Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Average faces: How does the averaging process change faces physically and perceptually?

Isabelle Bülthoff [a,*], Mintao Zhao [a,b]

[a] Max Planck Institute for Biological Cybernetics, Germany
[b] University of East Anglia, United Kingdom

ARTICLE INFO

Keywords:
Average face
Face recognition
Face norm
Eye movements
Face similarity

ABSTRACT

Average faces have been used frequently in face recognition studies, either as a theoretical concept (e.g., face norm) or as a tool to manipulate facial attributes (e.g., modifying identity strength). Nonetheless, how the face averaging process— the creation of average faces using an increasing number of faces —changes the resulting averaged faces and our ability to differentiate between them remains to be elucidated. Here we addressed these questions by combining 3D-face averaging, eye-movement tracking, and the computation of image-based face similarity. Participants judged whether two average faces showed the same person while we systematically increased their average level (i.e., number of faces being averaged). Our results showed, with increasing averaging, both a nonlinear increase of the computational similarity between the resulting average faces and a nonlinear decrease of face discrimination performance. Participants' performance dropped from near-ceiling level when two different faces had been averaged together to chance level when 80 faces were mixed. We also found a nonlinear relationship between face similarity and face discrimination performance, which was fitted nicely with an exponential function. Furthermore, when the comparison task became more challenging, participants performed more fixations onto the faces. Nonetheless, the distribution of fixations across facial features (eyes, nose, mouth, and the center area of a face) remained unchanged. These results not only set new constraints on the theoretical characterization of the average face and its role in establishing face norms but also offer practical guidance for creating approximated face norms to manipulate face identity.

1. Introduction

Average face, the arithmetic mean of the texture and shape of two or more faces, has been frequently used to formulate theories of face recognition. According to the influential face space framework (Valentine, 1991), the average of all faces one has seen can serve as a face norm, the prototype of faces in one's mind. Such a face norm acts as the origin of multi-dimensional face space, within which we encode, store, compare, and recognize faces that are all represented as locations in that space (e.g., Rhodes, Brennan, & Carey, 1987; Rhodes, Maloney, Turner, & Ewing, 2007; Valentine, 1991). Similarly, according to theories of category-specific face encoding, the average of all faces of a subcategory creates a subordinate-level face norm (e.g., prototype of female faces, happy faces, or Asian faces, Bülthoff & Newell, 2004; Jaquet, Rhodes, & Hayward, 2008; Leopold, O'Toole, Vetter, & Blanz, 2001; Little, DeBruine, & Jones, 2005; Papesh & Goldinger, 2010; Webster, Kaping, Mizokami, & Duhamel, 2004). Moreover, people can even establish identity-level face norms by averaging various faces of the same person, forming stable and robust representations of individual face identities (e.g., Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2011). Average faces may be also used to form a summary representation of a set of faces (i.e., ensemble coding), which offers a mechanism for us to overcome our limited working memory capacity (Whitney & Yamanashi Leib, 2018). Our visual system can extract the mean identity, expression, race, and gender of a group of faces rapidly and accurately (de Fockert & Wolfenstein, 2009; Haberman & Whitney, 2007; Jung, Bülthoff, & Armann, 2017). Neural coding of face identity relative to an average face (i.e., norm-based encoding of face identity) has also been observed in the face-selective brain areas in humans and nonhuman primates (Chang & Tsao, 2017; Leopold, Bondar, & Giese, 2006; Loffler, Yourganov, Wilkinson, & Wilson, 2005).

Although many theories of face recognition share the view that average faces are special, little research has empirically investigated the recognition of average faces per se and validated the hypothetical role of

[*] Corresponding author at: Max Planck Institute for Biological Cybernetics, Max-Planck-Ring 8, D-72076 Tübingen, Germany.
E-mail address: isabelle.buelthoff@tuebingen.mpg.de (I. Bülthoff).

https://doi.org/10.1016/j.cognition.2021.104867
Received 14 September 2020; Received in revised form 27 July 2021; Accepted 28 July 2021
Available online 4 August 2021

face averaging in the development of face norms. Previous studies of average faces have mainly focused on the social perception of faces (e.g., attractiveness or trustworthiness, Rhodes, 2006; Sofer, Dotsch, Wigboldus, & Todorov, 2015) but have rarely investigated how we recognize and discriminate average faces. While average faces have been widely hypothesized as a mathematical approximation of face norms, it remains unclear how the averaging process changes the physical and perceptual properties of average faces, and when, with increasing levels of face averaging, different average faces become perceptually indistinguishable from each other. Similarly, although many studies have used average faces to manipulate the identity, gender or race of faces (e.g., Bülthoff & Zhao, 2020; Jaquet et al., 2008; Jiang, Blanz, & O'Toole, 2006), whether the chosen average faces truly represent the origin of a face space or its sub-space (e.g., depicting neutral identity, gender or race) has seldomly been validated. Therefore, investigating how the face averaging process changes faces physically and perceptually may not only advance our understanding of average faces as a theoretical concept (i.e., face norm) but also help improve our use of average faces as a tool to manipulate face identity (e.g., creating caricatures or antifaces using face morphing, Bülthoff & Zhao, 2020; Jiang et al., 2006; Leopold et al., 2001).

In the present study, we investigated how the face averaging process—creating average faces with more and more faces—affects the resulting faces and changes our ability to discriminate between them. Specifically, we aimed to address four questions. Firstly, we examined whether average faces created with an increasing number of faces become more similar to each other, and, if so, how this similarity evolves with increasing average levels. According to the face space framework (Valentine, 1991), the more faces are mixed together, the closer the resulting average faces are to the hypothetical facial norm. Hence, increasing the average level should make average faces increasingly more similar to each other and more similar to the facial norm. Although it is fundamental to the face space framework, how the similarity between the resulting averaging faces changes with increasing average levels has not yet been systematically tested (cf. Busey, 1998). We addressed this question by creating average faces at 13 average levels (using 2 to 80 faces) and then computed their image-based similarity. By quantifying the relationship between average level and face similarity, we could examine whether, and, if so, how the averaging process gradually pushes the average faces toward the facial norm. Note that although the change of image-based similarity can be computed via a variety of methods, how the face averaging process changes the perceptual similarity between average faces, how image-based face similarity determines perceptual face similarity, and when the perceptual similarity reaches the limit of our face discrimination ability are not readily predicted by computational similarity (e.g., Beale & Keil, 1995; Bülthoff & Zhao, 2020; Jacques & Rossion, 2006; Rotshtein, Henson, Treves, Driver, & Dolan, 2005).

Our second question was related to both the theoretical characterization and the empirical use of average faces as face norms: when do average faces become identity-neutral so that they can serve as operationally defined face norms? Theoretically, face norms are assumed to be established based on all encountered faces (Rhodes et al., 1987; Valentine, 1991; Leopold et al., 2001, 2006). In empirical research, however, creating such facial norms is technically impossible, because we know neither the exact number nor the specific identities of all the faces a person has seen. Consequently, researchers often create average faces using an arbitrary number of faces to act as operationally defined face norms (e.g., 20 to 200 faces, Bülthoff & Zhao, 2020; Loffler et al., 2005; Jeffery, Burton, Pond, Clifford, & Rhodes, 2018). We approached this question by investigating when average faces become perceptually indistinguishable from each other (i.e., approaching a zero-identity strength, Leopold et al., 2001, 2006). We had participants perform a face discrimination task, deciding whether two average faces created using the same number of unique faces (i.e., same average level) depicted the same face identity. This task allowed us to infer at which average level our participants reached their limit to differentiate between these average faces. Here we focused on the estimation of the limit of face discrimination ability for a relatively homogeneous set of faces (i.e., adult Caucasian faces). For one thing, previous studies have suggested that people may form distinct face norms for different categories of faces (e.g., own vs other-race faces; Jaquet et al., 2008; Papesh & Goldinger, 2010). For the other, understanding the effect of face averaging on a homogenous set of faces may set a baseline for investigating more diverse face sets (e.g., a mixture of own- and other-race faces).

Our third question asked whether image-based face similarity between average faces determines our ability to discriminate between them, and, if so, how. Answers to this question will help reveal whether image-based face similarity mediates the influence of the face averaging process on face discrimination performance and, if so, when different average faces are too similar to each other to be distinguished by our visual system. The image-based similarity between two faces has been associated with their perceptual similarity (Dobs et al., 2014; Lades et al., 1993; Yue, Biederman, Mangini, Von Der Malsburg, & Amir, 2012). Nonetheless, assessment of such an association often assumes a linear relationship (e.g., based on correlation analysis), although our perception of face identity is categorical and does not change linearly with the physical change of faces (e.g., Beale & Keil, 1995; Bülthoff & Zhao, 2020; Jacques & Rossion, 2006; Rotshtein et al., 2005). In the present study, we measured image-based face similarity and participants' ability to discriminate between average faces at multiple average levels and quantified how one may change with the other using curve fitting. Moreover, our task also provided an empirical test of how recognition of unfamiliar faces may vary depending on stimulus-based similarity, as our participants were unfamiliar with the average faces and the "parent faces" used to create those average faces. In this respect, our findings may have implications for the theoretical distinction between recognition of familiar and unfamiliar faces (Bruce & Young, 1986; Hancock et al., 2000; Gobbini & Haxby, 2007; Rossion, 2018; Young & Burton, 2018).

Finally, we investigated how the face averaging process affects participants' gaze behavior when discriminating between faces becomes increasingly difficult. Previous studies have shown that how people look at faces during face recognition may differ between cultures (Blais, Jack, Scheepers, Fiset, & Caldara, 2008), between faces of own- vs. other-ethnicities (Brielmann, Bülthoff, & Armann, 2014; Hills & Pake, 2013), and even between individual observers (Mehoudar, Arizpe, Baker, & Yovel, 2014; Peterson & Eckstein, 2013). The landing location of fixations when viewing faces can be used to identify two gaze strategies: holistic fixations and analytic fixations; with the former fixating primarily at the center area of the face, which reflects the processing of a face as a whole, and the latter fixating mainly at the eyes and the mouth areas, which is indicative of part-based processing of faces (Blais et al., 2008; Chuk, Crookes, Hayward, Chan, & Hsiao, 2017). Discriminating between two faces can be relatively easy (e.g., persons of different ethnicities) or more challenging (e.g., identical twins). It remains unclear whether, and, if so, how people adjust their fixations when a face recognition task becomes increasingly challenging. One possibility is that we employ more holistic fixation strategy for an easier task when scrutinizing individual face parts is not necessary, and we use more analytic fixations for a more challenging task when sampling more detailed facial information for face comparison is required. Alternatively, we may look at faces in a similar fashion in terms of holistic versus analytic fixations, and when face discrimination task becomes more demanding, we may simply perform more fixations to either spot the difference or confirm the similarity between two faces. We tested these hypotheses by monitoring participants' eye movements and tested how the number and the landing location of fixations may change with increasing average levels.

2. Methods

2.1. Participants

Twenty-five people participated in the study (17 females 8 males, aged between 19 and 55 years old, mean = 27.5). Data from one participant who did not follow the instructions were discarded. All participants provided written informed consent and received a monetary compensation for their participation. The study was approved by the *Ethics Council* of the *Max Planck* Society.

2.2. Stimuli

We created the average faces using the face database of the Max Planck Institute for Biological Cybernetics and a 3D Morphable Model (Blanz & Vetter, 1999; Troje & Bülthoff, 1996). The shape and the texture of each 3Dlaser-scanned face were defined by the position and color values at 70,000 vertices in a 3D space. To create an average face, we computed the arithmetic means of position and color values at each vertex across a set of "parent faces" and then rendered them as a new 3D face (morph). Therefore, for each average face, all parent faces contributed equally during the face averaging process (e.g., each parent faces weighted 50% when two faces were averaged together, 25% when four faces were averaged, and so forth). In contrast to average faces created by morphing 2D images, our face averaging process also considers depth information and the resulting faces can be rotated in 3D space to display different views.

We used 320 Caucasian faces (half were male and half female) to create average faces at 13 average levels (AL2 to AL10, AL12, AL16, AL32 and AL80). The age of all 160 male faces varied between 19.50 and 36.75 years old (M = 27.02, SD = 4.21). The age of all 160 female faces varied between 18.33 and 32.67 years old (M = 25.44, SD = 3.65). For each average level, we created 20 different average faces (10 males, 10 females) except for two higher average levels (AL32, 10 average faces, 5 males, 5 females; AL 80, 4 average faces, 2 males, 2 females). Each parent face was used only once to create a morph at each average level, which resulted in a smaller number of morphs being created for AL32 and AL80 condition. This strict avoidance of repeated usage of any parent faces ensured that potential differences in the responses observed at different average levels could not be attributed to specific "parent faces" used to create average faces.

For each trial, we showed two average faces side by side on a computer screen; these two faces were always chosen from the same average level, had the same sex and similar mean age across their "parent faces" (Fig. 1). Both faces measured about 8 x 11 cm (7° x 10° visual angle) each and were 6 cm apart. For the *same condition*, the two faces showed the same average face from two different orientations (e.g., trial 2S in Fig. 1). For the *different condition*, we showed two different average faces (i.e., they were created with completely separated sets of faces) in two different orientations (e.g., trial 2D in Fig. 1). To ensure that our task tapped into identity processing and prevented participants from using an image-matching strategy, we showed one face slightly turning to the right (−5°) and the other to the left (+7.5°). The two faces in a trial were slightly turned toward each other for half of the trials and were slightly turned away from each other for the other half of trials (Fig. 1).

As mentioned above, we matched the two average faces used in the different condition in terms of their mean age and the age range of the faces used to create them. To do so, we sorted all 160 faces of each sex according to their age and then paired faces that had a very similar age to create 80 face pairs. The mean age difference between two faces in a pair (i.e., face A and face B) was 1.99 months (SD = 2.83 months; see **Appendix A** for details). For each average face that we created using A faces, there was a paired average face that was created with the corresponding B faces. Consequently, these paired average faces had the same race, gender, and a very similar mean and range of age of "parent faces". Therefore, in the different condition, we maximized the similarity of the average faces in terms of categorical facial information (i.e., they were created using the same number of faces that share the same race, same gender, and similar age). Meanwhile, we also maximized their difference in term of the composition of face identities used to create them (i.e., they never shared a single "parent" identity). This design helped to minimize the potential influence of other face-relevant factors (race, sex, and age) on our task, thereby providing a well-controlled way to assess the limit of face discrimination ability.

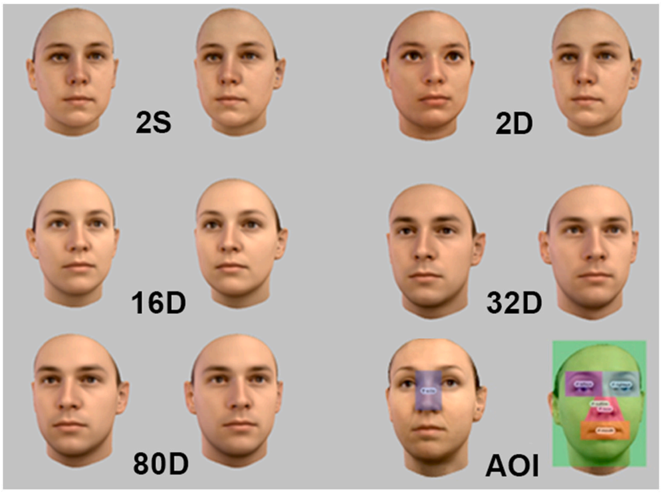

Fig. 1. Example of test stimuli and areas of interests (AOI) for eye-movement tracking. Each trial displays a pair of average faces in different orientations. The numbers (2 to 80) refer to the average level (i.e., how many faces are used to create the average face). The letters S and D represent the *same* and *different* condition respectively. For instance, trial 2S depicts the same average face that had been created with two faces, whereas trial 80D depicts two different average faces, each created with a set of non-overlapping 80 different faces. For AOIs, the eyes, mouth, nose, and whole face are indicated on the right face, whereas the center area that overlaps with the eyes and nose areas is shown on the left face. The sizes of AOIs in pixels are: eyes, 16,500; nose, 6525; mouth, 7975; center, 8250; and whole face, 89,460.

Each participant had 496 trials in total, resulting from 40 trials (20 same and 20 different trials) for each average level from AL2 to AL32 plus 16 trials for AL80 (8 same and 8 different trials). For each average level, each average face was used only once in a different trial and once in a same trial, except for AL32 (each average face used twice) and AL80 (each average face used four times) due to the limited number of average faces. The position (left/right) and orientation (+7.5° or − 5.0°) of the morphs were changed across trials.

2.3. Eye tracking

Eye movements were monitored using the Tobii Pro Spectrum (Tobii Pro AB, Sweden), with a sampling frequency of 600 Hz. Tobii Studio's default fixation filter of 30°/s was used to select fixations and remove saccades. We defined five areas of interests (AOIs) in all faces to pool together fixations located near one of the main facial features: the eyes, nose, mouth, center, and the whole face (Fig. 1, AOI). These AOIs are similar to those used in our previous study (Brielmann et al., 2014). We chose eyes, nose, and mouth regions because they are critical internal facial features of face identity. We included the center region because previous studies have suggested that our gaze tends to land on this area (Peterson & Eckstein, 2012) and that two fixations at this region may suffice to face recognition (Hsiao & Cottrell, 2008). Fixation at the center region has also been associated with holistic processing of faces whereas fixations at the eyes and mouth may reflect analytic processing of faces (Blais et al., 2008; Chuk et al., 2017). We recorded the number of fixations within each AOI during each trial, which allowed us to test whether increasing task difficulty (average level) changes gaze behavior from holistic fixations (i.e., a glimpse to a central face location) to analytic fixations (i.e., a more distributed gaze at eyes, nose, and mouth regions). Note that the center region overlapped with most of the nose area and the area between the eyes. These AOIs were identical in size across faces and trials (except a few exceptions at average levels 2 and 3 for less than 2% of all trials), which assured that fixation calculations across faces are comparable.

2.4. Procedure

After a gaze calibration procedure, participants read an on-screen text instructing them that their task was to decide whether two faces presented simultaneously showed the same person or not. Participants were notified that the difficulty of the task would increase gradually during the test. They then performed a short practice session with trials representing all average levels. Thereafter, all trials of AL2 were shown first, then those of AL3 and so forth up to AL80. We blocked the trials according to average level to maximize the consistency of participants' response and gaze strategy within each average level, which would help maximize the chance of observing any potential difference in gaze behavior due to the change of average level (i.e., task difficulty). We told participants that task difficulty would increase systematically to help participants reach maximal performance. Participants could take self-timed breaks approximatively every 80 trials, followed by a new gaze calibration. Each trial started with a 1-s fixation cross followed by a test image shown for 4 s maximally. The next trial started once participants made a response by pressing corresponding keys or when the 4 s-time ran out.

2.5. Data analysis

2.5.1. Image similarity

We used Gabor dissimilarity between two faces to measure image-based face similarity, as it often correlates with perceptual similarity of faces (Bülthoff & Zhao, 2020; Dobs et al., 2014; Lades et al., 1993; Yue et al., 2012). We followed the method described previously to compute Gabor dissimilarity (Bülthoff & Zhao, 2020; Dobs et al., 2014). Specifically, we converted face images into grayscale (256 by 256 pixels) and then filtered each face image with a Gabor jet [5 scales × 8 orientations × 2 phases (sine and cosine), cantered at the intersections of a 10 × 10 uniform grid] to generate a feature vector. We then computed the Euclidean distance between the two resulting feature vectors (one for each image) as their Gabor dissimilarity. Therefore, higher Gabor dissimilarity represents lower image similarity and a zero value means that the two images are identical.

2.5.2. Linear mixed model (LMM) analysis

As we have 13 ordered and unevenly spaced levels for the within-participants factors such as average level and image-based face similarity, we used a linear mixed model type III tests of fixed effects to statistically test the effect of these factors and their interaction with other factors (e.g., AOIs for eye-tracking data). The LMM analyses were performed using the linear mixed model analysis function in the IBM® SPSS® Statistics (v25). For face discrimination performance (e.g., d prime or accuracy), we used average level (or its associated mean Gabor similarity) as a fixed and repeated factor and the related performance measure as the dependent variable. For the fixation data (e.g., proportion or density of fixations at individual AOIs), we used average level and AOIs as the fixed and repeated factors and the corresponding measures of fixations as the dependent variable. We chose the default diagonal matrix to model repeated covariance (i.e., within-subjects variance-covariance). All reported p values were Bonferroni corrected when multiple comparisons were performed.

2.5.3. Curve fitting

We used the Matlab *fit* function to generate the fitting curves. We fitted the data with four candidate functions [power function: f(x) = a*x^b, f(x) = a*x^b + c; exponential function: f(x) = a*exp(−x*b) + c and its equivalent function f(x) = a*(1-exp(−(x + b)/c)); logarithmic function: f(x) = a*log(x) + b]. We report the best fitting results with the values of the coefficient of determination (R^2) and root mean square error (*RMSE*). For detailed results of all curve fitting reported here, please see **Appendix B**, Table B.1**.**

3. Results

3.1. Image-based similarity between average faces increased with increasing average levels

We first investigated whether average faces created with more faces become more physically similar to each other and, if so, how. For each average level, we computed the image-based similarity between two average faces in all trials of the different condition. We first computed the Gabor similarity using the frontal views. The results showed that face similarity increased (i.e., Gabor dissimilarity decreased) with increasing average levels (Fig. 2A). Differences between average faces diminished rapidly at the lower average levels and this reduction slowed down at higher average levels. This nonlinear relationship is fitted nearly perfectly with a power function ($R^2 = 0.99$, $RMSE = 5.89$). We then computed face similarity by rotating both faces in each trial to the same view (i.e., rotating the face on the right to the same view as the face shown on the left). The same pattern of results was observed ($R^2 = 0.99$, $RMSE = 4.30$, Fig. 2B) and the two measures were highly correlated ($r = 0.99$, $p < .001$). Note that the fitted power functions in Fig. 2A and B are remarkable close to the mathematical description of how average level (i.e., the sample size for the averaging process, X) may affect the difference (i.e., inverse of similarity) between the resulting averages (Y), which decreases following a power function of Y = a* $X^{-0.5}$(see **Appendix C** for more details).

To investigate whether face similarity continues to increase significantly at the two highest average levels, we directly compared AL32 vs AL80. Independent t-tests revealed significantly higher image similarity (i.e., lower Gabor dissimilarity) at AL80 compared to AL32, whether for frontal views, $t(21.43) = 8.28$, $p < .001$, or for the same-rotation view, $t(24.57) = 11.70$, $p < .001$ (equal variances not assumed).

To examine whether average faces created at higher average levels are increasingly closer to the facial norm, we used the four average faces generated with 80 faces as an approximation to the face norm, and calculated image similarity between all average faces and these operationally-defined gender-matched face norms. The good fit of a power function to the results indicate that Gabor dissimilarity values decreased nonlinearly with linearly increasing average levels ($R^2 = 0.99$, $RMSE = 5.06$, Fig. 2C). This result confirms that average faces created at higher average levels are closer to the face norm than those at lower average levels.

3.2. Face discrimination performance decreased with increasing average levels

We measured participants' performance using response sensitivity (d prime), response bias, overall accuracy, and response time (RT). Trials that received no response (2.97% of all 11,904 trials) were excluded from the following analyses.

Participants' response sensitivity decreased with increasing average levels (Fig. 3A). A LMM analysis revealed a significant influence of average level on response sensitivity, $F(12, 46.46) = 44.03$, $p < .001$). Such a drop in response sensitivity was rapid for lower average levels (e.g., AL2 to AL10) and slowed down for higher ones (e.g., AL32 and AL80). This pattern of response is fitted nearly perfectly by an exponential function ($R^2 = 0.99$, $RMSE = 0.11$). Moreover, participants showed significantly above-chance performance (i.e., $d' > 0$) for all but the highest average levels [AL2-AL32, $ts(23) \geq 7.21$, $ps < 0.001$; AL80, t

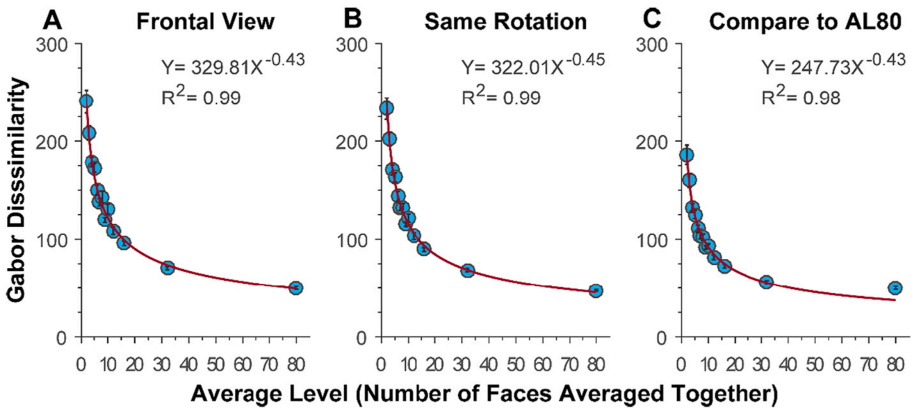

Fig. 2. Mean Gabor dissimilarity between two morphs as a function of average level. Panels A and B show the image similarity computed based on frontal view and the same-rotation view (i.e., both faces turned to the same orientation as the left face in a trial), respectively. Panel C shows the mean image similarity between morphs at each average level and the morphs of same sex at AL80. Lines represent nonlinear least squares fit with a power function. Error bars represent standard errors of the means (SEM).

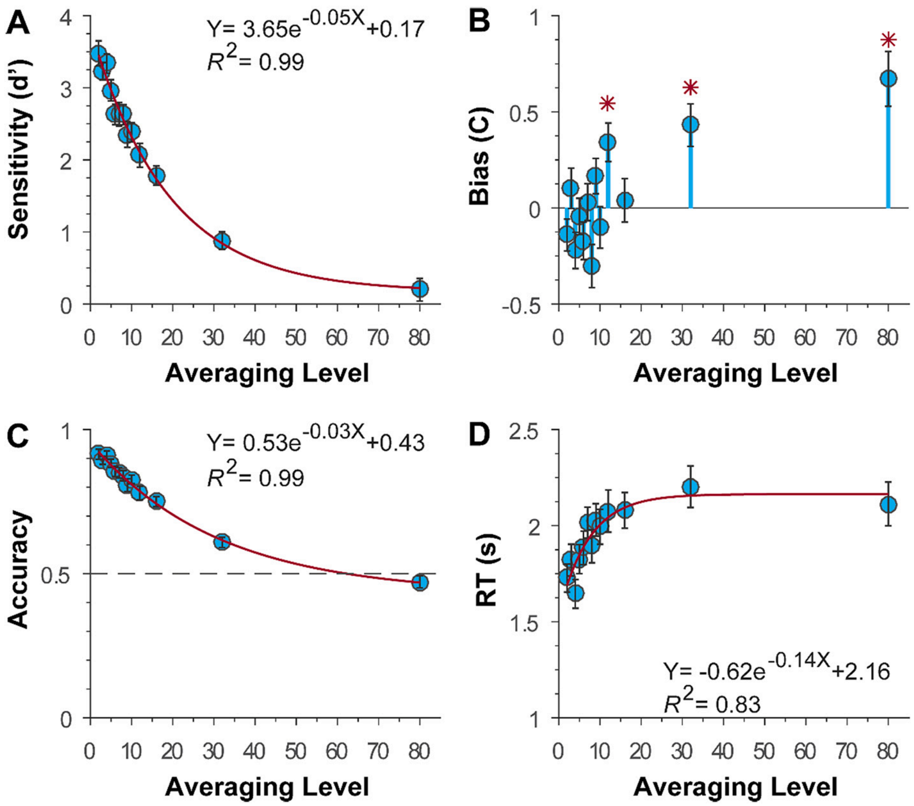

Fig. 3. Participants' performance as a function of average levels. (A) Response sensitivity; (B) Response bias; (C) Overall accuracy; and (D) Response time. Curves represent nonlinear least squares fit with an exponential function. Asterisks indicate significant response bias compared to zero. The dashed line represents chance level performance. Error bars represent SEM.

(23) $= 1.26$, $p = .220$; one-sample t-tests]. Average levels also significantly affected response bias (Fig. 3B), $F(12, 44.76) = 6.01$, $p < .001$. One-sample t-tests revealed that participants showed no significant response bias at lower average levels (AL2-AL10 and AL16), $ts(23) \leq 2.71$, $ps \geq 0.162$, but a significant conservative bias (i.e., tendency to make a 'same' response) at higher average levels (AL12, AL32 and AL80), $ts(23) \geq 3.39$, $ps \leq 0.032$.

Response accuracy mirrored the results of response sensitivity. Accuracy decreased with increasing average levels, dropping rapidly at lower average levels and then more slowly at higher ones (Fig. 3C). This effect of average level on accuracy was significant [$F(12, 40.417) = 46.44$, $p < .001$], and the pattern of response was fitted nearly perfectly with an exponential function ($R^2 = 0.99$, $RMSE = 0.01$). One-sample t-tests revealed above chance-level performance (i.e., accuracy = 0.5) for all average levels except AL80 [AL2-AL32, $ts(23) \geq 6.03$, $ps < 0.001$; AL80, $t(23) = 1.36$, $p = .188$]. This decrease in performance was not due to a speed-accuracy trade-off. With increasing average levels, RT gradually increased up to a ceiling level around 2.1 s (Fig. 3D). This increase was fitted relatively well with an exponential function ($R^2 = 0.83$, $RMSE = 0.07$). Again, the effect of average level was significant, [$F(12, 45.67) = 3.50$, $p = .001$]. We also analyzed the accuracy and RT data separately for the same and different conditions, which showed similar patterns of responses to those measured using overall accuracy and RT (see **Appendix D** for details).

To examine whether participants reached their limit of face discrimination ability before AL80, we compared their performance at the two highest average levels (AL32 vs AL80). Paired t-tests revealed significant differences in terms of response sensitivity, $t(23) = 3.02$, $p = .006$, bias, $t(23) = 2.11$, $p = .046$, accuracy, $t(23) = 4.21$, $p < .001$, but not RT, $t(23) = 1.86$, $p = .076$. These results indicate that, for a homogeneous set of faces like the one we used here, participants were still able to differentiate between average faces created with 32 faces but lost such capability when the average faces were created using 80 faces.

3.3. Face discrimination performance decreased nonlinearly with increasing image similarity

To minimize the 12.5° difference between two face orientations in a trial (one rotated 5° to the right and one 7.5° to the left, Fig. 1), participants might mentally rotate both faces to frontal views, rotate one face to match another face's orientation, or mirror-flipped one face to reduce the difference to 2.5°. We simulated these three potential strategies to calculate image-based face similarity, and then tested how it may determine participants' face discrimination performance. As shown in Fig. 4, the higher the image-based difference between two faces (in term of Gabor dissimilarity), the better the face discrimination performance, regardless of what type of face similarity was calculated and whether the performance was measured as response sensitivity or accuracy. This observation is supported by significant correlations between any combination of performance and face similarity measures (for sensitivity, all *rs* ≥ 0.864, all *ps* < 0.001; for accuracy, all *rs* ≥ 0.780, all *ps* < 0.002). Nonetheless, their relationship is clearly nonlinear. The same magnitude of change in image similarity has a larger effect on performance when two faces are computationally similar (towards the left end of the X-axis) than when they are computationally more different (towards the right end of the X-axis). This nonlinear relationship was captured nearly perfectly with an exponential function (for sensitivity, all $R^2 = 0.99$, all $RMSE \leq 0.12$; for accuracy, all $R^2 = 0.99$, all $RMSE = 0.01$), suggesting that the relationship between face similarity and discrimination performance follows Fechner's law.

Gabor dissimilarity was larger for mirror flipped faces than when computed with frontal view faces or when one of the faces rotated to the same orientation as the other. This result suggests that a small difference between face orientations (e.g., 2.5° for mirror-flipped faces) can dramatically reduce image–based face similarity. Paired t-test confirmed this observation at all 13 average levels (frontal view vs. mirror flipped view, all *ts* > 5.65, all *ps* < 0.001; same rotation vs. mirror flipped view, all *ts* > 7.02, all *ps* < 0.001). This finding may offer an additional account for the well-documented effect of viewpoint change on face recognition (O'Toole, Edelman, & Bülthoff, 1998; Swystun & Logan, 2019).

3.4. Increasing task difficulty induced more fixations but did not change the distribution of fixations

Our task became more challenging at higher average levels, as revealed by the increasing similarity between different average faces and participants' decreasing face discrimination performance. To investigate how task difficulty modulates gaze strategies during face comparison, we first counted the total number of fixations participants performed during each trial and tested whether more difficult trials required more fixations. Fig. 5A shows the mean total number of fixations for the same and the different condition at each average level. A LMM analysis revealed a significant effect of average level, $F(12, 80.30) = 3.858$, $p < .001$, and a significant effect of trial type, $F(1, 546.64) = 27.658$, $p < .001$. The total number of fixations during a face discrimination trial increased with increasing average levels, and participants performed more fixations when the two faces were the same than when they were different (see Fig. 5B for representative examples). The increasing number of fixations echoes the increasing RT reported above (Fig. 3D), as revealed by a significant correlation between RT and the total number of fixations (*rs* ≥ 0.92, *ps* < 0.001, Fig. 5C). The interaction between average level and trial type was not significant, $F(12, 80.30) = 0.904$, $p = .547$, although the difference between the same and different conditions diminished at high average levels (i.e., AL32 and AL80).

Next, we tested whether average level affects the distribution of fixations across facial features. For each participant and each average level, we calculated the percentage of fixations landing at each AOI

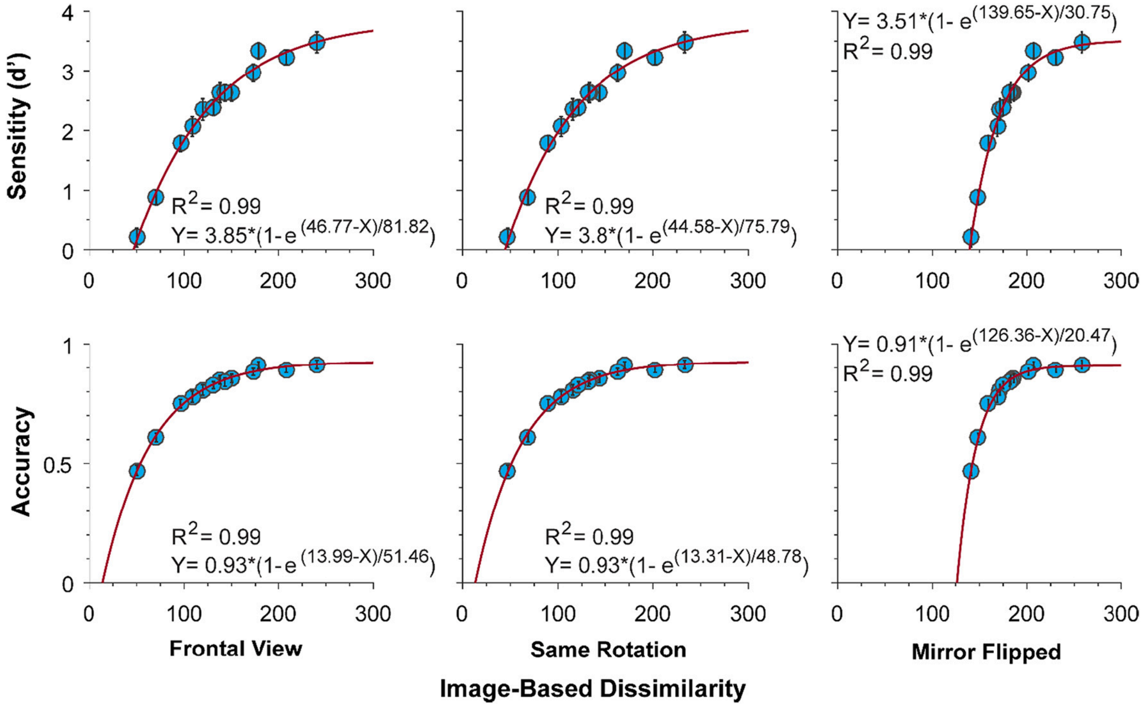

Fig. 4. Face discrimination performance as a function of image-based dissimilarity. Face discrimination performance at each average level was measured using response sensitivity (upper row) and accuracy (lower row). Image-based face dissimilarity was computed when two faces were both in frontal view (left column), when they were rotated to the same view (middle column), or when one face was mirror-flipped to minimize viewpoint difference (right column). Note that larger values along the X-axis represent lower image similarity. Curves represent nonlinear least squares fit with an exponential function (in the format of cumulative density function). Error bars represent SEM.

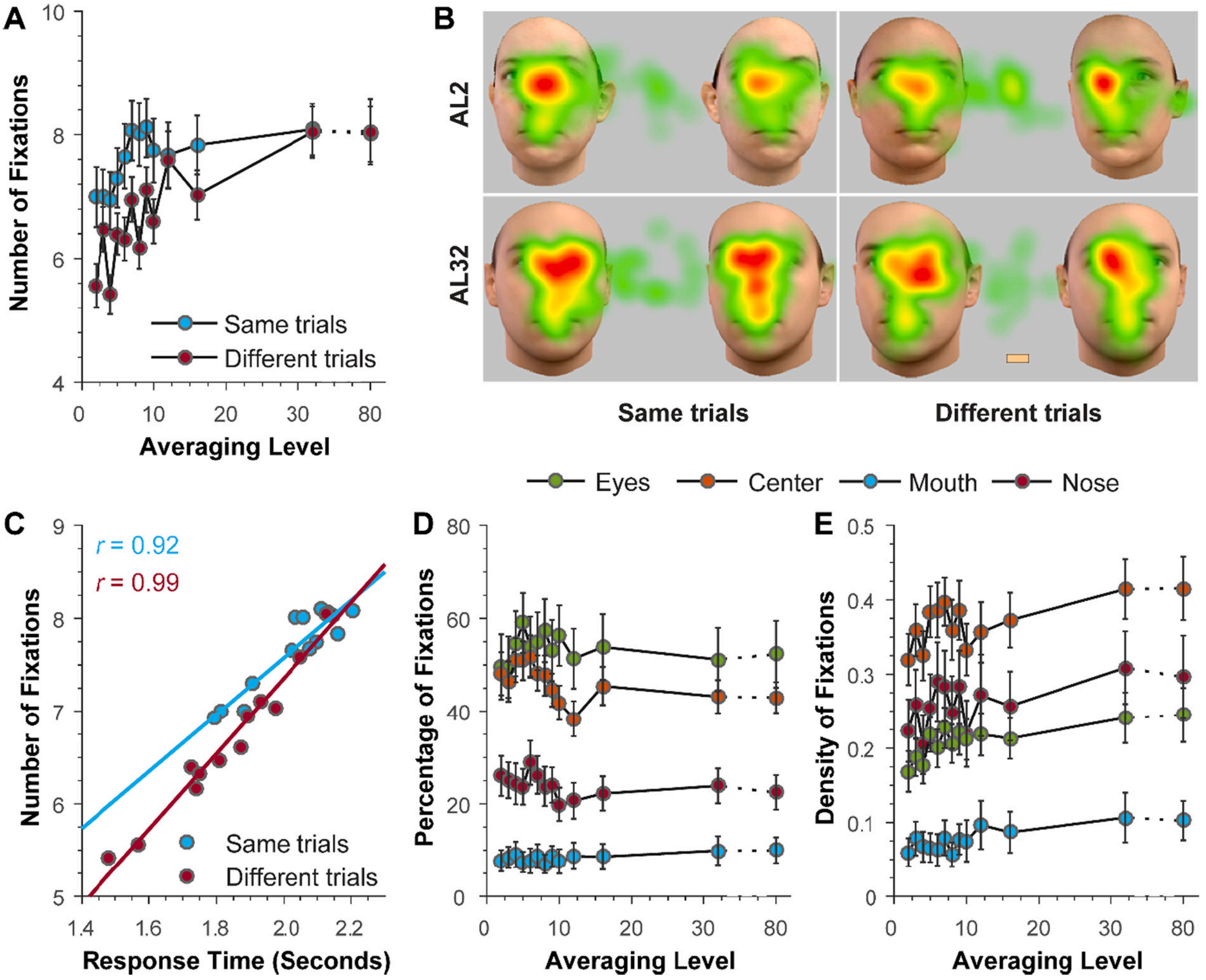

Fig. 5. Fixations recorded during the face discrimination task. (A) The total number of fixations as a function of average level for same and different trials. (B) Representative examples of eye movement distribution for same and different trials at average levels AL2 and AL32. The orange rectangle in the lower right panel illustrates a reference area that covers 1000 pixels. (C) Correlations between the number of fixations and the response time. (D) Distribution of fixations among the AOIs as a function of average level. (E) Density of fixations for each AOI (i.e., number of fixations per 1000 pixels of AOI size) as a function of average level. Error bars represent SEM. Note that average level AL80 is not shown to scale on the abscissa in panels A, D and E.

when the two average faces were different. We focused on the different condition because in the same condition the image-based face similarity did not vary across average levels and did not contribute to the increased task difficulty. For the results about the distribution and density of fixations obtained for the same condition, see **Appendix E**. A LMM analysis with average level and AOI as repeated factors revealed a significant effect of AOI on the proportion of fixations, $F(3,694.70) = 368.847$, $p <$ .001. Neither the effect of average level, $F(12,125.44) = 0.548$, $p = .879$, nor its interaction with AOI, $F(36,105.83) = 0.394$, $p = .999$, was significant. As shown in Fig. 5D, across all average levels, the eyes attracted about one-half of all fixations (53.64 ± 1.86%), and the proportion dropped to about a quarter (23.94 ± 1.11%) for the nose area and less than one-tenth for the mouth region (8.43 ± 0.71%). Pairwise comparisons confirmed that the eyes attracted more fixations than the nose, $t(483.49) = 13.695$, $p < .001$, and the nose received more fixations than the mouth, $t(494.974) = 11.727$, $p < .001$. The center area, which overlapped with part of the eyes and nose regions, received nearly half of all fixations (46.20 ± 1.13%). The proportion is lower than that for the eyes, $t(487.049) = 3.415$, $p = .004$, but higher than that for the nose or the mouth ($ts > 14.037$, $ps < 0.001$). Therefore, while participants made more fixations with increasing task difficulty, the distribution of fixations across AOIs remained unchanged (e.g., eyes > center > nose > mouth).

Finally, we assessed if there was a tendency of fixating the center of faces when considering the size of AOIs. The observed larger proportion of fixation to the eyes area might be due to its larger size compared to other AOIs (Fig. 1). To address this issue, for each AOI we computed the density of fixations as the number of fixations per 1000 pixels (Fig. 5E) and then performed the same LMM analysis as above. We found a significant effect of AOI on the density of fixations, $F(3,548.05) = 203.525$, $p < .001$. The density of fixations was higher for the center area (0.37 ± 0.01) than for the nose (0.26 ± 0.013), $t(542) = 6.412$, $p < .001$, which was higher than that for the eyes (0.21 ± 0.008), $t(461.37) = 3.288$, $p =$ .007, and the eyes showed a higher density than that for the mouth (0.08 ± 0.007), $t(526.55) = 12.798$, $p < .001$. These results indicate that participants' fixations landed more likely at the center area. The effect of average level, $F(12,132.534) = 1.741$, $p = .065$, and its interaction with AOI, $F(36,82.656) = 0.216$, $p = .999$, were not significant. Thus, the density of fixations in each AOI remains unchanged across average levels.

4. Discussion

When average faces are created with an increasing number of faces, they become computationally more similar to each other and more similar to the hypothetical face norm (which is approximated using the average faces of 80 identities in our study). Consistent with the view that face norms can be established through face averaging, our results show that the higher the average level, the more the average faces exhibit the hypothetical properties of a face norm (e.g., they become more typical and are located increasingly closer to the center of face space, Busey, 1998; Valentine, 1991). Moreover, we show that gradually increasing average level leads to a nonlinear increase of image-based face similarity, irrespective of how we calculated such face similarity (e.g., based on frontal view or the same rotated view, Fig. 2). The physical difference between two average faces diminishes rapidly at lower average levels and then the reduction slows down at higher average levels, following nicely a power law. These results provide new insights into how face norms are established and stabilized and set constraints on how to create an operationally defined face norm to manipulate various properties of faces.

If our visual system uses face averaging to establish identity-neutral face norms, our results suggest that we may not use all encountered faces the same way to form face norms. The increase of image-based similarity between average faces slows down with increasing average levels and the corresponding change of face discrimination performance suggest that faces encountered earlier may play a more important role in establishing a stable face norm than those seen later. This view is consistent with the pivotal role of early face exposure in the development of face space and expert face processing system (Crookes & McKone, 2009; de Heering, de Liedekerke, Deboni, & Rossion, 2010; McKone, Crookes, Jeffery, & Dilks, 2012). Faces seen earlier (e.g., the critical period during early infancy) may be used to rapidly construct a face prototype, which will then undergo further fine-tuning with the input of faces encountered later. Consistent with this idea, qualitative difference between the encoding of frequently encountered faces (e.g., own-race faces) and faces rarely seen (e.g., other-race faces) emerges within the first year of life (Kelly et al., 2007). Similarly, children who have been blind since early infancy can learn to differentiate between human faces and face-like objects in about 6 months after their sight recovery (Gandhi, Singh, Swami, Ganesh, & Sinha, 2017). These results suggest that our visual system can create an effective face prototype, based on the exposure to early-encountered faces, to differentiate between familiar and unfamiliar faces and between human faces and face-like objects. These results also raise an interesting question, that is, whether our ability to form a face norm using a limited number of faces is associated with our face recognition ability. Previous studies have shown that norm-based face encoding, measured indirectly using the face identity aftereffect, is correlated with individual differences in face recognition ability (Dennett, McKone, Edwards, & Susilo, 2012; Rhodes, Jeffery, Taylor, Hayward, & Ewing, 2014). Therefore, it is possible that the development of face norms, such as the importance of early and late encountered faces in establishing a stable face norm, may vary between super-recognizers and prosopagnosics. To reveal the potential link between face norm creation and face recognition ability, future study could test how the averaging process affects the physical and perceptual properties of average faces with people varying in their face recognition ability.

Increasing the average level produced both a nonlinear increase of the computational similarity between the resulting average faces and a nonlinear decrease of face discrimination performance. Participants' face discrimination performance dropped from a near-perfect level when two faces were averaged to chance level when 80 faces were averaged (Fig. 3). Increasing the average level also changes participants' response strategy. They tended to make a “same” response at higher average levels (e.g., AL32 and AL80), even though the two different average faces were created using completely separated sets of faces. Moreover, the perception of face similarity (hereby indicated by performance on a face identity matching task) follows a logarithm scale of image-based face similarity, as indicated by a good fitting of exponential functions (Fig. 4). These results indicate that the relationship between physical and perceptual similarity during the face averaging process also follows Fechner's law, as observed in other aspects of face perception (McKone, Aitkin, & Edwards, 2005).

The influence of face averaging on face discrimination performance also sets constraints on the creation of operationally defined face norms. Our participants reached the limit of their face discrimination ability at AL80. In terms of the face space metaphor (e.g., Busey, 1998; Valentine, 1991), this result indicates that average faces created with 80 distinct face identities are located too close in face space to be distinguished from each other. Note that the fitting of our data suggests that such perceptual limit may be reached at a lower average level than the AL80 (Figs. 2, 3A, and C). Future research with more fine-grained average levels may help pinpoint a more precise cut-off point for the limit of our face discrimination ability. Therefore, to safely approximate the face norm of a homogeneous population as we tested here (e.g., same race, sex, and similar age), one may need to create an average face out of 80 or more different face identities. Given that a person may know thousands of people and encounter much more faces in real life (Jenkins, Dowsett, & Burton, 2018), our result suggests that people could establish an identity-neutral face norm using a small fraction of faces they have encountered. Meanwhile, even for our well-controlled homogeneous set of faces, participants' face discrimination performance was still significantly above chance at AL32. This result suggests that the average of 32 face identities is not close to the hypothetic face norm enough to be identity neutral. Therefore, caution should be taken when such average faces are used as face norms to modify the strength of face identity.

Our study also demonstrates that image-based face similarity is tightly related to face discrimination performance. We found a nonlinear relationship between average level and participants' face discrimination performance (Fig. 3), which was mediated by a nonlinear relationship between image-based face similarity and face discrimination performance (Fig. 4). These results were consistently observed across three simulated strategies that participants may use to offset the viewpoint difference between the two average faces (i.e., rotate both faces to frontal view, rotate one face to the same view of another face, or mirror flip one of the two faces). Discrimination between two unfamiliar faces, such as the average faces created in the present study, varies substantially from chance-level to near-perfect performance depending on stimulus-based face similarity. Such stimulus-based variation in unfamiliar face processing should be considered when investigating the distinction between familiar and unfamiliar face recognition (Bruce & Young, 1986; Gobbini & Haxby, 2007; Hancock et al., 2000; Rossion, 2018; Young & Burton, 2018). For instance, caution should be taken when such distinction is based on the different performance of the same group of participants on separate sets of familiar and unfamiliar faces. This is because the difference between the processing of familiar and unfamiliar faces can vary remarkably depending on their intra-stimulus similarity. To rule out the potential influence of face similarity on the qualitative distinction between familiar and unfamiliar face recognition, the same set of faces should be used. The effect of face similarity can then be controlled by testing participants familiar or unfamiliar with those faces (e.g., Bruce, Henderson, Newman, Burton, & M., 2001; Noyes & Jenkins, 2017; Ritchie et al., 2015), by testing participants before and after a face familiarization procedure (e.g., Bonner, Burton, & Bruce, 2003. Clutterbuck & Johnston, 2005), or by a cross over design (as used in the study of the other-race effect in face recognition; e.g., Hills & Pake, 2013; Zhao, Hayward, & Bülthoff, 2014).

The relationship between image-based face similarity and face discrimination performance also offers new insights into the limit of face discrimination ability. Leopold et al. (2001) have manipulated the identity strength of faces by morphing between an original face and the average face (i.e., face norm), and they found that people needed about 11% of identity information for 50% accuracy in recognition. Wilson and colleagues (Gao & Wilson, 2013; Wilson, Loffler, & Wilkinson, 2002) have shown that 6–8% of geometric change between two synthetic faces is required to reach a 75% accuracy in face discrimination. In our study, the mean Gabor dissimilarity between two average faces is about 70 at AL32 and about 50 at AL80. The latter is close to the expected Gabor dissimilarity for a chance level performance according to the fitting of our data (about 45, Fig. 4). For the face stimuli used in our study, the mean Gabor dissimilarity between all same-gender faces was 300. In comparison to this typical difference between two faces, our participants can still differentiate between two very similar faces when the face-based difference drops to a quarter of the typical difference (i.e., 70/300) but they are unable to do so when it further drops to one-sixth of the typical difference (i.e., 50/300). These results provide an additional criterion for choosing an appropriate average level to approximate a face norm.

When discriminating between two average faces became more challenging, participants performed more fixations to scrutinize the faces at a fine-grained level. However, the distribution of fixations across core face features remained unchanged. Whether the two average faces were created using two faces (i.e., visually different) or 80 faces (i.e., very similar), the eyes region always attracted most of the fixations (a half), followed by the nose (a quarter) and the mouth (one-tenth). These results are consistent with previous studies showing that the eyes region is important for face identification and receives the majority of fixations during face comparison (Armann & Bülthoff, 2009; Schyns, Bonnar, & Gosselin, 2002). Therefore, to cope with the increasing task difficulty, our participants adopted an efficient gaze strategy —sampling more detailed facial information (i.e., with additional fixations) without changing the way of information sampling (i.e., stable distribution of fixations). Note that participants performed more fixations for the same trials than for the different trials. This finding suggests that participants may initially use their fixations to spot the diagnostic difference between two faces rather than to confirm their similarity. Once such diagnostic difference is identified, participants stop scanning.

Our participants also showed a tendency to fixate at the center area of the face, particularly when taking the sizes of AOIs into consideration (Fig. 5). Fixating at the center area of faces is argued to be optimal for face recognition (Eckstein & Peterson, 2013; Peterson & Eckstein, 2012; Hsiao & Cottrell, 2008) and is indicative of holistic face processing (Blais et al., 2008; Bombari, Mast, & Lobmaier, 2009; Chuk et al., 2017). That is, faces are perceived as a whole rather than as a combination of independent facial parts (Maurer, Le Grand, & Mondloch, 2002; Rossion, 2013; Zhao, Bülthoff, & Bülthoff, 2016). Our participants used such holistic fixations similarly across all average levels, irrespective of task difficulty. This finding contradicts the idea that participants would use less holistic fixations and more analytic fixations when the face matching task becomes increasingly demanding. This unvarying distribution of fixations across all average levels suggests that we do not scan averaged faces (approaching face norms) and individual faces differently. Therefore, changing task difficulty alone is insufficient to alter the unique and stable eye movement patterns of individual observers (e.g., Mehoudar et al., 2014). In contrast, qualitative change of fixation strategy (e.g., from holistic to analytic) often occurs when the faces being viewed are different from what we usually see (e.g., faces of other races or upside-down orientation, Brielmann et al., 2014; Chuk et al., 2017; Hills & Pake, 2013).

Face averaging, by definition, is affected by both the quantitative aspect (e.g., number of faces and the order of exposure) and the qualitative aspect (e.g., the typicality/distinctiveness of faces) of faces that are being averaged. Our study focuses primarily on the quantitative aspect and counterbalances the influence of the quality aspect (e.g., we used all "parent faces" non-repeatedly at all average levels). How exactly the quality of faces may modulate the effect of face averaging remains unclear. For instance, it has been shown that distinctive faces are located further away from the more typical faces and the face norm (Valentine, 1991; Wallis, 2013), however, how distinctive faces may influence the formation of face norms remains to be elucidated. One possibility is that newly encountered very distinctive faces are discounted in the formation of face norms (i.e., as outliers), as they deviate substantially from both the individual face exemplars and the norm created from those faces. Consistent with this idea, our visual system can form a summary representation of facial expression by averaging most exemplars while discarding extreme outliers (Haberman & Whitney, 2010; see also Alvarez, 2011). Another possibility is that newly encountered distinctive faces contribute more to the representation of face variability (i.e., variance) than the face norm (i.e., the mean). By systematically varying the distinctiveness of faces added to the face averaging process, future research may help differentiate between these possibilities.

Our results about the effects of the face averaging process on perceptual and image-based face similarity are based on a homogeneous set of faces (i.e., faces of the same race, sex, and similar age from a database of 320 Caucasian adult faces), as we aimed to minimize the influence of identity-independent information (e.g., sex, age and race) on face averaging. Whether our findings apply when more diverse parent faces are used (e.g., faces of different races or age) requires further research. Previous studies have shown that people form separable face norms for different face categories (race, sex, or age, Jaquet et al., 2008; Little et al., 2005; Papesh & Goldinger, 2010). Therefore, establishing face norms for a more diversified face set may not involve face averaging across face categories (e.g., own- and other-race faces). Given that other-race faces are located densely close to each other in face space and close to their norm, due to a lack of expertise with these faces (Papesh & Goldinger, 2010; Valentine, 1991; Wallis, 2013), we may reach the limit of our face discrimination ability at a lower average level for other-race faces than for own-race faces. Hence, if people form multiple face norms for faces of different races, how the face averaging process affects faces and face discriminations for each race might vary.

In summary, although average faces have been frequently used both as a theoretical concept and a practical tool in face recognition research, some fundamental aspects of average faces remain to be empirically validated. By combining 3D-face averaging, eye movement tracking, and image-based similarity analysis, here we show that an increasing average level leads to both a nonlinear increase of image-based face similarity and a nonlinear decrease of face discrimination performance. The influence of face averaging on face discrimination performance (i.e., perceptual similarity) is mediated by image-based face similarity, and the relationship between the perceptual and physical face similarity follows Fechner's law. Our ability to differentiate between average faces drops rapidly at lower average levels and then the drop rate slows down at higher average levels. For the faces of a homogeneous population (e.g., the same race, gender and similar age), our visual system seems to reach its limit to discern between two average faces when they are created with 80 different identities. Moreover, when differentiation between two average faces becomes increasingly demanding, we adapt our fixation strategies accordingly by increasing the number of fixations without changing its distribution across facial features. These results not only help improve the use of average faces as an approximate of face norms but also help us understand how faces may be represented and recognized in a modernized face space framework (e.g., Chang & Tsao, 2017; O'Toole, Castillo, Parde, Hill, & Chellappa, 2018).

Credit authors statement

Isabelle Bülthoff: Idea and Conceptualization

Isabelle Bülthoff and **Mintao Zhao** worked equally on all other aspects of the study and its writing

Acknowledgements

This research was supported by the Max Planck Society. We would like to thank the reviewers for their very constructive comments and thank Karin Bierig for synthesizing the stimuli and collecting the data.

Appendix A

Age matching for creating average faces for the *different* condition

To ensure that the mean parent ages of both average faces in a trial of the Different condition were similar to each other and similar to the mean parent ages of the other average faces in trials of the corresponding Same condition, we paired all “parent faces” according to their age. To create any average face at any average level, we used the paired faces to create each yoked average face. Fig. A.1 shows that 159 of 160 pairs had an age difference smaller than 1 year.

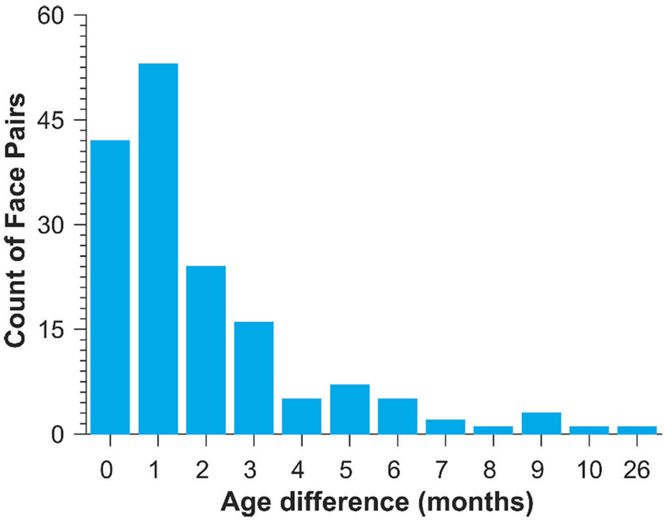

Fig. A.1. Age difference between the 160 pairs of “parent faces” used to create the average faces.

Appendix B

Table B.1
Results of all the curve fitting reported in Figs. 2, 3, and 4 in the main text.

Image similarity as a function of AL						
	Frontal view		Same rotation		To AL80	
Function	R2	RMSE	R2	RMSE	R2	RMSE
f(x) = a*xˆb	**0.99**	**5.89**	**0.99**	**4.30**	**0.98**	**5.06**
f(x) = a*xˆb + c	0.99	6.11	0.99	4.51	0.99	3.47
f(x) = a*exp(−x*b) + c	0.97	9.68	0.98	8.78	0.98	6.11
f(x) = a*log(x) + b	0.93	14.78	0.92	14.94	0.86	14.43
Performance as a function of AL						
	Sensitivity (d’)		Accuracy		RT	
Function	R2	RMSE	R2	RMSE	R2	RMSE
f(x) = a*xˆb	0.87	0.36	0.87	0.05	0.68	0.10
f(x) = a*xˆb + c	0.98	0.15	0.98	0.02	0.75	0.09
f(x) = a*exp(−x*b) + c	**0.99**	**0.11**	**0.99**	**0.01**	**0.83**	**0.07**
f(x) = a*log(x) + b	0.97	0.16	0.92	0.04	0.70	0.09
Performance (d’) as a function of image similarity						
	Frontal View		Same Rotation		Mirror Flipped	
Function	R2	RMSE	R2	RMSE	R2	RMSE
f(x) = a*xˆb	0.88	0.35	0.87	0.36	0.69	0.56
f(x) = a*xˆb + c	0.96	0.20	0.96	0.22	0.98	0.13
f(x) = a*exp(−x*b) + c	0.99	0.11	0.99	0.12	0.99	0.10
f(x) = a*log(x) + b	0.98	0.15	0.97	0.17	0.82	0.42
f(x) = a*(1-exp(−(x + b)/c))	**0.99**	**0.11**	**0.99**	**0.12**	**0.99**	**0.10**
Performance (accuracy) as a function of image similarity						
	Frontal View		Same Rotation		Mirror Flipped	
Function	R2	RMSE	R2	RMSE	R2	RMSE
f(x) = a*xˆb	0.87	0.05	0.86	0.05	0.63	0.08
f(x) = a*xˆb + c	0.90	0.04	0.91	0.04	0.92	0.04
f(x) = a*exp(−x*b) + c	0.99	0.01	0.99	0.01	0.99	0.01
f(x) = a*log(x) + b	0.93	0.04	0.92	0.04	0.69	0.07
f(x) = a*(1-exp(−(x + b)/c))	**0.99**	**0.01**	**0.99**	**0.01**	**0.99**	**0.01**

Note. Results of the goodness of fit for the functions plotted in the figures are shown in bold. The two exponential functions are mathematically equivalent.

Appendix C

Mathematical expectation of the difference between the averages of increasing number of samples from a Gaussian distribution.

Assume we have a Gaussian distribution, N(μ, σ^2), then the averages (means) of X samples from this distribution, A, also follows a Gaussian distribution, N(μ, σ^2/X). The expected value of the square difference between two such averages, Ai and Aj, will be $E[(Ai - Aj)^2] = 2\sigma^2/X$. So, its root mean square distance, $\sqrt{2}\,\sigma * X^{-0.5}$, goes down with the increasing sample size with a rate of $X^{-0.5}$, which is close to our fitting of power function as shown in Fig. 2A, B.

Appendix D

Accuracy and RT data for the same and the different trials, respectively.

Fig. D.1 below shows that both the same and the different condition showed a nonlinear decrease of response accuracy and a nonlinear increase of RT. Response accuracy on the different condition dropped faster with increasing average levels compared to that for the same condition. The patterns of responses are similar to the overall accuracy and RT data reported in Fig. 3 in the main text.

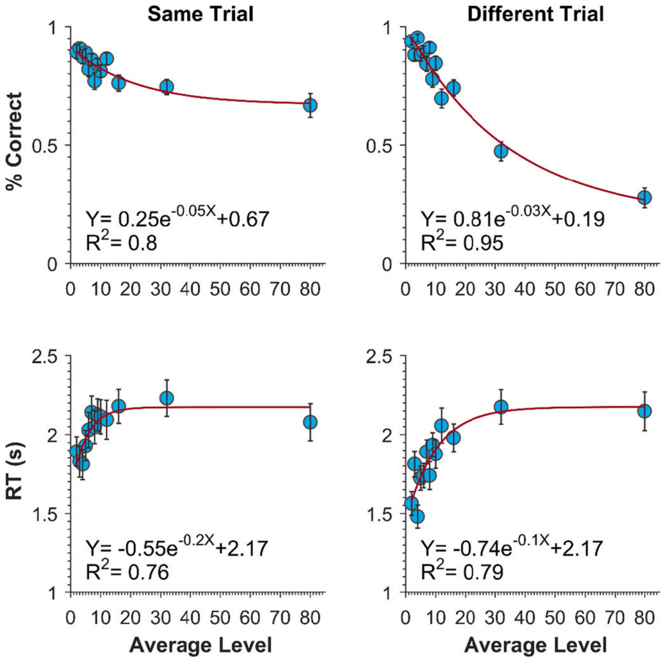

Fig. D.1. Participants' response accuracy and RT as a function of average level and trial condition (same vs different). Curves represent nonlinear least squares fit with an exponential function. Error bars represent SEM.

Appendix E

Distribution and density of fixations observed for the same trials.

The proportion and density of fixations recorded on the same condition (Fig. E.1) are similar to those observed for the different condition (Fig. 5D and E in main text). For the proportion of fixations, A LMM analysis with average level and AOI as repeated factors revealed a significant effect of AOI, $F(3,676.64) = 347.091$, $p < .001$. Neither the effect of average level, $F(12,105.16) = 0.375$, $p = .970$, nor its interaction with AOI, $F(36,96.69) = 0.275$, $p = .999$, was significant. Across all average levels, about one-half of all fixations (51.99 ± 1.77%) landed at the eyes region, a quarter (24.47 ± 1.01%) at the nose area and one-tenth (9.61 ± 0.71%) at the mouth region. For the density of fixations, the same analysis revealed a significant effect of AOI, $F(3,567.21) = 196.24$, $p < .001$. The effect of average level, $F(12,136.67) = 0.769$, $p = .681$, and its interaction with AOI, $F(36,90.43) = 0.225$, $p = .999$, were not significant. Pairwise comparison revealed a higher density of fixations at the center (0.41 ± 0.011) than at all other AOIs (eyes: 0.23 ± 0.008; nose: 0.30 ± 0.014; mouth: 0.10 ± 0.008), all *ps* < 0.001. Therefore, increasing average level (and consequently task difficulty) does not alter the gaze strategies used to differentiate between two average faces.

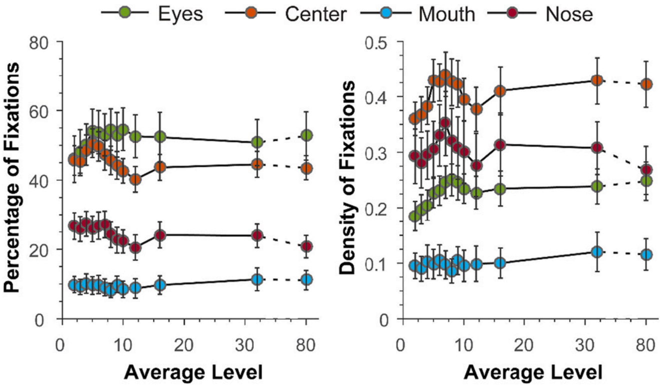

Fig. E.1. The distribution and density of fixations across AOIs as a function of average level. Note that average level AL80 is not shown to scale on the abscissa. Error bars represent SEM.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104867.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003.

Armann, R., & Bülthoff, I. (2009). Gaze behavior in face comparison: The roles of sex, task, and symmetry. *Attention, Perception, & Psychophysics, 71*(5), 1107–1126. https://doi.org/10.3758/App.71.5.1107.

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition, 57*, 217–239.

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture Shapes How We Look at Faces. *PLoS One, 3*(8), Article e3022. https://doi.org/10.1371/journal.pone.0003022.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99* (pp. 187–194). New York: ACM Press. https://doi.org/10.1145/311535.311556.

Bonner, L., Burton, A. M., & Bruce, V. (2003). Getting to know you: How we learn new faces. *Visual Cognition, 10*(3), 527–536.

Brielmann, A. A., Bülthoff, I., & Armann, R. (2014). Looking at faces from different angles: Europeans fixate different features in Asian and Caucasian faces. *Vision Research, 100*, 105–112.

Bruce, V., Henderson, Z., Newman, C., Burton, A., & M.. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*(3), 207–218.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327.

Bülthoff, I., & Newell, F. N. (2004). Categorical perception of sex occurs in familiar but not unfamiliar faces. *Visual Cognition, 11*(7), 823–855. https://doi.org/10.1080/13506280444000012.

Bülthoff, I., & Zhao, M. (2020). Personally familiar faces: Higher precision of memory for idiosyncratic than for categorical information. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 46*(7), 1309–1327.

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.

Busey, T. A. (1998). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science, 9*(6), 476–483. https://doi.org/10.1111/1467-9280.00088.

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell, 169*(6), 1013–1028.

Chuk, T., Crookes, K., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2017). Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures. *Cognition, 169*, 102–117. https://doi.org/10.1016/j.cognition.2017.08.003.

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology, 14*(1), 97–116.

Crookes, K., & McKone, E. (2009). Early maturity of face recognition: No childhood development of holistic processing, novel face encoding, or face-space. *Cognition, 111*(2), 219–247. https://doi.org/10.1016/j.cognition.2009.02.004.

Dennett, H. W., McKone, E., Edwards, M., & Susilo, T. (2012). Face aftereffects predict individual differences in face recognition ability. *Psychological Science, 23*(11), 1279–1287.

Dobs, K., Bülthoff, I., Breidt, M., Vuong, Q. C., Curio, C., & Schultz, J. (2014). Quantifying human sensitivity to spatio-temporal information in dynamic faces. *Vision Research, 100*, 78–87. https://doi.org/10.1016/j.visres.2014.04.009.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology, 62*(9), 1716–1722. https://doi.org/10.1080/17470210902811249.

Gandhi, T. K., Singh, A. K., Swami, P., Ganesh, S., & Sinha, P. (2017). Emergence of categorical face perception after extended early-onset blindness. *Proceedings of the National Academy of Sciences of the United States of America, 114*(23), 6139–6143.

Gao, X., & Wilson, H. R. (2013). The neural representation of face space dimensions. *Neuropsychologia, 51*(10), 1787–1793. https://doi.org/10.1016/j.neuropsychologia.2013.07.001.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia, 45*(1), 32–41. https://doi.org/10.1016/j.neuropsychologia.2006.04.015.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751–R753. https://doi.org/10.1016/j.cub.2007.06.039.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics, 72*(7), 1825–1838. https://doi.org/10.3758/App.72.7.1825.

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

de Heering, A., de Liedekerke, C., Deboni, M., & Rossion, B. (2010). The role of experience during childhood in shaping the other-race effect. *Developmental Science, 13*(1), 181–187. https://doi.org/10.1111/j.1467-7687.2009.00876.x.

Hills, P. J., & Pake, J. M. (2013). Eye-tracking the own-race bias in face recognition: revealing the perceptual and socio-cognitive mechanisms. *Cognition, 129*(3), 586–597. https://doi.org/10.1016/j.cognition.2013.08.012.

Hsiao, J. H. W., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological Science, 19*(10), 998–1006.

Jacques, C., & Rossion, B. (2006). The speed of individual face categorization. *Psychological Science, 17*(6), 485–492.

Jaquet, E., Rhodes, G., & Hayward, W. G. (2008). Race-contingent aftereffects suggest distinct perceptual norms for different race faces. *Visual Cognition, 16*(6), 734–753. https://doi.org/10.1080/13506280701350647.

Jeffery, L., Burton, N., Pond, S., Clifford, C., & Rhodes, G. (2018). Beyond opponent coding of facial identity: Evidence for an additional channel tuned to the average face. *Journal of Experimental Psychology: Human Perception and Performance, 44*(2), 243–260. https://doi.org/10.1037/xhp0000427.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 366*(1571), 1671–1683. https://doi.org/10.1098/rstb.2010.0379.

Jenkins, R., Dowsett, A. J., & Burton, A. M. (2018). How many faces do people know? *Proceedings of the Royal Society B: Biological Sciences, 285*, 20181319. https://doi.org/10.1098/rspb.2018.1319.

Jiang, F., Blanz, V., & O'Toole, A. J. (2006). Probing the visual representation of faces with adaptation - A view from the other side of the mean. *Psychological Science, 17*(6), 493–500.

Jung, W., Bülthoff, I., & Armann, R. (2017). The contribution of foveal and peripheral visual information to ensemble representation of face race. *Journal of Vision, 17*(13), 11. https://doi.org/10.1167/17.13.11.

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L. Z., & Pascalis, O. (2007). The other-race effect develops during infancy - Evidence of perceptual narrowing. *Psychological Science, 18*(12), 1084–1089.

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., Vandermalsburg, C., Wurtz, R. P., & Konen, W. (1993). Distortion Invariant Object Recognition in the Dynamic Link Architecture. *IEEE Transactions on Computers, 42*, 300–311.

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature, 442*(7102), 572–575.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience, 4*, 89–94.

Little, A. C., DeBruine, L. M., & Jones, B. C. (2005). Sex-contingent face after-effects suggest distinct neural populations code male and female faces. *Proceedings of the Royal Society B: Biological Sciences, 272*(1578), 2283–2287.

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neuroscience, 8*(10), 1386–1390.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6), 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4.

McKone, E., Aitkin, A., & Edwards, M. (2005). Categorical and coordinate relations in faces, or Fechner's law and face space instead? *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1181–1198. https://doi.org/10.1037/0096-1523.31.6.1181.

McKone, E., Crookes, K., Jeffery, L., & Dilks, D. D. (2012). A critical review of the development of face recognition: Experience is less important than previously believed. *Cognitive Neuropsychology, 29*(1–2), 174–212. https://doi.org/10.1080/02643294.2012.660138.

Mehoudar, E., Arizpe, J., Baker, C. I., & Yovel, G. (2014). Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. *Journal of Vision, 14*(7:6), 1–11. https://doi.org/10.1167/14.7.6.

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition, 165*, 97–104.

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face Space Representations in Deep Convolutional Neural Networks. *Trends in Cognitive Sciences, 22*(9), 794–809.

O'Toole, A. J., Edelman, S. Y., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research, 38*(15–16), 2351–2363. https://doi.org/10.1016/S0042-6989(98)00042-X.

Papesh, M. H., & Goldinger, S. D. (2010). A multidimensional scaling analysis of own- and cross-race face spaces. *Cognition, 116*(2), 283–288.

Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences of USA, 109*(48), E3314–E3323.

Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological Science, 24*(7), 1216–1225. https://doi.org/10.1177/0956797612471684.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology, 57*, 199–226.

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and Ratings of Caricatures - Implications for Mental Representations of Faces. *Cognitive Psychology, 19*(4), 473–497.

Rhodes, G., Jeffery, L., Taylor, L., Hayward, W. G., & Ewing, L. (2014). Individual differences in adaptive coding of face identity are linked to individual differences in face recognition ability. *Journal of Experimental Psychology. Human Perception and Performance, 40*(3), 897–903.

Rhodes, G., Maloney, L. T., Turner, J., & Ewing, L. (2007). Adaptive face coding and discrimination around the average face. *Vision Research, 47*(7), 974–989.

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*, 162–169.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition, 21*(2), 139–253. https://doi.org/10.1080/13506285.2013.772929.

Rossion, B. (2018). Humans Are Visual Experts at Unfamiliar Face Recognition. *Trends in Cognitive Sciences, 22*(6), 471–472. https://doi.org/10.1016/j.tics.2018.03.002.

Rotshtein, P., Henson, R. N. A., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience, 8*(1), 107–113.

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science, 13*(5), 402–409. https://doi.org/10.1111/1467-9280.00472.

Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What Is Typical Is Good: The Influence of Face Typicality on Perceived Trustworthiness. *Psychological Science, 26*(1), 39–47.

Swystun, A. G., & Logan, A. J. (2019). Quantifying the effect of viewpoint changes on sensitivity to face identity. *Vision Research, 165*, 1–12.

Troje, N. F., & Bülthoff, H. H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research, 36*(12), 1761–1771. https://doi.org/10.1016/0042-6989(95)00230-8.

Valentine, T. (1991). A Unified Account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition. *The Quarterly J. Exp. Psychol. Section A, 43*, 161–204. https://doi.org/10.1080/14640749108400966.

Wallis, G. (2013). Toward a unified model of face and object recognition in the human visual system. *Frontiers in Psychology, 4*, 497. https://doi.org/10.3389/Fpsyg.2013.00497.

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*, 557–561.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology, 69*, 105–129. https://doi.org/10.1146/annurev-psych-010416-044232.

Wilson, H. R., Loffler, G., & Wilkinson, F. (2002). Synthetic faces, face cubes, and the geometry of face space. *Vision Research, 42*, 2909–2923.

Young, A. W., & Burton, A. M. (2018). Are We Face Experts? *Trends in Cognitive Sciences, 22*(2), 100–110.

Yue, X., Biederman, I., Mangini, M. C., Von Der Malsburg, C., & Amir, O. (2012). Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vision Research, 55*, 41–46. https://doi.org/10.1016/j.visres.2011.12.012.

Zhao, M., Bülthoff, H. H., & Bülthoff, I. (2016). Beyond faces and expertise: Facelike holistic processing of nonface objects in the absence of expertise. *Psychological Science, 27*(2), 213–222.

Zhao, M., Hayward, W. G., & Bülthoff, I. (2014). Holistic processing, contact, and the other-race effect in face recognition. *Vision Research, 105*, 61–69.