Each of these three methods were sensitive to our manipulations of pixelation and averaging, and show broadly similar patterns of results. In Experiments 2 and 3, we have shown that the accuracy of two different implementations of automatic face recognition systems can be improved by using the average of multiple pixelated images. For the automatic systems, average images outperformed single exemplars, and the averages of unpixelated exemplars gave rise to near-perfect accuracy. In Experiment 1, we tested human observers on a face matching task using pixelated and unpixelated exemplars and their averages. Performance was poorer for pixelated than unpixelated exemplars, with a greater increase in accuracy when averaging was applied to pixelated images compared to individual exemplars.

Pixelation is often used as a method of masking identity for privacy purposes (Boyle, Edwards & Greenberg, 2000; Kitahara, Kogure & Hagita, 2004; Padilla-López, Chaaraoui & Flórez-Revuelta, 2015). It has been shown, however, that the effect of pixelation can be overcome by various computer algorithms so as to achieve accurate face identification from individual pixelated images (Newton, Sweeney & Malin, 2005) and when comparing a de-pixelated image to a very similar high quality image of the same person (Gross, Sweeney, De la Torre & Baker, 2006). The averaging technique we have used here provides a computationally inexpensive route to improving identification from pixelated images, provided that multiple images are available. Our results provide further evidence to suggest that pixelation is not a reliable form of image redaction for masking identity, in cases where multiple images are available.

The results of this study have clear and important implications for face identification in applied settings, particularly where automatic face recognition algorithms are in use. In settings such as police identification of suspects, it is common to compare a poor quality image to a database of high quality images using face recognition software. From the results of the experiments presented here, we suggest that creating an average of several poor quality images which have been obtained from different sources may improve face identification performance. We have also shown that this technique improves human face matching performance, which adds to a growing literature showing that image averaging can improve identification accuracy (e.g. Burton et al. 2005; Bruce, Ness, Hancock, Newman, & Rarity, 2002; Frowd, Bruce, Plenderleith, & Hancock, 2006; Hasel & Wells, 2007, White et al. 2014).

We have shown that averaging improves machine and human face identification, especially when image quality is low. These findings have implications for law enforcement where suspects are often identified from poor quality images. The face averaging method we have used is computationally inexpensive, easy to achieve, and yields clear benefits for both human and computer face recognition.

References

Beveridge, J. R., Phillips, P. J., Givens, G. H., Draper, B. A., Teli, M. N., & Bolme, D. S. (2011). When high-quality face images match poorly. *IEEE Proceedings of the International Conference on Automatic Face and Gesture Recognition,* 572-578.

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar-face matching. *Applied Cognitive Psychology, 27,* 707-717.

Blackburn, D., Bone, J. M., & Phillips, P. J. (2001). FRVT 2000 Evaluation Report. Technical report. 2001. Available from: http://www.frvt.org Accessed 5/5/2017

Boyle, M., Edwards, C., & Greenberg, S. (2000). The effects of filtered video on awareness and privacy. *Proceedings of the ACM Conference on Computer Supported Cooperative Work,* 1–10.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*(4), 339-360.

Bruce, V., Ness, H., Hancock, P. J. B., Newman, C., & Rarity, J. (2002). Four heads are better than one. Combining face composites yields improvements in face likeness. *Journal of Applied Psychology, 87*, 894-902.

Buciu, I., & Gacsadi, A. (2011). Noise suppression methods for low quality images with application to face recognition. *IEEE Proceedings ELMAR,* 21-24.

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51(3),* 256-284.

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40(1),* 202-223.

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

Cognitec FaceVACS DBScan. 2017. Available from: http://www.cognitec.com/facevacs-dbscan.html Accessed 1/8/2016

Davies, G., & Thasen, S. (2000). Closed-circuit television: How effective an identification aid? *British Journal of Psychology, 91,* 411–426.

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

Demanet, J., Dhont, K., Notebaert, L., Pattyn, S., & Vandierendonck, A. (2007). Pixelating familiar people in the media: Should masking be taken at face value? *Psychologica Belgica, 47*(4), 261-276.

Fronthaler, H., Kollreider, K., & Bigun, J. (2006). Automatic image quality assessment with application in biometrics. *IEEE Conference on Computer Vision and Pattern Recognition,* 30-35.

Frowd, C. D., Bruce, V., Plenderleith, Y., & Hancock, P. J. B. (2006). Improving target identification using pairs of composite faces constructed by the same person*. IEE Conference on Crime and Security*, 386-395, IET: London.

Garvie, C., Bedoya, A., & Frankle, J. (2016). The perpetual line-up: Unregulated police face recognition in America. Available from: http://www.perpetuallineup.org Accessed 5/5/2017

Gross, R., Sweeney, L., De la Torre, F., & Baker, S. (2006). Model-based face de-identification. *IEEE Conference on Computer Vision and Pattern Recognition Workshop,* 161–168.

Grother, P., & Ngan, M. (2014). Face Recognition Vendor Test (FRVT). Performance of Face Identification Algorithms. NIST, Interagency Report 8009.

Hasel, L. E., & Wells, G. L. (2007). Catching the bad guy: Morphing composite faces helps. *Law and Human Behavior, 31,* 193-207.

Huang, G., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical Report No: 07-49.

Jenkins, R., Burton, A. M. (2008). 100% accuracy in automatic face recognition. *Science, 319,* 435.

Keval, H., & Sasse, M. A. (2008). Can we ID from CCTV? Image quality in digital CCTV and facial identification performance. *Proceedings of SPIE International Society for Optical Engineering*, 6982.

Kitahara, I., Kogure, K., & Hagita, N. (2004). Stealth vision for protecting privacy. *IEEE Proceedings of the 17th International Conference on Patter Recognition,* 404–407.

Kramer, R. S. S., Young, A. W., Day, M. G., & Burton, A. M. (2017). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review, 124*(2), 115-129.

Kramer, R. S. S., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods, 49*(6), 2002-2011.

Lander, K., Bruce, V., & Hill, H. (2001). Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology, 15*, 101-116.

Luo, H. (2004). A training-based no-reference image quality assessment algorithm. *IEEE Proceedings International Conference on Image Processing,* 2973-2976.

Newton, E., Sweeney, L., & Malin, B. (2005). Preserving privacy by de-identifying facial images. *IEEE Transactions on Knowledge and Data Engineering, 17(2),* 232–243.

Norell, K., Lathen, K. B., Bergstrom, P., Rice, A., Natu, V., & O'Toole, A. (2015). The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Science, 60,* 331–340.

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

Padilla-López, J. R., Chaaraoui, A. A., & Flórez-Revuelta, F. (2015). Visual privacy protection methods: A survey. *Expert Systems with Applications, 42(9),* 4177-4195.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., & Worek, W. (2006). Preliminary face recognition grand challenge results. *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition,* 15-24.

Phillips, P. J., Hill, M. Q., Swindle, J. A., & O'Toole, A. J. (2015). Human and algorithm performance on the PaSC face recognition challenge. *IEEE 7th International Conference on Biometrics, Theory, Applications and Systems,* 1-8.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 1090-1104.

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32,* 74-85.

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition, 141,* 161-169.

Robertson, D. J., Kramer, R. S. S., & Burton, A. M. (2015). Face averages enhance user recognition for smartphone security. *PLoS One, 10(3),* e0119460.

Rudrani, S., & Das, S. (2011). Face recognition on low quality surveillance images by compensating degradation. *Image Analysis and Recognition, 6754,* 212-221.

Tidemann, B., Burt, M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications, 21(5),* 42-50.

Walker, H., & Tough, A. (2015). Facial comparison form CCTV footage: The competence and confidence of the jury. *Science & Justice, 55,* 487-498.

White, D., Burton, A. M., Jenkins, R., & Kemp, R. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied, 20(2),* 166-173.

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PloS One, 10(10),* e0139827.

White, D., Norell, K., Phillips, P. J., & O'Toole, A. J. (2017). Human factors in forensic face identification. In: M. Tistarelli, & C. Champod (Eds.), Handbook of Biometrics for Forensic Science. (pp. 195-218). Springer International Publishing.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society of London B, 282*(1814), 20151292.

Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys, 35,* 399-459.

Fig 1. Example photographs and their average. Individual images vary in head angle, expression, lighting, etc. Averaging together multiple images of the same face produces a more stable representation. [Copyright restrictions prevent publication of the face images used in all experiments, though these are available from the authors. Images used in Figs 1, 2, 4 and 6 are illustrative of the experimental stimuli. The individuals pictured in these images did not appear in the experiments, and have given permission for their images to be reproduced here.]

71x20mm (300 x 300 DPI)

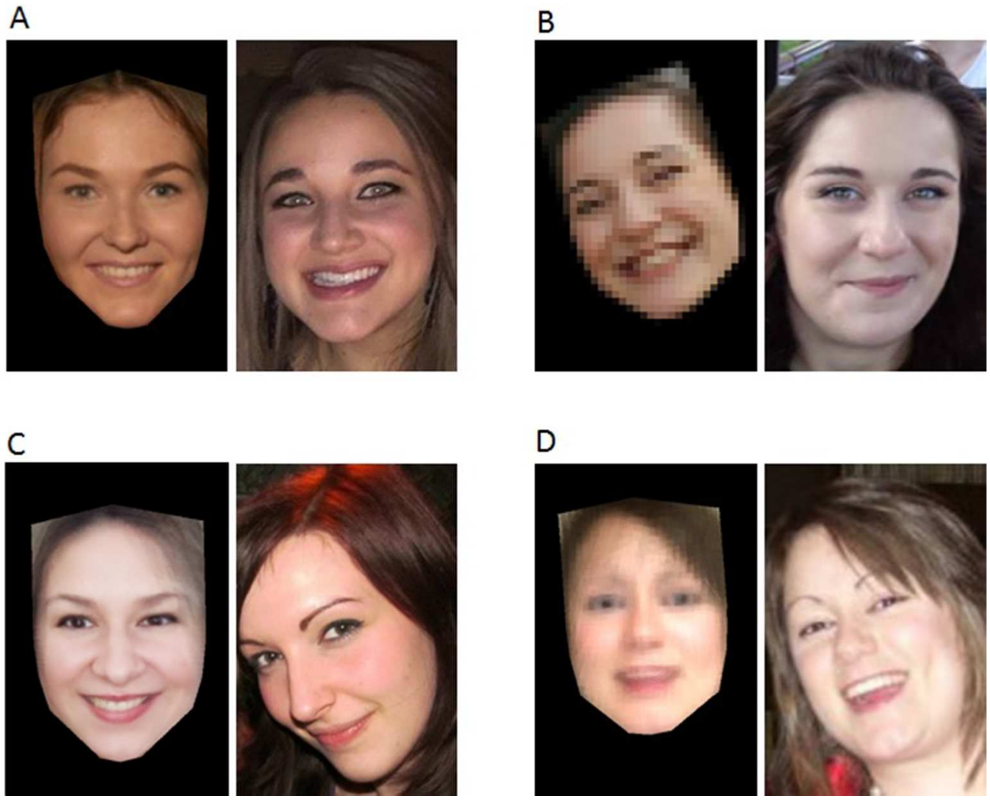

Fig 2. Example stimuli for Experiment 1. A) Unpixelated exemplar mismatch trial; B) Pixelated exemplar match trial; C) Unpixelated average mismatch trial; and D) Pixelated average match trial. The individuals pictured have given permission for their images to be reproduced here.

57x47mm (300 x 300 DPI)

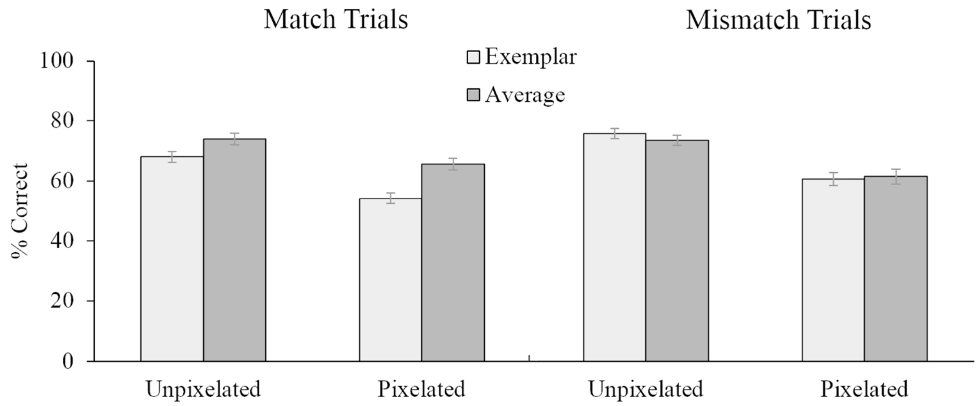

Fig 3. Face matching accuracy. Mean accuracy (percent correct) for the face matching task. Error bars denote standard error of the mean (SEM).

522x216mm (72 x 72 DPI)

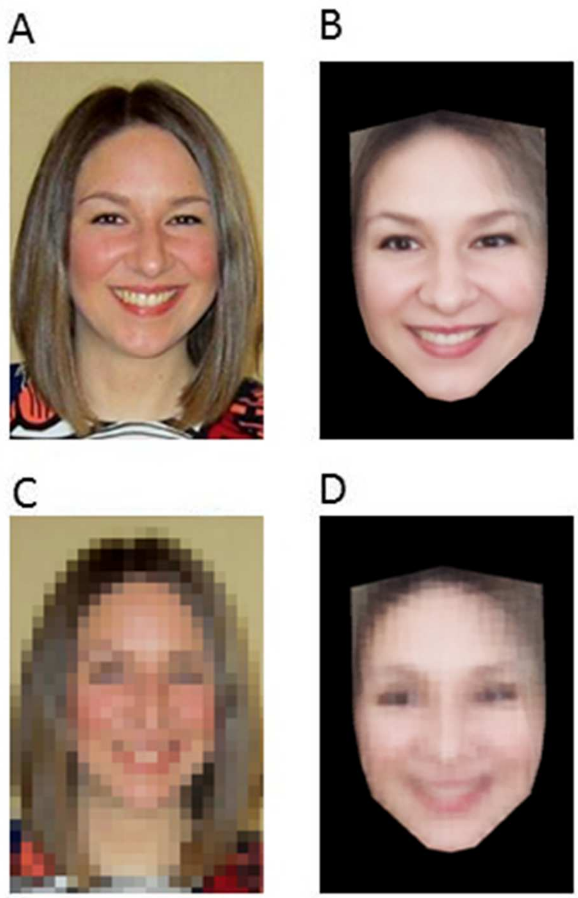

Fig 4. Example stimuli for Experiment 2. A) Unpixelated exemplar; B) Average of ten unpixelated images; C) Pixelated exemplar; and D) Average of ten pixelated images. The individuals pictured have given permission for their images to be reproduced here.

24x38mm (300 x 300 DPI)

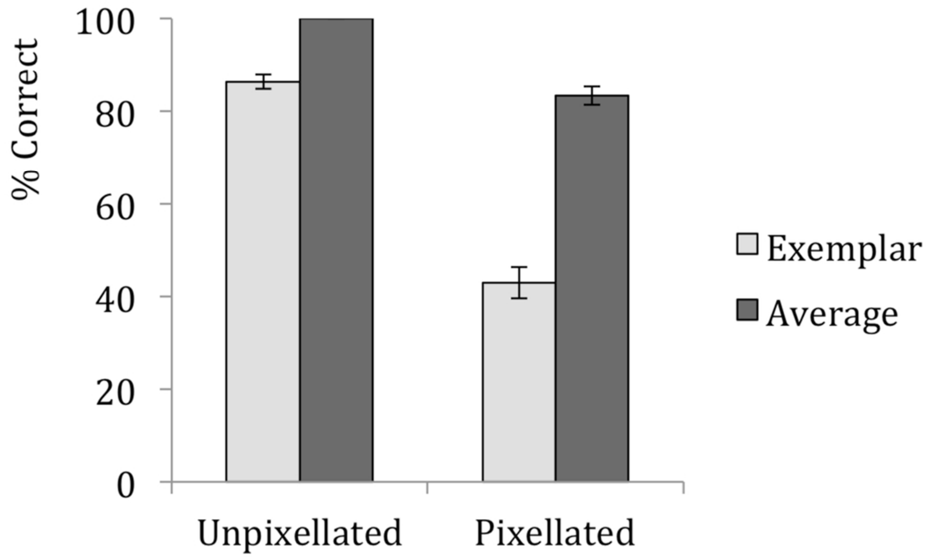

Fig 5. Accuracy of identity returned from Experiment 2 using the FaceDouble application. Error bars denote standard error of the mean (SEM).

93x55mm (300 x 300 DPI)

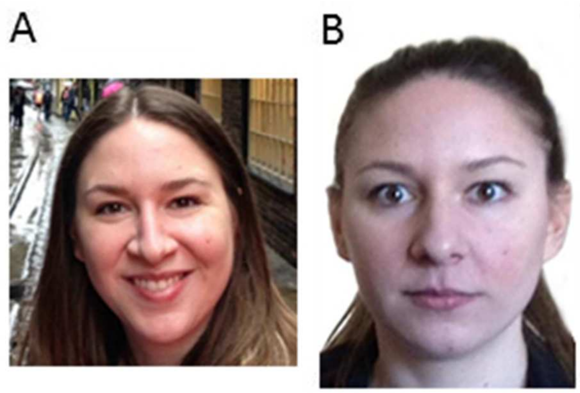

Fig 6. Example stimuli for Experiment 3. A) Image of a target identity cropped to be consistent with the ambient image database images from the 'labelled faces in the wild' set (Huang et al., 2007). B) Image of a target identity chosen to meet passport photo guidelines and edited to remove the background to be consistent with the passport photo database. Images are representative of the stimuli used in Experiment 3 but for reasons of privacy we are not able to provide examples of the passport images used in the database. The individuals pictured have given permission for their images to be reproduced here.

24x16mm (300 x 300 DPI)

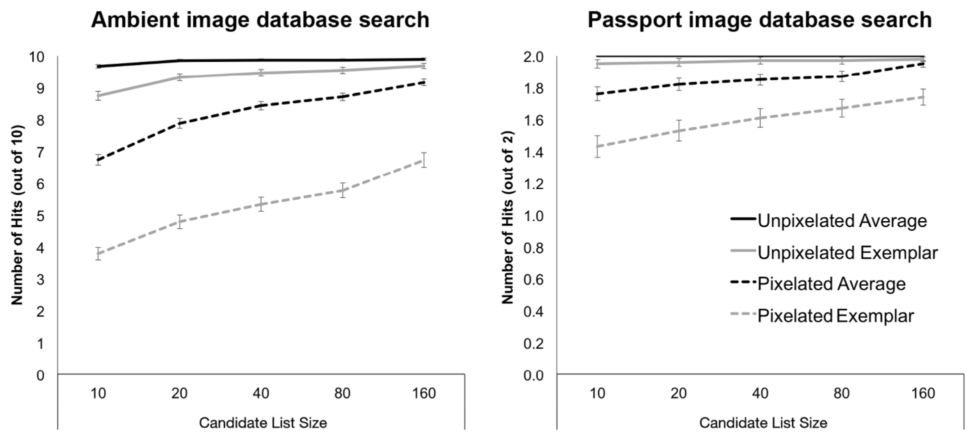

Fig 7. Results of Experiment 3. Identification performance is shown as a function of Gallery size for the Ambient Image test (left) and the Passport Image test (right). Error bars represent standard errors of the mean (SEM).

190x94mm (300 x 300 DPI)