Facial metrics generated from manually and automatically placed image landmarks are highly correlated

Alex L. Jones[1], Christoph Schild[2], & Benedict C. Jones[3]

[1]Swansea University, United Kingdom

[2]University of Copenhagen, Denmark

[3]University of Strathclyde, United Kingdom

Corresponding author: Alex Jones, Swansea University, United Kingdom (alexjonesphd@gmail.com)

Manuscript accepted at *Evolution and Human Behaviour* (September 2020)

Facial metrics generated from manually and automatically placed image landmarks are highly correlated

Abstract

Research on social judgments of faces often investigates relationships between measures of face shape taken from images (facial metrics), and either perceptual ratings of the faces on various traits (e.g., attractiveness) or characteristics of the photographed individual (e.g., their health). A barrier to carrying out this research using large numbers of face images is the time it takes to manually position the landmarks from which these facial metrics are derived. Although research in face recognition has led to the development of algorithms that can automatically position landmarks on face images, the utility of such methods for deriving facial metrics commonly used in research on social judgments of faces has not yet been established. Thus, across two studies, we investigated the correlations between four facial metrics commonly used in social perception research (sexual dimorphism, distinctiveness, bilateral asymmetry, and facial width to height ratio) when measured from manually and automatically placed landmarks. In the first study, in two independent sets of open access face images, we found that facial metrics derived from manually and automatically placed landmarks were typically highly correlated, in both raw and Procrustes-fitted representations. In study two, we investigated the potential for automatic landmark placement to differ between White and East Asian faces. We found that two metrics, facial width to height ratio and sexual dimorphism, were better approximated by automatic landmarks in East Asian faces. However, this difference was small, and easily corrected with outlier detection. These data validate the use of automatically placed landmarks for calculating facial metrics to use in

research on social judgments of faces, but we urge caution in their use. We also provide a tutorial for the automatic placement of landmarks on face images.

Keywords

face processing; computer graphics; sexual dimorphism; person perception

The human face is an important social stimulus. From a multitude of signals within faces, we can infer information about an individual that is often critical for social interaction, such as their age (Imai & Okami, 2019) and sex (Burton et al., 1993). People also make inferences regarding social social traits, such as attractiveness (Rhodes, 2006), health (Jones, 2018), and trustworthiness (Sutherland et al., 2013), from facial characteristics. Although the veracity of these perceptions is often questionable, they can influence important social outcomes, such as hiring and voting decisions and romantic partner choice (Todorov et al., 2015).

Researchers investigating social judgments of faces will often take specific shape measurements from face images and examine associations between these measurements and either perceived or physical characteristics of the photographed individual. For example, many studies have used this facial metric approach to investigate putative relationships between sexual dimorphism, distinctiveness, bilateral asymmetry, or facial width to height ratio (fWHR) and ratings of traits such as attractiveness, health, or dominance of face images (Holzeitner et al., 2019; Jones, 2018; Komori et al., 2009, 2011; Said & Todorov, 2011; Scheib et al., 1999). Other studies have used this approach to investigate putative relationships between these metrics and qualities of the photographed individuals such as their physical health, hormonal profile, or body size (Cai et al., 2018; Geniole et al., 2014; Lefevre et al., 2013; Wolffhechel et al., 2015). This approach has been invaluable for providing insights into the nature of the relationships among facial shape, person perception, and physical condition and, in doing so, has helped identify factors that drive social judgments of faces.

A significant barrier to addressing these research questions, and more importantly, addressing them well, is the length of time it takes to manually place the

landmarks that are essential for calculating these facial metrics. Indeed, this cost may explain why studies investigating relationships among measured face shape and perceived or physical characteristics of the photographed individual are often underpowered (Cai et al., 2018; Holzeitner et al., 2019). Manual placement of landmarks on face images is also arguably a barrier to the reproducibility of facial metrics, since some research demonstrates that different people place key landmarks in different locations on face images (Geniole et al., 2014; Grammer & Thornhill, 1994; Rikowski & Grammer, 1999; Scheib et al., 1999). With many open face image sets now available (for a comprehensive list of open access face image sets, see https://rystoli.github.io/FSTC.html), these issues represent a significant block on research progress. In addition, these landmarks are also often used to create facial averages that individual images can be warped between to test the effect on perceptions (Scott et al., 2013; Sutherland et al., 2017), highlighting the essential nature and involvement of manual landmarking in many avenues of face perception research.

An alternative approach to manual placement of landmarks is to use fully automated landmark placement. Computer vision research has developed powerful face recognition algorithms trained to place landmarks quickly, automatically, and reproducibly, using regression tree methods (King, 2009). While they have seen extensive use in computer vision work (Baddar et al., 2016; Damer et al., 2019; Özseven & Düğenci, 2017; Schroff et al., 2015), these methods have not yet been validated for use in social perception research. Given that these automatically placed landmarks capture shape information vital for facial recognition (Juhong & Pintavirooj, 2017; Shi et al., 2006), they may capture equally well the metrics of interest to social perception. If validated for measurement of facial metrics, automatic

landmark placement would substantially decrease the time cost that manual landmark placements require, produce fully reproducible facial metrics, and ultimately improve the quality of research using facial metrics to investigate social perception.

In light of the above, in our first study, we investigated the correlations between four facial metrics that are commonly used in social perception research, sexual dimorphism, distinctiveness, bilateral asymmetry, and fWHR, derived from manually and automatically placed landmarks. As these shape-dependent measures are sensitive to scaling, translation, and rotation, we also examined these correlations between these manual and automatic landmarks after submitting them to a Generalized Procrustes Analysis (GPA; see Kleisner et al., 2014; Mitteroecker et al., 2015). Finally, to investigate the generalizability of our results across image sets, we investigated these correlations in two independent open-access image sets (DeBruine & Jones, 2017; DeBruine & Jones, 2020). In our second study, we investigated whether these facial metric generated from manual and automatic landmarks show any systematic biases when measured on faces of different ethnicities, to test whether automatic methods may be generalizable to different study populations without introducing biases that can be present in facial detection algorithms (O'Toole et al., 2012).

Study One - Estimating correlations between manual and automatic landmark measures

Method

All data and analyses (including code for calculating facial metrics) can be found on the Open Science Framework (osf.io/5e3qp). Analyses were conducted using Python 3.6 and JupyterLab notebooks that detail the measurements and

statistical analysis. We have also provided a tutorial notebook for automatic landmark of faces, also available on the Open Science Framework.

Image sets

The first open access image set used in our study was the Face Research Lab London Set (DeBruine & Jones, 2017). This image set consists of 102 faces (49 females, age *M* = 27.72 years, *SD* = 7.11 years) of various ethnicities. The second was the Three DSK image set (DeBruine & Jones, 2020). This image set consists of 100 White faces (50 females, age *M* = 24.25 years, *SD* = 3.98 years). Photographs were taken against a white background in both image sets, and distance to the camera was also standardised in both.

All faces were delineated by a single annotator to minimize inter-observer error. Landmarks were placed using Webmorph (DeBruine, 2017). The landmark template used was one built for transforming and averaging images, which includes a variety of anatomical landmarks (e.g. outer lip edges, widest point of the face, edges of nose, and so on) as well landmarks that are linked to soft tissue areas, such as the cheek bones and nasolabial folds. These semi-landmarks were not subjected to sliding procedures. Across images, these manually placed landmarks were aligned by inter-pupillary distance. The GPA conducted here translated, rotated, and scaled shapes, with these steps going some way to remove variations in size that are not accounted for by the standardisation in photographic capture and interpupillary alignment. However, as we do not have access to absolute size measures of the photographed individuals, we are unable to fully remove body size information (i.e., allometry), of which facial correlates have been shown to directly affect social perception independently of measures such as sexual dimorphism (Holzleitner et al., 2014).

Following previous research that used manually placed landmarks to generate facial metrics (e.g., Cai et al., 2018; Holzeitner et al., 2019), landmarks describing non-facial characteristics, such as hairstyle, that are not typically used to derive facial metrics, were removed. The average configuration of the remaining 164 landmarks is shown in Figure 1 (left panel).

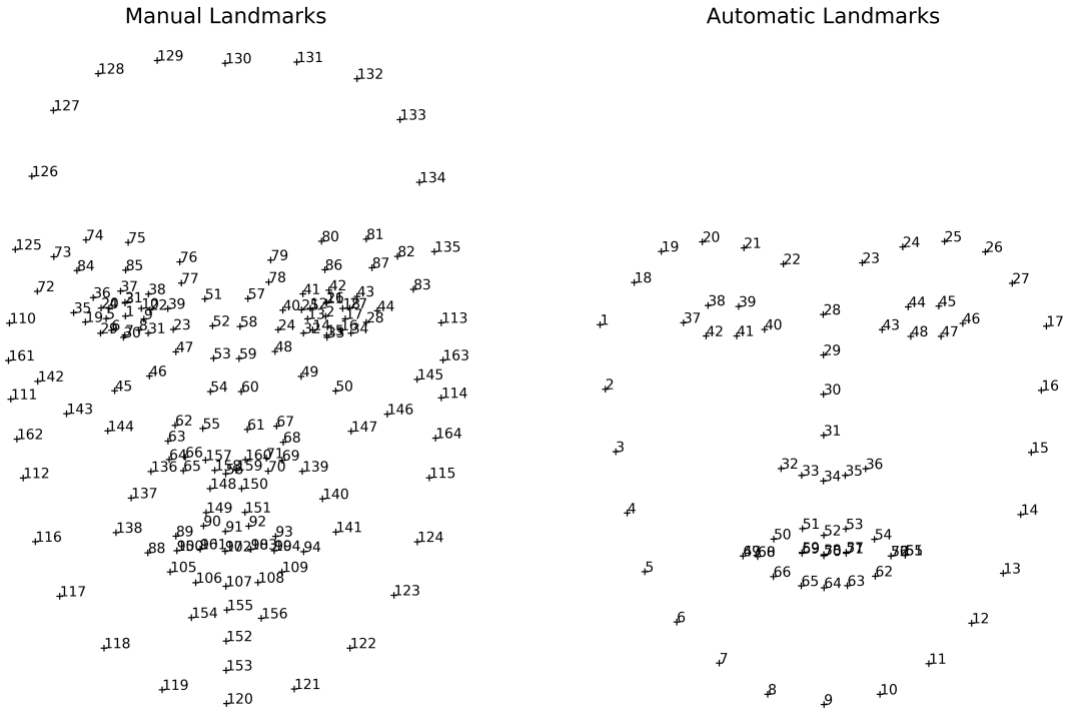

Figure 1. The average configuration of manually and automatically placed landmarks used to derive facial metrics in our study.

Automatic landmark placement

Each of the two image sets were automatically landmarked using the Python face recognition module (https://github.com/ageitgey/face_recognition), which is built on the Dlib machine learning package (King, 2009). Each face was detected and a set of 72 landmark points were placed, recovered, and saved to file. The average configuration of these 72 landmarks is shown in Figure 1 (right panel).

Measures

Each measure described below was taken twice, from the manual and automatically placed landmarks. Measurements were taken from the original landmarks (which retained some elements of rotation which is uncorrected for by interpupillary alignment, as well as scaling and translation differences), and once more from the landmarks resulting from a GPA of the original landmarks. Thus, for each face, there were 16 scores – one for each of four traits, under two landmark placement types, and for raw and Procrustes configurations.

Facial asymmetry. Following Jones and Jaeger (2019) and Komori et al. (2009), asymmetry of face shape was calculated using a method that treats the landmark coordinates as a vector in *n*-dimensional space (e.g. 328 dimensions for manually placed landmarks with 164 *xy* points, and 144 dimensions for automatically placed landmarks with 72 points). Asymmetry is then calculated as the distance between the original vector and a version of the vector that is mirror reflected about the origin. Greater distances between these vectors indicate greater asymmetry.

Facial distinctiveness. Also following Jones and Jaeger (2019) and Komori et al. (2009), distinctiveness of face shape was measured in the following steps. First, the vector representation of all faces of the same sex were averaged to produce an average male and female vector. Next, the vector representation of each individual face image's landmarks was subtracted from the average vector for that sex. The unsigned magnitude of this vector measures the distance of a given face from the average configuration. Greater distances indicate greater distinctiveness – the facial shape is further from the average configuration.

Facial sexual dimorphism. Following Jones and Jaeger (2019), sexual dimorphism of face shape was measured with a vector projection approach (Mitteroecker et al., 2015), using multivariate regression. Biological sex (male faces

coded as zero, female faces coded as one) was first regressed against the shape vectors for each face. This analysis produces a coefficient vector that describes how facial shape changes with biological sex. Each individual face's vector was then projected onto this axis, resulting in an objective dimorphism score that indicates how far along the dimorphism axis a face is. Greater scores indicate greater femininity. For each face, sexual dimorphism was calculated twice (once from the manually placed landmarks and once from the automatically placed landmarks).

Measuring fWHR. Following Zhang et al. (2018) and Lefevre et al. (2013), face width was first calculated as the Euclidean distance between the landmarks describing bizygomatic width. Face height was then calculated as the distance between the averaged points describing the top lip and the averaged points describing the highest arch of each eyebrow. fWHR was then calculated for each face from these measurements. For each face, fWHR was calculated twice (once from the manually placed landmarks and once from the automatically placed landmarks).

Power and analytical strategy

As we used open-access databases, we were limited in the sample of faces available, and thus the number of observations available for statistical tests. As such, we conducted a sensitivity analysis to give us an estimate of the smallest correlation we could detect. For a simple correlation between measures derived from measures placed manually and automatically, with alpha set to .05 and beta at .80, and with 100 observations (our smallest sample size), we can detect a correlation of .27. To provide a convincing argument for the use of automatically placed landmarks to measure faces, we would expect to see correlations far higher than this – detecting such a small effect would suggest there is considerable divergence between the placement types.

As such, our sample size allows for comfortable detection of large and meaningful correlations.

Results

For each of the four facial metrics (sexual dimorphism, distinctiveness, bilateral asymmetry, and fWHR), we computed the Pearson correlation between scores generated from manually and automatically placed landmarks. We did this separately for each raw and Procrustes landmark coordinates, and separately for each image set for internal replication purposes. Figure 2 shows these correlations in full, which were all statistically significant at $p < .001$, and the results are summarized in Table 1.

While correlations between metrics derived from manually and automatically placed landmarks were generally very high, ranging from .654 to .996 for standard landmarks, and from .558 to .930 for Procrustes aligned landmarks, the correlation for distinctiveness was substantially lower in the Three DSK set than it was in the London Set. By using ordinary least squares to predict Manual distinctiveness scores from Automatic distinctiveness scores in the Three DSK face set, we identified the only face with a studentized residual above or below ± 3 (for a model using standard landmarks, score = -10.42, for the model with Procrustes landmarks, the same face had a score of = -4.31). Examining its automatic landmark configuration revealed that the face detector had made significant errors in placing points along the jaw and mouth.

Consequently, we first recomputed the correlations in the Three DSK set when this face was removed. Correlations increased for all standard landmark measures – fWHR $r = .873$, distinctiveness $r = .852$, symmetry $r = .926$, and dimorphism $r = .911$. For Procrustes landmarks, only symmetry $r = .863$, and distinctiveness $r = .639$,

increased. Dimorphism did not change, nor did fWHR (which is identical between Procrustes and standard landmarks, as it relies on relative distances between individual points). All correlations remained significant at $p < .001$.

We also conducted a simple geometric morphometrics analysis to assess the relationship between the way the different landmark sets captured facial form. To do this, we took the Procrustes fitted landmarks in each set, for manual and automatic placements, and submitted them to a principal components analysis, limiting the analysis to the first two components. We then correlated the manual components with their automatic counterparts. Geometrically, this operation is equivalent to finding the angles between these components. That is, if the resulting landmark arrangement PC's capture similar information about the faces, they should be significantly correlated. These correlations may be positive or negative, as direction of principal components is arbitrary. For the London set, the manual and automatic PC1 was strongly correlated, $r = .89$ (26.88º), $p < .001$, as was PC2, $r = .81$ (36.19º), $p < .001$. For the Three DSK set, the manual and automatic PC1 was strongly negatively correlated $r = -.87$ (150.62º), $p < .001$. The correlation here for PC2 was also significant, but was somewhat weaker, $r = -.51$ (120.93º), $p < .001$. These correlations indicate, particularly for the maximal axes of variability in faces, that both landmark sets capture similar variance.

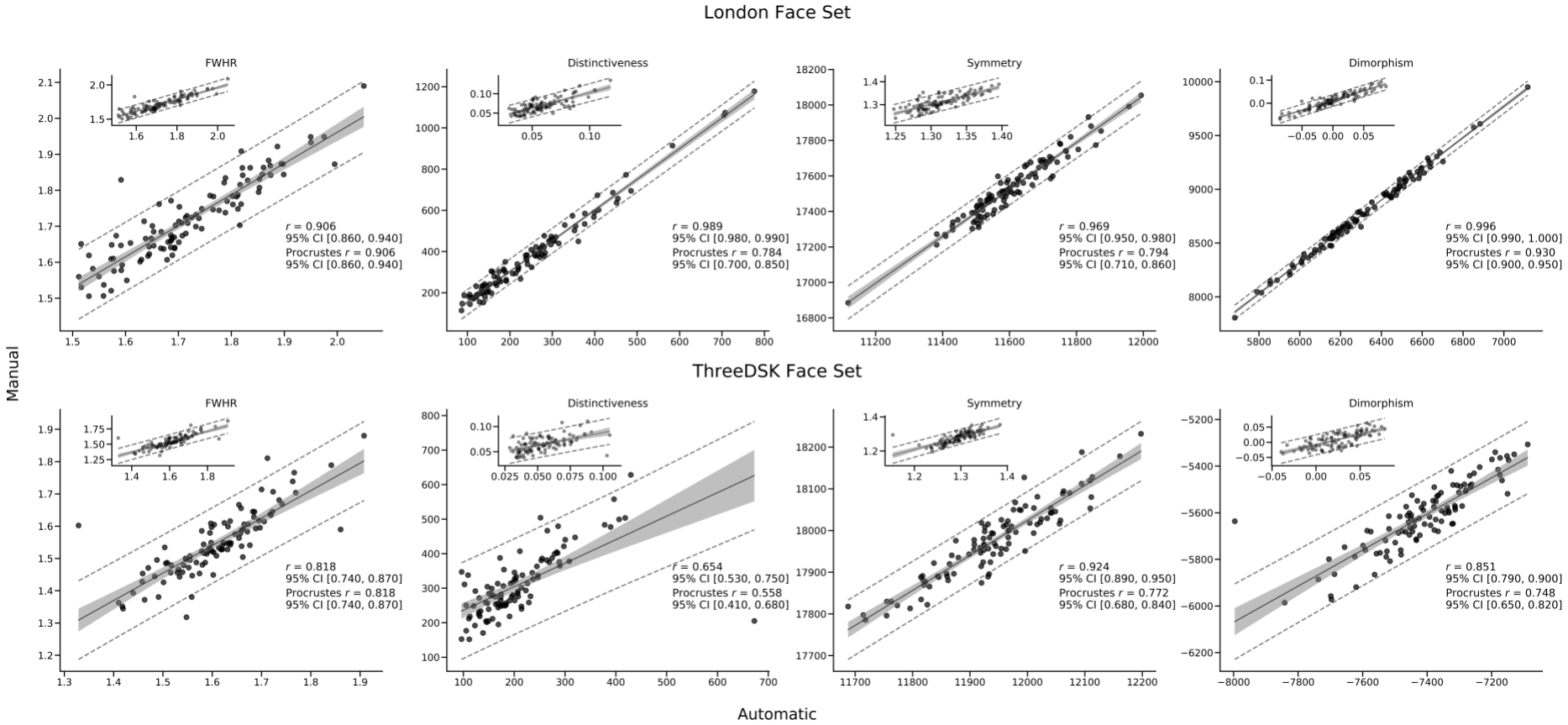

Figure 2. Correlations between measures produced from manually and automatically placed landmarks for the Face Research Lab London (top row) and Three DSK (bottom row) image sets. Main axes represent the standard landmarks, while inset axes represent Procrustes-fitted landmarks. The solid black lines represent the ordinary least squares fit, the dashed lines represent the 95% confidence interval of the coefficient, and the shaded areas represent the 95% prediction intervals (where new predicted values would fall).

Study Two - Testing for potential biases in automatic landmark placement

We have demonstrated that strong correlations emerge between commonly used facial metrics measured from manual and automatically placed landmarks. Aside from errors in automatic landmarking on certain faces, automatic placement appears to be accurate and capable of deriving metrics of interest. However, automatic landmark placement of the kind leveraged here is a critical step in face detection and recognition algorithms (Damer et al., 2019; Juhong & Pintavirooj, 2017; Köstinger et al., 2011; Shi et al., 2006). Moreover, there has been controversy and research around how these algorithms are biased in a multitude of ways. Ethnicity is a salient example, with findings indicating variability in face recognition algorithms for faces of different ethnicities (O'Toole et al., 2012), with studies demonstrating poorer performance on different demographic cohorts (Klare et al., 2012). This issue has received significant attention in computer vision (Abdurrahim et al., 2018; Garcia et al., 2019), and is an active area of research (Wang & Deng, 2020). While face recognition networks are comprised of multiple steps, the algorithms that find and place landmarks may be a generator of these biases. Thus, automaticing landmarking may introduce systematic errors in labelling that could bleed into metrics calculated from these landmarks. If present, this bias would significantly hamper the use of automatic landmarks for facial metrics, as they may induce spurious correlations between metrics for certain demographics. To be clear, we do not claim that ethnicity is the only factor that face detection algorithms may induce bias on, when other prominent examples include age and sex (Das et al., 2018), but we focus here on a critically important factor for both computer vision and psychological research.

In the following study, we included two separate samples of White and East Asian faces and examined the correlation between metrics derived from automatic

and manually placed landmarks for these faces. We test how closely automatic landmark measures approximate manual measures and, importantly, whether this approximation is moderated by face ethnicity. Indeed, it has been shown that some facial recognition algorithms have poorer accuracy for East Asian faces compared to White faces (Cavazos et al., 2020). As such, this study provides a useful test of wheter landmark placement seems systematically different between these ethnicities, and will provide insight into the extent these landmarks can be utilised for deriving facial metrics.

Method

Image sets

We used an image database reported in Zhang et al. (2019), which comprised 100 East Asian and 100 White individuals in their mid twenties. There was an even split of women and men in each ethnicity. Faces were manually landmarked in the same manner as study one, using the same set of landmarks to outline facial shape, which were aligned on interpupillary distance. All individuals were photographed facing the camera with a neutral expression.

Automatic landmark placement

Faces were landmarked using the same face detector and landmark set as study one.

Measures

The four measures from each face were calculated in exactly the same way as in the initial study, separately for each ethnicity. Given the generally strong associations between standard and Procrustes aligned shapes in study one, we restrict our analysis here to the results of standard landmarks. If biases are found, then it is this landmark representation that must be corrected.

Analytical strategy

To test for biases between ethnicity and automatically generated metrics, we used linear regression to model manual landmark metrics as a linear function of the mean-centred automatic landmark metric, ethnicity (dummy coded, White faces as zero, East Asians as one), and the interaction between the two of these predictors. Here, the interaction is the key test, as a significant coefficient will indicate a different strength of approximation in automatic landmarking metrics between ethnicities. Main effects of ethnicity may represent differences between ethnicities in certain facial metrics (Danel et al., 2018; Fang et al., 2011). Where a significant interaction is present, we use analysis of variance to examine the amount of variance it explains compared to a model with only the main effects (that is, automatic and ethnicity predictors).

Results

The main regression results are presented in Table 2. Across each trait, the models explained significant variance in the manual metrics, and had low root-mean-squared-error (*RMSE*). For the symmetry and distinctiveness models, we observed no significant main effects or interactions with ethnicity, and the magnitude of the coefficient of automatic ethnicity serves as a measure of the accuracy of the approximation in interpretable units – for example, the coefficients of the automatic landmark measurement vary between 0.603 and 1.476, meaning the automatic metrics both under and overestimate the manual landmarks, but all within a range of close to one unit – a coefficient of 1 would represent a direct one-to-one mapping. In all cases the *RMSE* of the predictions were small on the actual scale of the metrics.

Measures of symmetry and distinctiveness were unrelated to ethnicity and its interaction with automatic measures. For fWHR and dimorphism, we observed a

significant interaction, indicating the automatic landmark approximation of manual metrics differed between ethnicities, and we consider these in detail

fWHR. For fWHR, the automatic coefficient represents the slope (and thus approximation of manual metrics) for White faces, $\beta_1 = 0.603$. Thus, for East Asian faces, the approximation is significantly higher, equal to the sum of the main effect fit and interaction coefficients ($\beta_1 + \beta_3 = 0.894$). fWHR is thus approximated more closely in East Asian faces by automatic landmarks than in White faces. Using a model comparison approach (Type III sums of squares), we computed the magnitude of the variance this interaction term - and thus difference between ethnicities in landmarking approximation – contributed to the model, which was small but significant $\Delta R^2 = 0.024$, $F(1, 196) = 13.14$, $p < .001$. Finally, we examined the fWHR model for faces with high ($> \pm 3$) studentized residuals. Three White faces showed very large errors above three (studentized residuals = 7.36, 4.98, -3.96), and inspecting their automatic landmarks revealed errors in placement of the mouth area. Omitting these faces and refitting the model led to the interaction becoming non-significant ($p = .057$), and a much higher overall model fit, adjusted $R^2 = .787$.

Dimorphism. For sexual dimorphism, the main effect of ethnicity (reflecting mean differences in dimorphism between East Asian and White faces) was a strong predictor. Nonetheless, the interaction was significant, and the coefficient for automatic landmark dimorphism showed a tendency to over-estimate manual landmark dimorphism, $\beta_1 = 1.307$. The interaction term again showed East Asian faces were more closely approximated than White faces ($\beta_1 + \beta_3 = 1.19$). A model comparison approach revealed the interaction term contributed a significant yet very small proportion of variance, $\Delta R^2 < 0.001$, $F(1, 196) = 8.33$, $p = .004$. Examining the dimorphism model for faces with high studentized residuals again revealed three

White faces (two of which were the same cases as in the fWHR model, scores = 6.12, 4.17, 3.09). Removing them from the model did not affect the significance of the interaction.

The predictions of the models with interactions are shown in Figure 3.

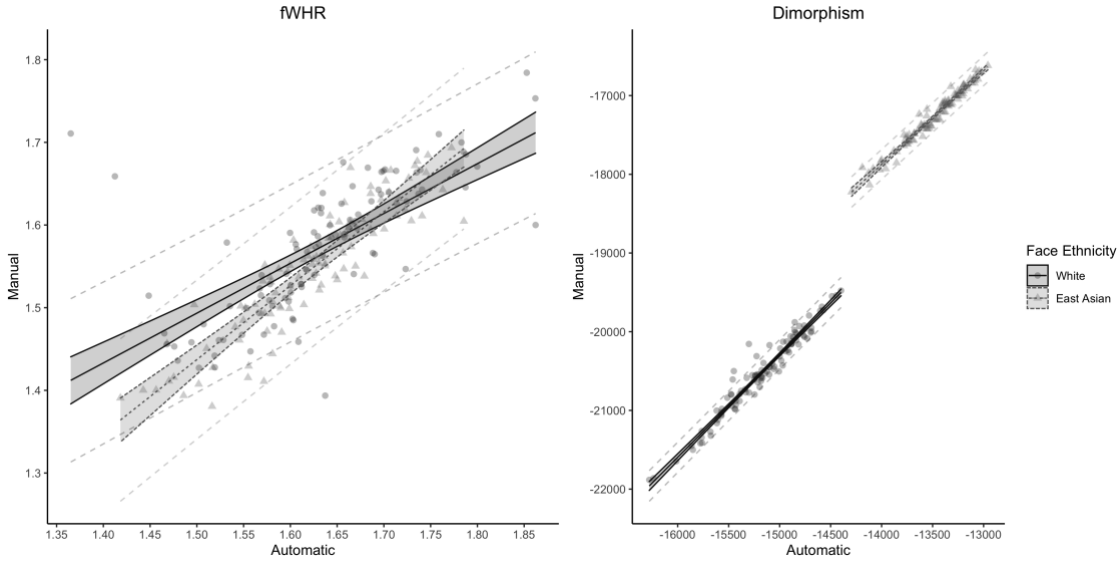

Figure 3. Model fits of the fWHR (left) and dimorphism (right) models illustrating the interactions between automatic landmark metrics and ethnicity on predicting manual landmark metrics. Dashed lines represent the 95% confidence interval of the coefficient, and the shaded areas represent the 95% prediction intervals (where new predicted values would fall).

Discussion

The current studies used several independent image sets to investigate the correlations between four facial metrics commonly used in social perception research (sexual dimorphism, distinctiveness, bilateral asymmetry, and fWHR) when they were derived from manually and automatically placed landmarks, as well as estimating the degree of bias that may occur if these landmarking procedures are used on faces of different ethnicities.

Figure 2 highlights the main finding that, across both image sets and all four facial metrics, and under the raw and Procrustes shape representations, correlations between measures derived from automatically and manually placed landmarks were

high. An encouraging perspective of these findings is to compare them with reported correlations in the literature, which detail the reliability of measures taken from two independent researchers placing landmarks separately. For example, these have been reported as high as r = .87 for fWHR (Geniole et al., 2014), r = .80 and r = .85 for asymmetry (Grammer & Thornhill, 1994; Rikowski & Grammer, 1999), and r = .85 for sexual dimorphism (Scheib et al., 1999). That we find similar values here is relatively unsurprising, as earlier efforts in computer vision research achieved accuracy in landmark placement of within five pixels to manually labelled images (e.g., Efraty et al., 2011). These results are therefore good evidence that automatic landmarks closely approximate manual landmarks, and the measures derived from those automatic landmarks are similar to those from manual labels.

Across both image sets, the correlations between manual and automatic landmark measures were lower for Procrustes landmarks, indicating a divergence in automatic and manual measures when shapes had been translated, scaled, and rotated. This makes geometric sense across measures. Using symmetry as an example, consider a face that is posing with a slight head tilt. The reflection of its landmarks to measure symmetry would yield a greater asymmetry score, as some of the asymmetry is attributable to simple head tilt. The fewer landmarks placed by face detectors will capture less of the variability in actual morphometry here, and thus be more sensitive to head tilt, while the greater number of landmarks in the manual condition will be less sensitive and offer a better measure.

For measures like distinctiveness, the translation, rotation, and scaling may contribute to the measure of distinctiveness itself – a face can be distinctive due to its rotation compared to the mean configuration and not due to any morphology divergences. In Procrustes space, when these factors are removed, the reduced

number of landmarks afforded by automatic placement could be a somewhat weaker approximation of appearance. Dimorphism measures may similarly be affected as the distance between female and male average configurations may change with Procrustes analysis. However, here we were unable to remove shape variation that is linked with body size – that is, allometry. This has been shown to vary with certain metrics like dimorphism (Mitteroecker et al., 2015), and indeed can affect social perception independently of measures like sexual dimorphism (Holzleitner et al., 2014). Thus, we cannot rule out that the differences that emerged between the raw and Procrustes configurations (which have been scaled) could be linked to this source of variation. Despite this, the correlations for measures derived from Procrustes analysis are still high, and we again urge caution in checking the landmark configurations if users utilise this method.

A risk of systematic bias in facial recognition algorithms, which rely heavily on automatic landmark placements, is apparent and well researched (Cavazos et al., 2020; Das et al., 2018). In our second study, we found that, for symmetry and distinctiveness measures, automatically derived metrics were a close approximation (though slightly upwardly biased) of manual metrics, and showed no evidence of significantly interacting with ethnicity. However, for fWHR and sexual dimorphism, this interaction was significant. Contrary to popular expectations of algorithmic bias, we found that fWHR and dimorphism were more poorly approximated in White faces. Further investigating these differences by way of outlier analysis revealed White faces with significant automatic landmarking errors. Removing these for fWHR resulted in a more equal approximation between ethnicities, but not for dimorphism.

Our results appear to validate the use of automatically placed landmarks for deriving facial metrics to employ in social perception research, but with several very

important caveats. This validation is important for two reasons. First, it suggests that automatically placed landmarks can be used to derive facial metrics, removing the substantial time costs that the manual placement method is subject to and, potentially, allowing researchers to use this timesaving to increase the number of faces tested. Second, it suggests that automatically placed landmarks could be used to produce more replicable facial metrics, by contrast with those derived from manually placed landmarks. Automatically placed landmarks might also be usefully employed for other common methods in face research that require placement of landmarks, such as averaging and transforming face images.

The caveats are both practical and theoretical. The first is that, in both of our studies, the automatic landmarks did not always delineate faces correctly, and substantial errors were made on a small percentage of faces (< 2%). We analysed our data without inspecting these landmarks directly, taking with us a naïve assumption in our inferential approach, which we corrected with outlier detection. A primary conclusion is that researchers who wish to leverage the rapid generation of landmarks cannot treat them as error-free, and each face should be carefully checked before continuing with analysis. The second and most important caveat is that, while we found mild evidence for bias in automatic landmark placement across two ethnicities, without testing multiple other potential sources of bias, such as age or different ethnicities, there is no guarantee that automatic landmark placement does not have biases that can influence metrics or tests derived from those landmarks. Conversely, while there is no empirical evidence that we are aware of that shows human raters are unbiased when manually landmarking faces of different ethnicities, ages, or various other parameters, we suggest that it is premature to rule out other biases that automatic landmarking may induce within social perception research. Likely, the best

approach is semi-supervised landmarking, whereby automatic landmarks are critically and carefully checked by researchers. Fortunately, this is likely to be a significantly faster process than manual delineation and with very similar outcomes.

We have provided a tutorial for the installation and use of the software we employed for automatic placement of facial landmarks, together with the code that we used to calculate the facial metrics we investigated, on the Open Science Framework (osf.io/5e3qp). We hope that these resources will allow researchers to more easily make use of the many large image sets now being made open access and more easily increase the sample sizes they employ in research using facial metrics. In doing so, we hope to see substantial improvements in the reliability of social perception research employing facial metrics.

References

Abdurrahim, S. H., Samad, S. A., & Huddin, A. B. (2018). Review on the effects of age, gender, and race demographics on automatic face recognition. *The Visual Computer*, *34*(11), 1617–1630. https://doi.org/10.1007/s00371-017-1428-z

Baddar, W. J., Son, J., Kim, D. H., Kim, S. T., & Ro, Y. M. (2016). A deep facial landmarks detection with facial contour and facial components constraint. *2016 IEEE International Conference on Image Processing (ICIP)*, 3209–3213. https://doi.org/10.1109/ICIP.2016.7532952

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *ArXiv:1912.07398 [Cs]*. http://arxiv.org/abs/1912.07398

Damer, N., Boller, V., Wainakh, Y., Boutros, F., Terhörst, P., Braun, A., & Kuijper, A. (2019). Detecting Face Morphing Attacks by Analyzing the Directed Distances of Facial Landmarks Shifts. In T. Brox, A. Bruhn, & M. Fritz (Eds.), *Pattern Recognition* (pp. 518–534). Springer International Publishing. https://doi.org/10.1007/978-3-030-12939-2_36

Danel, D. P., Valentova, J. V., Sánchez, O. R., Leongómez, J. D., Varella, M. A. C., & Kleisner, K. (2018). A cross-cultural study of sex-typicality and averageness: Correlation between frontal and lateral measures of human faces. *American Journal of Human Biology*, *30*(5), e23147. https://doi.org/10.1002/ajhb.23147

Das, A., Dantcheva, A., & Bremond, F. (2018). *Mitigating Bias in Gender, Age and Ethnicity Classification: A Multi-Task Convolution Neural Network Approach*. 0–0. http://openaccess.thecvf.com/content_eccv_2018_workshops/w5/html/Das_Miti

gating_Bias_in_Gender_Age_and_Ethnicity_Classification_a_Multi-Task_ECCVW_2018_paper.html

DeBruine, L. (2017). *WebMorph*. https://webmorph.org/#P23669

DeBruine, L., & Jones, B. C. (2020). *3DSK face set with webmorph templates*. https://doi.org/10.17605/OSF.IO/A3947

DeBruine, L. M., & Jones, B. C. (2017). *Face Research Lab London Set.* https://doi.org/10.6084/m9.figshare.5047666.v3

Efraty, B. A., Papadakis, M., Profitt, A., Shah, S., & Kakadiaris, I. A. (2011). Facial component-landmark detection. *Face and Gesture 2011*, 278–285. https://doi.org/10.1109/FG.2011.5771411

Fang, F., Clapham, P. J., & Chung, K. C. (2011). A Systematic Review of Inter-ethnic Variability in Facial Dimensions. *Plastic and Reconstructive Surgery*, *127*(2), 874–881. https://doi.org/10.1097/PRS.0b013e318200afdb

Garcia, R. V., Wandzik, L., Grabner, L., & Krueger, J. (2019). The Harms of Demographic Bias in Deep Face Recognition Research. *2019 International Conference on Biometrics (ICB)*, 1–6. https://doi.org/10.1109/ICB45273.2019.8987334

Holzleitner, I. J., Hunter, D. W., Tiddeman, B. P., Seck, A., Re, D. E., & Perrett, D. I. (2014). Men's Facial Masculinity: When (Body) Size Matters. *Perception*, *43*(11), 1191–1202. https://doi.org/10.1068/p7673

Jones, A. L., & Jaeger, B. (2019). Biological Bases of Beauty Revisited: The Effect of Symmetry, Averageness, and Sexual Dimorphism on Female Facial Attractiveness. *Symmetry*, *11*(2), 279. https://doi.org/10.3390/sym11020279

Juhong, A., & Pintavirooj, C. (2017). Face recognition based on facial landmark detection. *2017 10th Biomedical Engineering International Conference (BMEiCON)*, 1–4. https://doi.org/10.1109/BMEiCON.2017.8229173

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., & Jain, A. K. (2012). Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

Kleisner, K., Chvátalová, V., & Flegr, J. (2014). Perceived intelligence is associated with measured intelligence in men but not women. *PLOS ONE*, *9*(3), e81237. https://doi.org/10.1371/journal.pone.0081237

Köstinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2144–2151. https://doi.org/10.1109/ICCVW.2011.6130513

Mitteroecker, P., Windhager, S., Müller, G. B., & Schaefer, K. (2015). The Morphometrics of "Masculinity" in Human Faces. *PLOS ONE*, *10*(2), e0118374. https://doi.org/10.1371/journal.pone.0118374

O'Toole, A. J., Phillips, P. J., An, X., & Dunlop, J. (2012). Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, *30*(3), 169–176. https://doi.org/10.1016/j.imavis.2011.12.007

Özseven, T., & Düğenci, M. (2017). Face recognition by distance and slope between facial landmarks. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–4. https://doi.org/10.1109/IDAP.2017.8090258

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). *FaceNet: A unified embedding for face recognition and clustering*. *1*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Scott, N. J., Kramer, R. S. S., Jones, A. L., & Ward, R. (2013). Facial cues to depressive symptoms and their associated personality attributions. *Psychiatry Research*, *208*(1), 47–53. https://doi.org/10.1016/j.psychres.2013.02.027

Shi, J., Samal, A., & Marx, D. (2006). How effective are landmarks and their geometry for face recognition? *Computer Vision and Image Understanding*, *102*(2), 117–133. https://doi.org/10.1016/j.cviu.2005.10.002

Sutherland, C. A. M., Rhodes, G., & Young, A. W. (2017). Facial image manipulation: A tool for investigating social perception. *Social Psychological and Personality Science*, *8*(5), 538–551. https://doi.org/10.1177/1948550617697176

Wang, M., & Deng, W. (2020). *Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning*. 9322–9331. http://openaccess.thecvf.com/content_CVPR_2020/html/Wang_Mitigating_Bias_in_Face_Recognition_Using_Skewness-Aware_Reinforcement_Learning_CVPR_2020_paper.html

Zhang, L., Holzleitner, I. J., Lee, A. J., Wang, H., Han, C., Fasolt, V., DeBruine, L. M., & Jones, B. C. (2019). A Data-Driven Test for Cross-Cultural Differences in Face Preferences: *Perception*. https://doi.org/10.1177/0301006619849382

Table 1. Correlations between automatic and manual landmark placement derived facial metrics.

Set	Landmark Representation	fWHR	Distinctiveness	Symmetry	Dimorphism
London	Standard	0.906 [0.86, 0.94]	0.989 [0.98, 0.99]	0.969 [0.95, 0.98]	0.996 [0.99, 1.0]
	Procrustes		0.784 [0.7, 0.85]	0.794 [0.71, 0.86]	0.930 [.90, 0.95]
ThreeDSK	Standard	0.818 [0.74, 0.87]	0.654 [0.53, 0.75]	0.924 [0.89, 0.95]	0.851 [0.79, 0.9]
	Procrustes		0.558 [0.41, 0.68]	0.772 [0.68, 0.84]	0.748 [0.65, 0.82]

Note. All correlations significant at $p < .001$. Bracketed values indicated 95% confidence intervals.

Table 2. Results of multiple regression for each of the four facial metrics.

Metric	Model fit $F(3, 196)$	Adj R^2	*RMSE*	β_0 Intercept	β_1 Automatic	β_2 Ethnicity	β_3 Interaction
fWHR	121.20	0.644	0.047	1.570* [1.560, 1.579]	0.603* [0.502, 0.703]	-0.019* [-0.033, -0.006]	0.291* [0.133, 0.450]
Distinctiveness	541.30	0.891	78.60	633.972* [618.251, 649.694]	1.358* [1.263, 1.452]	1.861 [-20.371. 24.093]	-0.017 [-0.150, 0.115]
Symmetry	2855.10	0.977	135.64	23234.831 [23207.796, 23261.865]	1.476* [1.434, 1.518]	-31.206 [-69.441, 7.029]	-0.002 [-0.066, 0.061]
Dimorphism	21222.80	0.997	93.45	-19412.409* [-19457.522, -19367.269]	1.307* [1.258, 1.356]	1145.477* [1073.077, 1217.878]	-0.117* [-0.197, -0.037]

Note. Coefficients with asterisks are significant, highest observed p-value = 0.0051