MeshMonk: open-source large-scale intensive 3D phenotyping

Julie D. White[1*], Alejandra Ortega-Castrillón[2,3], Harold Matthews[4,5,6], Arslan A. Zaidi[1,7], Omid Ekrami[8], Jonatan Snyders[9], Yi Fan[4,10], Tony Penington[4,5,6], Stefan Van Dongen[8], Mark D. Shriver[1], Peter Claes[2,3,4*]

[1]Department of Anthropology, The Pennsylvania State University, University Park, PA, USA.

[2]Department of Electrical Engineering, KU Leuven, Leuven, Belgium

[3]Medical Imaging Research Center, UZ Leuven, Leuven, Belgium

[4]Murdoch Children's Research Institute, Melbourne, Australia

[5]Royal Children's Hospital, Melbourne, Australia

[6]Department of Pediatrics, University of Melbourne, Melbourne, Australia

[7]Department of Biology, The Pennsylvania State University, University Park, PA, USA.

[8]Department of Biology, University of Antwerp, Antwerp, Belgium

[9]WebMonks, Hasselt, Belgium

[10]Melbourne Dental School, University of Melbourne, Melbourne, Australia

*** Correspondence:**
jdw345@psu.edu; peter.claes@kuleuven.be

Keywords: Automated landmarking[1], automated phenotyping[2], non-rigid registration[3], phenomics[4], genomics[5], morphometrics[6], 3D[7], facial variation[8].

Abstract

Introduction

In the post-genomics era, an emphasis has been placed on disentangling 'genotype-phenotype' connections so that the biological basis of complex phenotypes can be understood. However, our ability to efficiently and comprehensively characterize phenotypes lags behind our ability to characterize genomes. Anthropometric studies of morphology have traditionally relied on sparse sets of landmarks manually placed on images, which is tedious, error-prone, and sensitive to individual differences among observers. Here, we report a toolbox for fast and reproducible high-throughput phenotyping of 3D images. While we demonstrate this toolbox using 3D facial images, the procedure can also be applied to 3D images of other complex morphological structures, such as bones.

Methods

Given a facial image (target), a rigid registration is first used to orient a template to the target scan. Then, using a symmetrical weighted k-nearest neighbors and a visco-elastic transformation model, the reference is transformed to fit the specific shape of the target. For facial scans, this results in homologous spatially dense (N=7,160) quasi-landmark configurations for all 3D images. As validation, a dataset (N=41) with 19 manually-placed landmarks was registered using MeshMonk and the manually placed landmarks were aligned to the template scan to identify the closest coordinate on the template. In a leave-one-out approach, the position of the manual landmarks on the template were averaged and this average was then placed back on the left-out face, resulting in an automatic indication of the sparse landmarks for comparison.

Results and Conclusion

We demonstrate that this method is highly accurate, with an average root mean squared error between the manual and automatic placements of 0.62 mm and no variation in landmark position or centroid size significantly attributable to landmarking method used. Though validated using 19 landmarks, for comparison with traditional methods, MeshMonk allows for automated dense phenotyping, freeing the researcher from the use of a limited number of landmarks and allowing for more comprehensive investigations of 3D shape variation. This expansion opens up an exciting avenue of study in assessing genomic and phenomic data to better understand the genetic contributions to complex morphological traits.

1 Introduction

The phenotypic complement to genomics is *phenomics*, which aims to obtain high-throughput and high-dimensional phenotyping in line with our ability to characterize genomes (Houle et al., 2010). The paradigm shift is simple and similar to the one made in the Human Genome Project: instead of 'phenotyping as usual' or measuring a limited set of simplified features that seem relevant, why not measure it all? In contrast to genomic technologies, which successfully measure and characterize complete genomes, the scientific development of phenomics lags behind. However, with the advent of new technologies, hardware exists for extensively and intensively collecting quantitative phenotypic data. For example, 3D image surface and/or medical scanners provide the optimal means to capture information of biological morphology and appearance. Today, the challenge is to establish standardized and comprehensive phenotypic representations from large scale image data that can be used to study phenotypic variation in the context of genetic variation (Walter et al., 2010). This is a challenge that we address with the development of the MeshMonk toolbox.

Dense correspondence phenotyping is important beyond genomics and could be employed by anthropologists, biologists, and medical clinicians to accurately and reproducibly characterize anatomical structures, like a femur, skull, or face, such that underlying qualities about the structure can be understood. The study of variation and covariation in anatomy can provide insights into the genetic causes and evolution of the anatomical structure. In addition, comparing the anatomy of an individual patient to a control population can indicate pathology to a medical practitioner. Traditionally, this has been achieved using visual clinical assessment or by taking measurements between manually placed anatomical 'landmarks', traditionally defined as precise locations on biological forms that hold some developmental, functional, structural, or evolutionary significance (Richtsmeier et al., 2002) and are unambiguously defined and reliably locatable (Aldridge et al., 2005; Corner et al., 1992; Richtsmeier et al., 1995). Some examples include the endo- and

exocanthi (the inner and outer corners of the eyes, respectively) and the pronasale (the tip of the nose).

However, manual landmarking is tedious to perform, difficult to standardize in practice, and prone to intra and inter-operator error (Fagertun et al., 2014; Toma et al., 2009; von Cramon-Taubadel et al., 2007; Weinberg et al., 2004; Wong et al., 2008). Furthermore, sparse landmark configurations can only quantify form at defined landmarks that can be reliably identified and indicated by a human and thus lack the resolution to fully characterize shape variation in between landmarks. An alternative is to automatically indicate quasi-landmarks across the entire surface of the structure. This is achieved by gradually warping a generic template (i.e. anthropometric mask) composed of thousands of points into the shape of each target image through a non-rigid registration algorithm (Andresen and Nielsen, 2001; Claes, 2007; Claes et al., 2012b; Hutton et al., 2003b; Snyders et al., 2014). The coordinates of these warped templates, now in the shape of each target, can then be assessed in geometric morphometric analysis. An automatic approach like this is preferable for the analysis of large datasets, avoiding the problems of manual landmarking at different sites by multiple operators. They are also more suitable for applications that require synthesis of a recognizable instance of the actual structure, such as predicting a complete shape from DNA (Claes et al., 2014), synthetic growth and ageing of a face (Imaizumi et al., 2015; Matthews et al., 2018), constructing 3D facial composites for forensic applications (Blanz and Vetter, 1999), and characterization of dysmorphology for clinical diagnosis (Baynam et al., 2015; Hammond et al., 2005). Here, we report the MeshMonk toolbox for fast and reproducible high-throughput phenotyping of 3D images, or quasi-landmark indication, which can be applied to 3D facial images as well as 3D scans of other complex morphological structures.

Surface registration, implemented in the MeshMonk toolbox, defines a warping of the vertices from one (template) image to their corresponding locations on another (target) and allows us to quantify and visualize both subtle and acute variation in surface form across a sample by finding the geometrical relationship (one-to-one correspondences) between 3D shapes (Andresen and Nielsen, 2001; Claes, 2007; Claes et al., 2012b; Hutton et al., 2003a; Snyders et al., 2014). The registration strategy is akin to fitting an elastic net onto a solid facial statue through a geometry-driven mapping of anatomically corresponding features. When the template is warped onto each target, the coordinates of any anatomical landmark, manually annotated on the template, can also be defined on each target, thus the complete quasi-landmark indication can also be considered a method for automatic placement of sparse anatomical landmarks (Wei et al., 2011). As a validation of the MeshMonk toolbox, we compare manual and automatic indications of a set of 19 sparse landmarks.

2 Materials and Methods

Registering a template surface to a target in a manner unique to the target is possible if for each point on the template a corresponding point on the target surface is known, or if an appropriate transformation model is known for registering each point on the template to the corresponding point on the target. Before the actual template registration takes place, however, the correspondences and transformation model are unknown, and therefore the registration procedure involves an iterative solution in which both the correspondences and transformation model update each other sequentially over each iteration step. Such an iterative approach was first introduced in the popular iterative closest point (ICP) algorithm (Besl and McKay, 1992), embedding into the

registration a joint optimization problem where the distance between the template's point set and the respective correspondences on the target is minimized in every iteration until an optimal minimum is reached. This also forms the basis of the surface registration implemented in MeshMonk, in which a specific symmetrical correspondence searching strategy and rigid as well as non-rigid transformation models are provided. The core functionality of the toolbox is implemented in C++, with a focus on computational speed and memory to enable the processing of large 3D images. Interaction with the toolbox is also provided using Matlab™, enabling an easy to use implementation and visualization environment for the user.

2.1 Explanation of process

Snyders et al., 2014, demonstrated that the best generic registration method is a combination of symmetrical weighted k-neighbor correspondences and a visco-elastic transformation model. A schematic of the complete surface registration algorithm is presented in Figure 1 and screenshots of the process on an example face are presented in Figure 2. A short video of the registration on this example face is also available in the Supplementary Information. To initiate the process, a rigid registration developed from the original ICP algorithm is performed. This will adapt the position, orientation and scale of the template to better align to the target surface (Figure 2B). Subsequently, a non-rigid registration is done that will alter the shape of the template to match the shape of the target surface (Figure 2C).

Finding correspondences: At any iteration during the process, for both the rigid and non-rigid registration steps, correspondences are updated by using pull-and-push forces (symmetrical correspondences) (Redert et al., 1999) and a weighted k-neighbor approach (SI Figure 1). The symmetrical correspondences are calculated by combining two affinity matrices: 1) from template points to points on the target surface (push forces – the typical one-to-one correspondences calculation), and 2) from target points to points on the template surface (pull forces). This ensures that potential protrusions present on the target surface are allowed to pull the corresponding structure on the template, as illustrated in SI Figure 1C. Binary correspondences are avoided by using the weighted k-nearest neighbor rule, allowing correspondence to be defined as an interpolation between existing surface points (anywhere on the surface). For each point, the k closest points on the opposite surface are searched for, and the inverse of the distance to each of its closest points is coded as a weight in the affinity matrix. The weights for the remaining non-closest points are set to zero. The distance to the closest points can be computed in terms of 3D position only or a combination of 3D position and 3D normal orientation in each point, rendering 6D distances as a definition for "closeness". The incorporation of the normal information better matches points with a similar orientation and avoids the inappropriate matching of opposite oriented points. For example, in skeletal surface data, the inner and outer surface have opposing normal orientations and the left and right flanks of the human nose also have opposing normal orientations.

Pruning correspondences: 3D surface images typically contain artifacts such as holes and large triangles indicating badly captured or missing parts. Any correspondence to such artifacts is meaningless and are indicated as correspondence outliers, not to be taken into account when updating the transformation model. The MeshMonk toolbox allows for the identification of outliers either deterministically or stochastically, or a combination of both. Deterministic outliers include correspondences to surface border points (this properly handles the situation when the template

and target surface have non-overlapping structures), large triangles (as identified using a z-score on triangle area, and therefore defined in function of the underlying surface resolution), angle between point surface normals (further excluding point correspondences with opposing normal orientations) and any manually tagged points on the target or template. Stochastic outliers are defined following (Claes et al., 2012a), using inlier versus outlier distribution estimations. The inlier distances are assumed to form a Gaussian distribution, and any point falling out of +/- a user-defined κ times the standard deviation is considered abnormal and flagged as an outlier. Then, the contributions of the outlier correspondences are fixed and the confidence values of all the points are updated.

Updating the transformation model: Given updated correspondences, an update of the transformation model parameters is done in each iteration. During the rigid registration, the transformation model is constrained to changing the position (translation), orientation (rotation), and scale of the template only. During the non-rigid registration, a visco-elastic model is enforced, controlling a regularization of the energy function to ensure that points that lie close to each other move coherently. This regularization also includes the outliers, which do not contribute to the transformation model estimation but should be consistently transformed along with the inliers. The smoothness of the transformation model is parametrized by convolving the displacement vectors between corresponding points with a Gaussian (Bro-Nielsen, 1996). The amount of smoothing is high (multiple Gaussian convolution runs) at the beginning iterations, when correspondences are still noisy and hard to define, and reduces gradually towards the later iterations, when correspondences are more accurately defined.

2.2 Parameters and tuning

Given a dataset of 3D images of interest, the entire MeshMonk procedure can be optimized by setting a variety of parameters in the toolbox, and a parameter tuning can be done based on two "quality" measures. First, a quality of "shape fit" is defined as the root mean squared distance of all template points to the target surface after registration. This essentially measures how well the shape of the template was adapted to the target shape and can be measured over multiple images to deduct an overall quality of shape fit from the dataset. Second, a quality of the consistency of point indications across the same dataset is obtained following the principle of minimum description length in shape modelling (Davies et al., 2002). Given two models explaining the same amount of variance, the model requiring fewer parameters is favored, or given two models with the same number of parameters, the one explaining more variance in the data is favored. As an underlying model in shape analysis, a principal component analysis (PCA) is a valid option from which the proportion of variance explained by a given number of principal components (PCs) reflects the model quality. Intuitively, lesser PCs are required to explain the same proportion of variation when the data are well correlated and contains good redundancy. In the opposite scenario, when presented with noisy data, more PCs are required to capture the same amount of variance. Therefore, if the point indications were preformed consistently, a good PCA model results. A parameter tuning was done for the facial data in this work prior to the validation and is described in the supplementary methods.

2.3 Validation

2.3.1 Sample and data curation

Over many years, our collaborative group has recruited study participants through several studies at the Pennsylvania State University and sampled in the following locations: State College, PA (IRB 44929 and 4320); New York, NY (IRB 45727); Urbana-Champaign, IL (IRB 13103); Dublin, Ireland; Rome, Italy; Warsaw, Poland; and Porto, Portugal (IRB 32341). Stereo photogrammetry was used to capture 3D facial surfaces of N~6,000 participants using the 3dMD Face 2-pod and 3-pod systems (3dMD, Atlanta, GA). This well-established method generates a dense 3D point cloud representing the surface geometry of the face from multiple 2D images with overlapping fields of view. During photo capture, participants were asked to adopt a neutral facial expression with their mouth closed and to gaze forward, following standard facial image acquisition protocols (Heike et al., 2010).

2.3.2 Manual placement of validation landmarks

Of the larger sample, N=48 surface images were chosen at random for validation. This number was then reduced by excluding surface images from participants that reported major facial injury or surgery. This resulted in N=41 surface images for validation, which were diverse with respect to sex (N_{Female} = 29, N_{Male} = 12), age (range: 18-79, *M* = 32.78), height (range: 149.86-184.00 cm, *M* = 167.13 cm), weight (range: 43.00-103.80 kg, *M* = 67.62 kg), and 3D camera system used (SI Table 1). Most participants reported being of European descent, with one person reporting to be of admixed African and European descent and another choosing not to report their ancestry. 3dMDpatient was used to record the 3D coordinates of 19 standard landmarks (7 midline and 12 bilateral) from each unaltered surface (i.e. still containing hair and clothing) in wavefront.obj format (Figure 3; SI Table 2). Two independent observers placed landmarks three times each, with at least 24 hours in-between landmarking sessions, resulting in six total landmark indications for each facial image. For each individual, we checked for gross landmark coordinate errors (e.g. mislabeling right and left side landmarks) before analysis. In the subsequent analysis, A_{ML} represents the average manual landmarks from observer A, B_{ML} represents the average manual landmarks from observer B, while the average of all six manual landmark indications (i.e. the combined average) is denoted as C_{ML}.

2.3.3 Automatic placement of validation landmarks

To obtain automatic indications of the 19 validation landmarks, a leave-one-out approach was used to identify the placement of the validation landmark on the template, then indicate these landmarks on the left-out face (Figure 4). Specifically, each of the validation faces was registered using MeshMonk and the manual landmark placements were transferred to the registered face using coordinate conversions (Figure 4A; Hille, 1982). Because the registered faces are now all in the same coordinate system as the original template, we can subsequently transfer the manual landmark indications to the original template (i.e. pre-registration), giving a set of 41 x 2 observers x 3 indications = 246 manual landmark positions on the template scan (Figure 4B). One by one, each face was left out while averaging the other 40 landmark placements to “train” the automatic landmarks (Figure 4C). These averages were then transferred back onto the left-out (target) face resulting in the automatic placement of the validation landmarks using a “training” set that did not include the target face (Figure 4D). Further detail on this process can be found in the Supplemental Methods.

The placement of automatic landmarks was performed three times, changing the manual landmark data used as input: once using the average of observer A's three manual landmark indications (A_{Auto}), again using the average of observer B's three manual landmark indications (B_{Auto}), and a final time using the combined average of all six manual landmark indications from both observers (C_{Auto}). This process resulted in three placements of automatic landmarks for comparison.

2.3.4 Validation

2.3.4.1 Accuracy

We assessed the accuracy of the MeshMonk automatic landmark placements by comparing them to manual landmark placements, using the root mean squared error (RMSE) between the manual and automatic *x*, *y*, and *z* coordinates. We also calculated Bland-Altman (Altman and Bland, 1983) and Intraclass Correlation Coefficient (ICC; Fisher, 1925) statistics to compare the manual and automatic landmark indications. The Bland-Altman method is preferred over correlation or regression as it is less influenced by the variance of the sample and ICC is preferred because it tests both the degree of correlation and agreement between methods.

In addition to absolute locations, geometric morphometric analyses also make use of centroid sizes as a means of understanding large scale differences among groups and to control for size variation. To assess differences manual and automatic methods, we compared estimates of centroid size calculated using each method and performed an analysis of variance (ANOVA) test on the centroid size calculations, with individual, observer, method, and individual x observer as predictors to determine if variation in centroid size could be attributable to variation in landmarking method.

We utilized several methods to determine if the variance structures produced by the two methods were similar. Fitting a multivariate analysis of variance (MANOVA) estimates the variance explained, in correlated outcome variables, by various factors included in the model. Here, we performed MANOVAs separately on the GPA-aligned average manual landmark indications from each observer (A_{ML} and B_{ML}) as well as on the GPA-aligned automatic landmark indications trained using the average of each observer's three landmark placements (A_{Auto} and B_{Auto}), with image and observer as predictors in both tests. By comparing the results of these two tests, we can determine how the explanation of shape variance changes given a different landmarking method. To directly determine if any variance in shape was attributable to landmarking method, we combined the average manual landmark placements of each observer with the automatic placements trained using each of these averages and aligned them using GPA (A_{ML}, B_{ML}, A_{Auto}, and B_{Auto}). We then tested the shape variation in this combined space as the response in a MANOVA, with individual, observer, method, and individual x observer as factors.

2.3.4.2 Reliability

In general, there are two sources of variation in landmark placement: variation between indications taken at different times by the same individual (intra-observer error); and the difference between indications made by different individuals (inter-observer error). We calculated the manual landmarking intra-observer error as the standard deviation between the *x*, *y*, and *z* coordinates of each observer's manual landmarking indications. The inter-observer error of the manual landmark indications was calculated as the standard deviation between each observer's average *x*, *y*, and *z* coordinates (A_{ML} vs. B_{ML}). As an additional method to understand the variation present in the manual landmark indications only, we performed a multivariate

analysis of variance (MANOVA) after aligning the six manual landmarking indications using a generalized Procrustes alignment (GPA; Rohlf and Slice, 1990). Study individual, observer, and landmarking iteration were used as factors and landmark configuration as the response.

To determine if the automatic indication process was more or less variable than manual landmarking, we compared the inter-observer error calculated using only the manual landmarks (A_{ML} vs. B_{ML}) to the standard deviation between one observer's manual landmarks and the automatic landmarks trained using the other observer's manual placements (A_{ML} vs. B_{Auto} and A_{Auto} vs. B_{ML}), as if the automatic indications replaced the manual indications in a calculation of inter-observer error. A paired T-test was used to determine whether the "inter-observer errors" calculated using the automatic indications were significantly different than the error calculated using only the manual indications. Standard deviation values calculated using both automatic placements (A_{Auto} vs. B_{Auto}) were compared to manual landmarking inter-observer error to illustrate the variance of automatic landmark indications. Levene's test was performed (Levene, 1960) to determine if the variances of the inter-observer errors calculated using the manual landmarks were equal to the standard deviation between the automatic landmarks (the null hypothesis) or unequal (the alternative hypothesis). Levene's test was chosen because the distribution of standard deviation values was non-normal.

All analyses were performed in R using the Geomorph (Adams and Otárola-Castillo, 2013), BlandAltmanLeh (https://cran.r-project.org/web/packages/BlandAltmanLeh/BlandAltmanLeh.pdf), and ICC (https://cran.r-project.org/web/packages/ICC/ICC.pdf) packages, as well as packages for data manipulation (readxl, reshape2, plyr, car, data.table, dplyr, broom) and graphing (ggplot2, GGally, GGpubr). Centroid sizes were calculated using Geomorph and MANOVAs for shape variation were implemented using the ProcD.lm function from Geomorph (Collyer et al., 2015). The 19 manual and automatic landmark indications as well as the code used to perform this analysis are available in the following GitHub repository: https://github.com/juliedwhite/MeshMonkValidation/.

3 Results

3.1 Accuracy

3.1.1 Direct comparison of manual and automatic landmark placements

As one measure of validation of the automatic landmark indications, we compared the raw coordinate values of the manual landmark indications with the raw coordinate values of the automatic landmark indications while considering the manual landmarks to be the "gold standard". Because of the leave-one-out nature of our approach, we can compare the manual and automatic landmark coordinates directly without fear of training bias. To compare landmark indications, we calculated the root mean squared error between the *x*, *y*, and *z* coordinates for manual and automatic indications (Table 1) and calculated the intraclass correlation coefficient between the *x*, *y*, and *z* coordinates produced by the two methods. When comparing the average of all six manual landmarking indications (C_{ML}) and the automatic landmarks trained using this average (C_{Auto}), the highest difference after averaging standard deviation values across all axes, was 0.85 mm for the right side exocanthion landmark (Table 1). Overall, the average standard deviation between C_{ML} and C_{Auto} across all landmarks was 0.62 mm. Bland-Altman comparisons showed that the 95% confidence intervals for the landmark indication between methods are within 1.5 mm of a mean

difference of 0 mm (Figure 5). Most individuals fall within these confidence limits, with only a few comparisons from each axis having differences greater than 3 mm. The intraclass correlation coefficients for each axis are around 0.99, representing very high correlation and agreement between manual and automatic landmark indications.

3.1.2 Centroid size comparison

We used estimates of centroid size (CS; the square root of the sum of squared distances from each landmark to the geometric center of each landmark configuration) as an additional assessment of the similarity between manual and automatic landmark placements, since centroid sizes feature heavily in geometric morphometric assessments. The ICC of centroid sizes calculated using the manual and automatic landmarks were all high (ICC_A = 0.9589, ICC_B = 0.9486, ICC_C = 0.9591; Figure 6A). ANOVA by individual, observer, and method shows that individual is the only significant factor in explaining variance in centroid size (F = 130.407, $P < 2 \times 10^{-16}$; Table 2). Bland-Altman comparison showed that the 95% confidence intervals for the centroid size estimates between methods are 2 mm relative to an average centroid size of about 165 mm (Figure 6B).

3.1.3 Analysis of shape variance

A MANOVA on shape, based on the average of each observer's manual landmark indications and automatic landmark configurations, separately, was performed to determine if the variance explained by individual and observer factors was similar in both methods (Table 3). In both methods, individual variation contributed to most of the variation in shape (R^2_{ML} = 94%; R^2_{Auto} = 97%). Differences in observer accounted for 1.9% of the variation in shape from manual landmarks and 2.6% of the variation in shape from automatic landmarks. In total, 3.9% of the variation present in manual landmark shape configurations was unexplained by our model while only 0.22% of the variation was unexplained when testing the automatic landmark configurations. A MANOVA on GPA-aligned manual and automatic configurations from each observer, with method, individual, observer, and individual x observer as predictors showed that landmarking method did not significantly account for variation in landmark placement (F = 0.3463; P = 0.987; Table 4)

3.2 Reliability

3.2.1 Intra- and inter-observer error of manual landmarks

The quantitative study of morphology using 3D coordinates requires specific attention to measurement error and has a robust presence in the literature. For each observer, we calculated the intra-observer error of the manual landmarks as the standard deviation between the *x*, *y*, and *z* coordinates of each observer's three landmarking iterations. SI Table 3 reports intra-observer standard deviations for the manual landmark indications along each axis, averaged only across images. The average standard deviation of observer A across all landmarks was 0.58 mm while the average standard deviation of observer B across all landmarks was 0.44 mm. The average inter-observer error, measured as the standard deviation between the average *x*, *y*, and *z* coordinates of each observer's landmarking iterations was 0.40 mm. This range of deviation is considered highly precise and is similar to previously reported measures of landmark error (Aldridge et al., 2005; von Cramon-Taubadel et al., 2007).

The analysis of measurement and observer error for the manual landmarks alone, assessed using a MANOVA for shape, with individual, observer, observer x individual, and nested observer x landmarking iteration as factors showed that non-individual factors contributed significantly to variation in shape (SI Table 4). Individual variation contributed to most of the variation in shape (85%), as expected. Simple measurement error accounted for 3.5% of the total variation in shape. Additional to this, differences in observer accounted for 1.8% of shape variation, and deviation across landmarking iterations contributed an additional 1.5% of the total variation in shape. In total, non-individual effects contributed to 15% of the total shape variation, with 8.3% of this variation unexplained by the model.

3.2.2 Comparison of inter-observer errors

By treating the automatic landmark indications as if they were performed by a third observer, we calculated "inter-observer" errors to compare the variation of automatic and manual landmarking. In this assessment, we compared inter-observer errors calculated using only the manual landmarks (A_{ML} vs. B_{ML}) with error estimates calculated by replacing one of the observer's manual landmark indications with the automatic indications trained using that observer's average. This resulted in two extra estimations of inter-observer error (A_{ML} vs. B_{Auto} and A_{Auto} vs. B_{ML}), calculated as the standard deviation between *x*, *y*, and *z* coordinates (Figure 7). The mean manual landmarking inter-observer error was 0.40 mm while both manual-automatic comparisons had mean standard deviation values of 0.53 mm (Table 5). A paired t-test between the manual landmark error values and each of the manual-automatic comparison showed that the standard deviation values for both manual-automatic comparisons were significantly different than their manual comparison counterparts at the chelion right, crista philtri left, endocanthion right, both exocanthi, glabella, and labiale superius landmarks. The standard deviation values calculated after replacing the B_{ML} landmarks with B_{Auto} landmarks were significantly different from the A_{ML} vs. B_{ML} comparison at the endocanthion left, labiale inferius, and pronasale landmarks. The standard deviation values for alar curvature left, chelion left, and subalare left landmarks were significantly different when comparing standard deviation values for A_{Auto} to B_{ML} with those of the manual indications. Overall, ten of the nineteen landmarks showed significant differences when comparing the manual landmark inter-observer error with standard deviation values of A_{ML} vs. B_{Auto}. Eleven of the nineteen landmarks showed significant differences when comparing manual landmark inter-observer error to the A_{Auto} vs. B_{ML} standard deviation values. The landmark indications that were significantly different between the two methods tended to be those where facial texture likely assisted in the placement of the manual landmarks (e.g. localizing the crista philtra by looking at the differences in color between the lips and the skin). This result indicates that automatic sparse landmarking using MeshMonk will likely produce more robust results when given input data that has a strong anatomical orientation (e.g. the nasion and pogonion). Even given these differences in variance, the manual-automatic comparisons did not produce errors that were completely outside the range of inter-observer errors, a sign of the reliability of the MeshMonk registration.

As an illustration of the low errors between automatic landmark indications trained using different observers, we calculated the standard deviation between automatic landmark indications trained using the average of observer A's three landmark indications and the average of observer B's three landmark indications (A_{Auto} vs. B_{Auto}; Table 6, Figure 7). The variance of the average standard deviation values were significantly different for all landmarks except labiale superius, where we could not reject the null hypothesis that the variances of the two standard deviation distributions

were equal (F = 2.4213, P = 0.1236). Figure 7 shows that the variance between automatic landmarking indications (A_{Auto} vs. B_{Auto}) is easily identified as being smaller than the manual landmark inter-observer error (A_{ML} vs. B_{ML}).

4 Discussion

Through studies utilizing manually placed sparse landmarks, we have begun to understand the biological basis and evolution of complex phenotypes, both normative and clinical. However, there is still much to be learned. One avenue for improvement is to expand and speed up the production and analysis of data using methods derived from engineering and computer vision, which allow for the description of shapes as "big data" structures instead of sparse sets of landmarks or linear distances, thus matching our ability to describe phenotypes with our ability to describe genomes. To this end, we introduce the MeshMonk registration framework, giving researchers the opportunity to quickly and reliably establish a homologous set of positions across the entire sample. We have validated this framework using a sparse set of landmarks, though the registration framework produces thousands of landmarks (7,160 for the face) to finely characterize the structure.

With respect to the overarching theme of this journal issue, MeshMonk represents a step forward in our ability to describe complex structures, like the human face, for clinical and non-clinical purposes. Consider Figure 8, showing the starting template for facial image registration (left) as well as three example faces (right). Each point on the images represents a quasi-landmark datapoint that is homologous and can be compared across faces. Researchers are no longer limited to a few homologous points, chosen because they can be reliably indicated over hundreds of hours of work. Instead, minute details of the face can be identified and compared across thousands of images in a few hours, and additional images can be incorporated just as easily, regardless of the camera system with which they were captured, allowing for the incorporation of images from different sources and databases (e.g. Facebase.org).

Because of the relative newness of dense correspondence phenotyping, few studies have focused on the accuracy and reliability of the resulting registrations. Previous studies using versions of the MeshMonk framework have shown that the error associated with the registration of the template onto facial images is 0.2 mm (Claes et al., 2012c) and parameters of the toolbox have been fine-tuned, as discussed elsewhere (Snyders et al., 2014) and in the supplemental methods. To provide some validation regarding the ability of the registration process to accurately identify anatomical positions of interest, we used a set of 40 faces with manual landmark indications to "train" positions of interest on the template, then automatically indicate these positions on a face that was not present in the training dataset. In the comparison of manual and automatic landmark indications, the positions of the manual landmarks were considered to be the gold standard, as they have a long history of use and validation in morphological studies (Aldridge et al., 2005; Weinberg et al., 2004). By limiting ourselves to a set number of sparse landmarks, we cannot necessarily speak to the accuracy of structures not involved in our validation (i.e. the cheek bones), but we argue that the results for our comparison speak highly of the fidelity with which the MeshMonk registration framework aligns to underlying anatomical structures.

In the direct comparison of sparse landmarks placed manually and using the MeshMonk toolbox, the average difference between the manual and automatic placements was low (Figure 5), with the

average root mean squared error across all landmarks from 0.62 to 0.68 mm (Table 1), which is well within the range of acceptable error for manual landmarks (Aldridge et al., 2005; von Cramon-Taubadel et al., 2007; Weinberg et al., 2004) and similar or below errors reported in other comparisons of manual and automatic landmarking methods (De Jong et al., 2016, 2018; Li et al., 2017; Subburaj et al., 2009). When assessing landmarking methods separately, the variance in landmark configuration attributable to individual and observer factors is similar, with considerably less variation left unexplained by a MANOVA model using automatic landmark configuration as the response (Table 3). When assessing manual and automatic landmark configurations in a single MANOVA, the landmarking method is a nonsignificant factor, indicating that variation in scans is not attributable to variation in landmarking method (Table 4). This result was also reproduced when comparing centroid sizes and variance-covariance matrices calculated using manual and automatically placed landmarks (Table 2), speaking highly to the high correspondence between landmark indications placed by human observers and those indicated by the MeshMonk toolbox.

The validation results together suggest that the MeshMonk toolbox is able to reliably reproduce information given by manual landmarking. Though the larger contribution of the MeshMonk toolbox is the ability to quickly and densely characterize entire 3D surfaces, our illustration using a small number of manually placed landmarks as a training set could be useful for studies seeking specifically to study a sparse set of landmarks, perhaps to add more images to a dataset that is already manually landmarked or to add additional landmarks to an analysis. Utilization of the MeshMonk toolbox also gives the opportunity to minimize variation due to different observers. Take, for example, datasets with manual landmarks indicated by two different observers. During the course of analysis, the inter-observer error of these observers would have to be calculated and taken into account when interpreting results. From our own study, the inter-observer error of the manual landmarks placed by two different observers was 0.40 mm (SI Table 3). With the automatic landmarking framework implemented during this study, we can minimize both intra-observer variance for a single scan (by averaging together all indications for that scan by a single observer) and intra-observer variance across scans by placing all indications from the training dataset on the template mesh and averaging the entire training set before carrying along these averages during the registration process to place them in an automatic fashion on the target image. This process finely tunes the position of the landmark, such that even if the training sets were indicated by two different observers, the variation in landmark indication is much smaller than the variation in manual landmark indication, averaging 0.2711 mm in our study (Table 6; Figure 7).

A visual hallmark of the ability of spatially dense surface registration to reliably represent anatomical structures is found in the crispness of “average shapes,” constructed by averaging together all registered surfaces in a study sample. Because the MeshMonk registration aligns closely with the underlying anatomical structure, averages across the study samples continue to cleanly resemble the structure and detail is not lost in the averaging process. As depicted in Figure 9, consider the sample average of the 41 faces in this work and 100 mandible scans. The sample averages on the left were registered using only rigid registration, then the template points were simply mapped exactly to their closest points on the target, giving a recognizable but rough matching of template and target. The sample averages on the right were registered using rigid plus non-rigid registration, gradually warping the template to closely fit the surface of the target. In the rigid-only average, fine details of the are overly smoothed compared to the level of detail present in the rigid plus non-rigid registration averages. For example, it is obvious to the naked eye that the sharpness of the eyes, nose, philtrum and mouth for the facial average, and the alveolar crest,

mental foramen, and coronoid and condylar processes for the mandible, are clearly better represented with the rigid plus non-rigid registration. Thus, non-expert readers can easily evaluate the quality of dense-correspondence morphometrics research by looking at the average surface shapes used, which are typically used in manuscript figures, with the understanding that high quality registration leads to sharp average scans where anatomical positions of interest are clearly defined.

Within a dense-correspondence framework like that supplied by MeshMonk, researchers can develop algorithms to recognize fine structures indicative of a specific dysmorphology, aiding clinicians in diagnoses which are typically reliant upon the experience of the examiner (Hopman et al., 2016; Ibrahim et al., 2016; Klingenberg et al., 2010; Suttie et al., 2013, 2017). Our own recent work is an example of the potential of the MeshMonk toolbox to contribute to our understanding of the underlying genetic contributions to normal-range 3D facial variation (Claes et al., 2018). With the increase in resolution offered by MeshMonk, we were able to utilize two different datasets and identify more loci than had previously been reported in a single GWAS of facial variation, even those with a larger sample size (Liu et al., 2012; Paternoster et al., 2012; Shaffer et al., 2016). Other works in this issue also using MeshMonk highlight our ability to finely localize facial variation and genetic effects associated with a common dysmorphology (Indencleef *et al.*, 2018, this issue) and to push forward our understanding of the heritability of the face in a family-based study (Hoskens *et al.*, 2018, this issue).

5 Conclusion

In this study, we present MeshMonk, an open-source resource for intensive 3D phenotyping on a large scale. Compared to a sparse set of manual landmarks, MeshMonk is able to accurately place the same set of landmarks with an average indication error of 0.62 to 0.68 mm. Through dense-correspondence registration algorithms, like MeshMonk, we can advance our ability to integrate genomic and phenomic data to explore variation in complex morphological traits and answer evolutionary and clinical questions about normal-range variation, growth and development, dysmorphology, and taxonomic classification.

6 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7 Author Contributions

JW performed all landmark based analyses and landmarked the 3D scans used for validation with AZ. PC and AO performed the parameter tuning on the facial data and provided the automatic landmark indications. JW, AO, and HM wrote the first draft of the manuscript under supervision of PC. HM, YF, and TP provided input and images using mandible scans. PC and JW conceptualized the design of the study. OE, SV, and MS provided input throughout the analyses and writing process. JS developed the MeshMonk code.

8 Funding

The sample collection and personnel involved in this work was supported by grants from the US National Institute of Justice (2008-DN-BX-K125), the US Department of Defense, the University of Illinois Interdisciplinary Innovation Initiative Research Grant, the Science Foundation of Ireland Walton Fellowship (04.W4/B643), the Penn State Center for Human Evolution and Development, the United States National Institutes of Health (1-RO1-DE027023), the Research Fund KU Leuven (BOF-C1, C14/15/081) and the Research Program of the Fund for Scientific Research - Flanders (Belgium) (FWO, G078518N).

9 Acknowledgments

MeshMonk was developed with WebMonks (https://webmonks.vision), a Belgian startup that works as a bridge to highly qualified developers in third-world countries, and we are very grateful for their high-quality implementation and support. We also thank the many participants who have volunteered their time and all the past and present members of the Shriver and Claes labs, without whom we would have never been able to develop this toolbox or perform the research that it has contributed to.

10 Ethics statement

Institutional review board (IRB) approval was obtained at all locations and all participants signed a written consent form before participation. The Pennsylvania State University IRB board approved the collection of the participants recruited at the following locations: State College, PA (IRB 44929 and 4320); New York, NY (IRB 45727); Urbana-Champaign, IL (IRB 13103); Dublin, Ireland; Rome, Italy; Warsaw, Poland; and Porto, Portugal (IRB 32341).

11 Data Availability Statement

The informed consent with which the data were collected does not allow for dissemination of identifiable data to persons not listed as researchers on the IRB protocol. Thus, the full surface 3D facial images used for validation cannot be made publicly available. In the interest of reproducibility, we have provided the 19 manual and automatic landmarks used for validation as well as the code used to analyze them. These data are available in the following GitHub repository: https://github.com/juliedwhite/MeshMonkValidation/. The MeshMonk code and tutorials are available at https://github.com/TheWebMonks/meshmonk.

12 References

Adams, D. C., and Otárola-Castillo, E. (2013). Geomorph: An r package for the collection and analysis of geometric morphometric shape data. *Methods Ecol. Evol.* 4, 393–399. doi:10.1111/2041-210X.12035.

Aldridge, K., Boyadjiev, S. A., Capone, G. T., DeLeon, V. B., and Richtsmeier, J. T. (2005). Precision and error of three-dimensional phenotypic measures acquired from 3dMD photogrammetric images. *Am. J. Med. Genet.* 138A, 247–253. doi:10.1002/ajmg.a.30959.

Altman, D. G., and Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. *Stat.* 32, 307–317. doi:10.2307/2987937.

572 Andresen, P. R., and Nielsen, M. (2001). Non-rigid registration by geometry-constrained
573 diffusion. *Med. Image Anal.* 5, 81–88. doi:10.1016/S1361-8415(00)00036-0.

574 Baynam, G., Walters, M., Claes, P., Kung, S., LeSouef, P., Dawkins, H., et al. (2015).
575 Phenotyping: Targeting genotype's rich cousin for diagnosis. *J. Paediatr. Child Health* 51,
576 381–386. doi:10.1111/jpc.12705.

577 Besl, P. J., and McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Trans.*
578 *Pattern Anal. Mach. Intell.* 14, 239–256. doi:10.1109/34.121791.

579 Blanz, V., and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. in
580 *Proceedings of the 26th annual conference on computer graphics and interactive*
581 *techniques* (New York, NY, USA: ACM Press/Addison-Wesley Publishing Co), 187–194.
582 doi:10.1145/311535.311556.

583 Bro-Nielsen, M. (1996). Medical image registration and surgery simulation. doi:0909-3192.

584 Chui, H., and Rangarajan, A. (2003). A new point matching algorithm for non-rigid registration.
585 *Comput. Vis. Image Underst.* 89, 114–141. doi:10.1016/S1077-3142(03)00009-2.

586 Claes, P. (2007). A robust statistical surface registration framework using implicit function
587 representations: application in craniofacial reconstruction. Available at:
588 http://mic.uzleuven.be/download/public/MIC/publications/2967/PHD_ pclaes.pdf.

589 Claes, P., Daniels, K., Walters, M., Clement, J. G., Vandermeulen, D., and Suetens, P. (2012a).
590 Dysmorphometrics: the modelling of morphological abnormalities. *Theor. Biol. Med.*
591 *Model.* 9, 5. doi:10.1186/1742-4682-9-5.

592 Claes, P., Hill, H., and Shriver, M. (2014). Towards DNA-based facial composites: preliminary
593 results and validation. *Forensic Sci. Int.* 13, 208–216. doi:10.1016/j.fsigen.2014.08.008.

594 Claes, P., Roosenboom, J., White, J. D., Swigut, T., Sero, D., Li, J., et al. (2018). Genome-wide
595 mapping of global-to-local genetic effects on human facial shape. *Nat. Genet.* 50, 414–423.
596 doi:10.1038/s41588-018-0057-4.

597 Claes, P., Walters, M., and Clement, J. (2012b). Improved facial outcome assessment using a 3D
598 anthropometric mask. *Int. J. Oral Maxillofac. Surg.* 41, 324–330.
599 doi:10.1016/j.ijom.2011.10.019.

600 Claes, P., Walters, M., Shriver, M., Puts, D., Gibson, G., Clement, J. G., et al. (2012c). Sexual
601 dimorphism in multiple aspects of 3D facial symmetry and asymmetry defined by spatially
602 dense geometric morphometrics. *J. Anat.* 221, 97–114. doi:10.1111/j.1469-
603 7580.2012.01528.x.

604 Collyer, M. L., Sekora, D. J., and Adams, D. C. (2015). A method for analysis of phenotypic
605 change for phenotypes described by high-dimensional data. *Heredity (Edinb).* 115, 357–
606 365. doi:10.1038/hdy.2014.75.

607 Corner, B. D., Lele, S., and Richtsmeier, J. T. (1992). Measuring Precision of Three-
608 Dimensional Landmark Data. *J. Quantative Anthropol.* 3, 347–359.

609 Davies, R. H., Twining, C. J., Cootes, T. F., Waterton, J. C., and Taylor, C. J. (2002). A
610 minimum description length approach to statistical shape modeling. *IEEE Trans. Med.*
611 *Imaging* 21, 525–537. doi:10.1109/TMI.2002.1009388.

612 De Jong, M. A., Hysi, P., Spector, T., Niessen, W., Koudstaal, M. J., Wolvius, E. B., et al.
613 (2018). Ensemble landmarking of 3D facial surface scans. *Sci. Rep.* 8, 1–11.
614 doi:10.1038/s41598-017-18294-x.

615 De Jong, M. A., Wollstein, A., Ruff, C., Dunaway, D., Hysi, P., Spector, T., et al. (2016). An
616 automatic 3D facial landmarking algorithm using 2D gabor wavelets. *IEEE Trans. Image*
617 *Process.* 25, 580–588. doi:10.1109/TIP.2015.2496183.

618 Fagertun, J., Harder, S., Rosengren, A., Moeller, C., Werge, T., Paulsen, R. R., et al. (2014). 3D
619 facial landmarks: Inter-operator variability of manual annotation. *BMC Med. Imaging* 14,
620 35. doi:10.1186/1471-2342-14-35.

621 Fisher, R. A. (1925). *Statistical methods for research workers*. 5th editio. , eds. F. A. E. Crew
622 and D. W. Cutler Edinburgh: Oliver & Boyd doi:10.1056/NEJMc061160.

623 Hammond, P., Hutton, T. J., Allanson, J. E., Buxton, B. F., Campbell, L. E., Clayton-Smith, J., et
624 al. (2005). Discriminating Power of Localized Three-Dimensional Facial Morphology. *Am.*
625 *J. Hum. Genet.* 77, 999–1010. doi:10.1086/498396.

626 Heike, C. L., Upson, K., Stuhaug, E., and Weinberg, S. M. (2010). 3D digital
627 stereophotogrammetry: a practical guide to facial image acquisition. *Head Face Med.* 6, 18.
628 doi:10.1186/1746-160X-6-18.

629 Hille, E. (1982). *Analytic Function Theory, Volume I*. 2nd editio. Providence, RI: AMS Chelea
630 Publishing Company.

631 Hopman, S. M. J., Merks, J. H. M., Suttie, M., Hennekam, R. C. M., and Hammond, P. (2016).
632 3D morphometry aids facial analysis of individuals with a childhood cancer. *Am. J. Med.*
633 *Genet. Part A* 170, 2905–2915. doi:10.1002/ajmg.a.37850.

634 Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: The next challenge. *Nat. Rev.*
635 *Genet.* 11, 855–866. doi:10.1038/nrg2897.

636 Hutton, T. J., Buxton, B. F., and Hammond, P. (2003a). Automated registration of 3D faces
637 using dense surface models. in *British Machine Vision Conference*, eds. R. Harvey and A.
638 Bangham (Norwich: Citeseer), 1–10. doi:10.5244/C.17.45.

639 Hutton, T. J., Buxton, B. F., Hammond, P., and Potts, H. W. W. (2003b). Estimating Average
640 Growth Trajectories in Shape-Space Using Kernel Smoothing. *IEEE Trans. Med. Imaging*
641 22, 747–753. doi:10.1109/TMI.2003.814784.

642 Ibrahim, A., Suttie, M., Bulstrode, N. W., Britto, J. A., Dunaway, D., Hammond, P., et al.
643 (2016). Combined soft and skeletal tissue modelling of normal and dysmorphic midface
644 postnatal development. *J. Craniomaxillofacial Surg.* 44, 1777–1785.
645 doi:10.1016/j.jcms.2016.08.020.

646 Imaizumi, K., Taniguchi, K., Ogawa, Y., Matsuzaki, K., Nagata, T., Mochimaru, M., et al.
647 (2015). Three-dimensional analyses of aging-induced alterations in facial shape: a
648 longitudinal study of 171 Japanese males. *Int. J. Legal Med.* 129, 385–393.
649 doi:10.1007/s00414-014-1114-x.

650 Klingenberg, C. P., Wetherill, L., Rogers, J., Moore, E., Ward, R., Autti-Rämö, I., et al. (2010).
651 Prenatal alcohol exposure alters the patterns of facial asymmetry. *Alcohol* 44, 649–657.
652 doi:10.1016/j.alcohol.2009.10.016.

653 Levene, H. (1960). "Robust tests for equality of variances," in *Contributions to Probability and*
654 *Statistics: Essays in Honor of Harold Hotelling*, eds. I. Olkin and H. Hotelling (Stanford:
655 Stanford University Press), 278–292.

656 Li, M., Cole, J. B., Manyama, M., Larson, J. R., Liberton, D. K., Riccardi, S. L., et al. (2017).
657 Rapid automated landmarking for morphometric analysis of three-dimensional facial scans.
658 *J. Anat.* 230, 1–12. doi:10.1111/joa.12576.

659 Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., et al. (2012).
660 A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in
661 Europeans. *PLoS Genet.* 8, e1002932. doi:10.1371/journal.pgen.1002932.

662 Matthews, H., Penington, A., Clement, J., Kilpatrick, N., Fan, Y., and Claes, P. (2018).
663 Estimating age and synthesising growth in children and adolescents using 3D facial
664 prototypes. *Forensic Sci. Int.* 286, 61–69. doi:10.1016/j.forsciint.2018.02.024.

665 Paternoster, L., Zhurov, A. I., Toma, A. M., Kemp, J. P., St. Pourcain, B., Timpson, N. J., et al.
666 (2012). Genome-wide association study of three-dimensional facial morphology identifies a
667 variant in PAX3 associated with nasion position. *Am. J. Hum. Genet.* 90, 478–485.
668 doi:10.1016/j.ajhg.2011.12.021.

669 Redert, A., Kaptein, B., Reinders, M., van den Eelaart, I., and Hendriks, E. (1999). Extraction of
670 semantic 3D models of human faces from stereoscopic image sequences. *Acta Stereol.* 18,
671 255–264.

672 Richtsmeier, J. T., Burke Deleon, V., and Lele, S. R. (2002). The promise of geometric
673 morphometrics. *Am. J. Phys. Anthropol.* 119, 63–91. doi:10.1002/ajpa.10174.

674 Richtsmeier, J. T., Paik, C. H., Elfert, P. C., Cole III, T. M., and Dahlman, H. R. (1995).
675 Precision, repeatibility, and validation of the localization of cranial landmarks using
676 computed tomography scans. *Cleft Palate-Craniofacial J.* 32, 217–227. doi:10.1597/1545-
677 1569_1995_032_0217_pravot_2.3.co_2.

678 Rohlf, F. J., and Slice, D. (1990). Extensions of the procrustes method for the optimal

superimposition of landmarks. *Syst. Zool.* 39, 40–59. doi:10.2307/2992207.

Shaffer, J. R., Orlova, E., Lee, M. K., Leslie, E. J., Raffensperger, Z. D., Heike, C. L., et al. (2016). Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. *PLoS Genet.* 12, 1–21. doi:10.1371/journal.pgen.1006149.

Snyders, J., Claes, P., Vandermeulen, D., and Suetens, P. (2014). Development and comparison of non-rigid surface registraion and extensions, Technical report KUL/ESAT/PSI/1401. Leuven, Belgium.

Subburaj, K., Ravi, B., and Agarwal, M. (2009). Automated identification of anatomical landmarks on 3D bone models reconstructed from CT scan images. *Comput. Med. Imaging Graph.* 33, 359–368. doi:10.1016/j.compmedimag.2009.03.001.

Suttie, M., Foroud, T., Wetherill, L., Jacobson, J., Molteno, C., Meintjes, E., et al. (2013). Facial Dysmorphism Across the Fetal Alcohol Spectrum. *Pediatrics* 131, e779–e788. doi:10.1542/peds.2012-1371.

Suttie, M., Wetherill, L., Jacobson, S. W., Jacobson, J. L., Hoyme, H. E., Sowell, E. R., et al. (2017). Facial curvature detects and explicates ethnic differences in effects of prenatal alcohol exposure. *Alcohol. Clin. Exp. Res.* 41, 1471–1483. doi:10.1111/acer.13429.

Toma, A. M., Zhurov, A. I., Playle, R., Ong, E., and Richmond, S. (2009). Reproducibility of facial soft tissue landmarks on 3D laser-scanned facial images. *Orthod. Craniofacial Res.* 12, 33–42. doi:10.1111/j.1601-6343.2008.01435.x.

von Cramon-Taubadel, N., Frazier, B. C., and Mirazon-Lahr, M. (2007). The problem of assessing landmark error in geometric morphometrics: Theory, methods, and modifications. *Am. J. Phys. Anthropol.* 134, 24–35. doi:10.1002/ajpa.

Walter, T., Shattuck, D. W., Baldock, R., Bastin, M. E., Carpenter, A. E., Duce, S., et al. (2010). Visualization of image data from cells to organisms. *Nat. Methods* 7, S26–S41. doi:10.1038/nmeth.1431.

Wei, R., Claes, P., Walters, M., Wholley, C., and Clement, J. G. (2011). Augmentation of linear facial anthropometrics through modern morphometrics: A facial convexity example. *Aust. Dent. J.* 56, 141–147. doi:10.1111/j.1834-7819.2011.01315.x.

Weinberg, S. M., Scott, N. M., Neiswanger, K., Brandon, C. A., and Marazita, M. L. (2004). Digital three-dimensional photogrammetry: Evaluation of anthropometric precision and accuracy using a Genex 3D camera system. *Cleft Palate-Craniofacial J.* 41, 507–518. doi:10.1597/03-066.1.

Wong, J. Y., Oh, A. K., Ohta, E., Hunt, A. T., Rogers, G. F., Mulliken, J. B., et al. (2008). Validity and reliability of craniofacial anthropometric measurement of 3D digital photogrammetric images. *Cleft Palate-Craniofacial J.* 45, 232–239. doi:10.1597/06-175.

13 Tables

Table 1. Root mean squared error between manual and automatic landmarks. Root mean squared error (mm) between the manual and automatic landmark indications. Values are presented for each axis, averaged across all faces, as well as averaged across the axes (mean).

Landmark	A_{ML} *vs.* A_{Auto}				B_{ML} *vs.* B_{Auto}				C_{ML} *vs.* C_{Auto}			
	X	*Y*	*Z*	*Mean*	*X*	*Y*	*Z*	*Mean*	*X*	*Y*	*Z*	*Mean*
Alar curvature left	0.17	0.54	0.59	**0.44**	0.19	0.65	0.76	**0.53**	0.16	0.52	0.61	**0.43**
Alar curvature right	0.18	0.53	0.67	**0.46**	0.18	0.58	0.61	**0.46**	0.17	0.52	0.57	**0.42**
Chelion left	1.23	0.70	0.64	**0.86**	1.26	0.74	0.66	**0.88**	1.11	0.71	0.61	**0.81**
Chelion right	0.93	0.70	0.53	**0.72**	1.15	0.65	0.62	**0.81**	0.98	0.66	0.55	**0.73**
Crista philtri left	0.69	0.85	0.44	**0.66**	0.89	1.01	0.51	**0.80**	0.75	0.89	0.45	**0.70**
Crista philtri right	0.66	0.95	0.50	**0.70**	1.00	1.13	0.47	**0.87**	0.76	1.00	0.44	**0.73**
Endocanthion left	0.84	0.64	0.53	**0.67**	0.83	0.62	0.42	**0.62**	0.78	0.54	0.40	**0.57**
Endocanthion right	1.05	0.74	0.62	**0.80**	1.09	0.62	0.45	**0.72**	1.04	0.65	0.50	**0.73**
Exocanthion left	0.92	0.78	0.91	**0.87**	0.97	0.75	0.88	**0.87**	0.91	0.74	0.88	**0.84**
Exocanthion right	0.93	0.67	0.93	**0.85**	0.98	0.68	0.97	**0.88**	0.94	0.65	0.95	**0.85**
Glabella	0.52	1.43	0.60	**0.85**	0.55	1.46	0.59	**0.87**	0.48	1.31	0.56	**0.78**
Labiale inferius	0.52	0.75	0.56	**0.61**	0.50	0.71	0.38	**0.53**	0.46	0.72	0.48	**0.55**
Labiale superius	0.57	0.72	0.31	**0.54**	0.59	0.98	0.37	**0.65**	0.59	0.81	0.33	**0.58**
Nasion	0.37	1.10	0.51	**0.66**	0.42	1.04	0.48	**0.65**	0.35	0.97	0.47	**0.60**
Pogonion	0.48	1.08	0.45	**0.67**	0.54	1.12	0.42	**0.69**	0.43	1.00	0.38	**0.60**
Pronasale	0.44	0.71	0.33	**0.49**	0.45	0.57	0.28	**0.44**	0.40	0.56	0.28	**0.41**
Subalare left	0.78	0.47	0.54	**0.60**	0.79	0.44	0.64	**0.62**	0.73	0.43	0.56	**0.57**
Subalare right	0.75	0.46	0.76	**0.66**	0.67	0.50	0.52	**0.56**	0.65	0.43	0.60	**0.56**
Subnasale	0.33	0.46	0.33	**0.37**	0.35	0.68	0.33	**0.46**	0.32	0.48	0.26	**0.35**
Mean	**0.65**	**0.75**	**0.57**	**0.66**	**0.71**	**0.79**	**0.55**	**0.68**	**0.63**	**0.72**	**0.52**	**0.62**

Table 2. ANOVA of centroid sizes. Results from an ANOVA with centroid size as the response variable and individual, observer, method and individual x observer as predictors.

Variable	*Df*	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
Individual	40	7936	198.39	130.407	$<2 \times 10^{-16}$
Observer	2	0	0.23	0.154	0.857
Method	1	0	0	0.002	0.962
Individual x Observer	80	12	0.15	0.101	1.000
Residuals	122	186	1.52		

Table 3. MANOVAs on average manual landmark configurations and automatic landmark configurations, separately. Results of two separate MANOVAs, one using the average manual landmark configurations from each observer as the response, and the other using the automatic landmark configurations as the response. In both cases, individual and observer were included as predictors. The interaction effect between individual and observer was not included because the residual degrees of freedom became zero when it was included.

Variable		*DF*	*SS*	*MS*	R^2	*F*	*Z*	*Pr(>F)*
Individual	*ML*	40	0.3937	0.0098	0.9413	23.987	22.515	0.001
	Auto	40	0.3152	0.0079	0.9714	435.70	27.609	0.001
Observer	*ML*	1	0.0082	0.0081	0.0195	19.853	11.563	0.001
	Auto	1	0.0085	0.0085	0.0264	472.98	8.1969	0.001
Residuals	*ML*	40	0.0164	0.0004	0.0392			
	Auto	40	0.0007	1.81×10^{-5}	0.0022			
Total	*ML*	81	0.4182					
	Auto	81	0.3245					

Table 4. MANOVA on manual and automatic landmarks, together. Results from a single MANOVA using the average manual landmark indications from each observer (A_{ML} and B_{ML}) and the automatic landmark indications using the observer level averages (A_{Auto} and B_{Auto}).

Variable	*DF*	*SS*	*MS*	R^2	*F*	*Z*	*Pr(>F)*
Method	1	0.0003	0.0003	0.0004	0.3463	-2.2135	0.987
Individual	40	0.6522	0.0163	0.8778	20.2019	23.3507	0.001
Observer	1	0.0167	0.0167	0.0224	20.6396	11.4067	0.001
Individual x Observer	40	0.0085	0.0002	0.0114	0.2623	13.7253	0.001
Residuals	81	0.0654	0.0008	0.0880			
Total	163	0.7430					

Table 5. Comparison of inter-observer errors. Standard deviation for only manual landmarks and for manual and automatic comparisons. Based on a paired T-test, comparisons that are significantly different using an alpha of 0.05 are in bold.

Landmark	A_{ML}-B_{ML} *Mean SD (mm)*	$A_{ML} - B_{Auto}$ *Mean SD (mm)*	*T statistic*	*P value*	A_{Auto} - B_{ML} *Mean SD (mm)*	*T statistic*	*P value*
Alar curvature left	0.31	0.31	-0.10	0.9197	0.37	-2.14	**0.0382**
Alar curvature right	0.33	0.37	-0.99	0.3275	0.37	-1.23	0.2266
Chelion left	0.42	0.63	-1.89	0.0665	0.64	-2.40	**0.0212**
Chelion right	0.30	0.51	-2.59	**0.0132**	0.56	-3.78	**5.20×10^{-4}**
Crista philtri left	0.39	0.60	-2.87	**0.0066**	0.59	-3.11	**0.0034**
Crista philtri right	0.47	0.62	-1.89	0.0661	0.67	-3.53	**0.0010**
Endocanthion left	0.44	0.55	-2.50	**0.0167**	0.52	-1.46	0.1527
Endocanthion right	0.36	0.64	-5.49	**2.50×10^{-6}**	0.51	-2.84	**0.0071**
Exocanthion left	0.29	0.62	-6.14	**3.00×10^{-7}**	0.61	-5.80	**8.93×10^{-7}**
Exocanthion right	0.29	0.60	-5.54	**2.09×10^{-6}**	0.62	-5.65	**1.49×10^{-6}**
Glabella	0.45	0.64	-2.63	**0.0121**	0.66	-3.12	**0.0033**
Labiale inferius	0.59	0.69	-2.14	**0.0381**	0.62	-0.54	0.5895
Labiale superius	0.32	0.50	-3.05	**0.0040**	0.49	-3.21	**0.0026**

Nasion	0.49	0.55	-0.93	0.3556	0.57	-1.25	0.2188
Pogonion	0.60	0.59	0.15	0.8835	0.61	-0.22	0.8245
Pronasale	0.33	0.41	-2.16	**0.0366**	0.36	-0.65	0.5169
Subalare left	0.40	0.48	-1.35	0.1842	0.51	-2.24	**0.0308**
Subalare right	0.43	0.52	-1.63	0.1114	0.47	-0.84	0.4083
Subnasale	0.35	0.32	0.69	0.4939	0.36	-0.26	0.7930
Mean	0.40	0.53			0.53		

Table 6. Comparison of error variance. The standard deviation of average landmark configurations for the manual (A_{ML} vs. B_{ML}) and automatic (A_{Auto} vs. B_{Auto}) landmarks, averaged across scans. Levene's test was performed per landmark to assess the difference between error variance.

Landmark	*Manual (mm)*	*Auto (mm)*	*F value*	*P value*
Alar curvature left	0.3067	0.0728	59.6244	**2.83×10^{-11}**
Alar curvature right	0.3287	0.2133	22.2346	**1.01×10^{-5}**
Chelion left	0.4182	0.1998	4.6453	**0.0341**
Chelion right	0.2984	0.0637	24.5101	**4.03×10^{-6}**
Crista philtri left	0.3881	0.3811	29.1832	**6.60×10^{-7}**
Crista philtri right	0.4737	0.4472	18.1685	**5.49×10^{-5}**
Endocanthion left	0.4362	0.3504	14.2000	**3.13×10^{-4}**
Endocanthion right	0.3608	0.2669	28.4103	**8.85×10^{-7}**
Exocanthion left	0.2946	0.0808	47.7334	**1.06×10^{-9}**
Exocanthion right	0.2855	0.0961	28.0100	**1.03×10^{-6}**
Glabella	0.4542	0.2938	41.5866	**7.95×10^{-9}**
Labiale inferius	0.5857	0.5773	26.3847	**1.93×10^{-6}**
Labiale superius	0.3185	0.3289	2.4213	0.1236
Nasion	0.4938	0.3511	87.7550	**1.67×10^{-14}**
Pogonion	0.5987	0.3478	23.9927	**4.95×10^{-6}**
Pronasale	0.3323	0.2376	38.2428	**2.49×10^{-8}**
Subalare left	0.4005	0.3239	16.4805	**1.14×10^{-4}**
Subalare right	0.4283	0.3113	25.6819	**2.54×10^{-6}**
Subnasale	0.3480	0.2072	42.6476	**5.57×10^{-9}**
Mean	0.3974	0.2711		

14 Figures

Figure 1. Schematic of the MeshMonk's surface registration algorithm. MeshMonk uses an initial rigid registration based on the ICP algorithm. This step might require an initial rough alignment to ensure similar orientation, which can be done by placing few landmarks on the target surface. Then, the symmetrical weighted k-neighbor correspondences are found, and

outliers are detected and removed. Finally, the visco-elastic transformation is applied. This is performed in an iterative manner, until either a pre-set number of iterations or a pre-set amount of coverage (e.g. a pre-defined root mean squared distance of all template points to the target surface after the transformation) has been reached. Otherwise, the correspondences are updated and the non-rigid registration starts over.

Figure 2. Depiction of MeshMonk registration process. (A) The target and template are separated and not necessarily aligned in space or scale. **(B)** The template is scaled to fit the target and is matched with the target using a rigid registration step. **(C)** The template is further modified to fit the target using a non-rigid registration step that allows for fine adjustment.

Figure 3. Manual validation landmarks. Seven midline and twelve bilateral landmarks indicated by two observers during validation of the MeshMonk software. Descriptions of the landmarks are present in SI Table 2.

Figure 4. Depiction of automatic landmark indication. (A) Each facial scan was manually landmarked six times, three times each by two observers (red and blue points). **(B)** These iterations were then averaged together and are placed on the template (purple points). **(C)** The average of all but the test face (N=40) placements on the template, serving as the foundation for the automatic landmark placements (magenta points). **(D)** Coordinate conversions, described in more detail in the Supplemental Methods, is used to subsequently transfer the automatic landmark placements from the template to the target (left-out) surface, serving as the automatic landmark indication for the target surface (magenta points). **(E)** The manual landmark indications from two observers (red and blue points) for the shown example face, for comparison to the automatic indication in (D).

Figure 5. Bland-Altman plot for similarity between manual and automatic landmark placements. For *x*, *y*, and *z* coordinates, Bland-Altman plot showing the differences between the manual (C_{ML}) and automatic (C_{Auto}) landmark indications against the averages of the two techniques. Blue lines represent the mean difference value and red lines represent the upper and lower 95% confidence limits. Also given are the intra-class correlation coefficient with ICC 95% confidence interval.

Figure 6. Comparison of centroid sizes. (A) Point plots for comparison of centroid sizes using automatic and manual landmarking methods, separated by observer. (B) Bland-Altman plot showing the differences between centroid sizes produced using the manual and automatic methods against the averages of the two techniques. Blue lines represent the mean difference value and red lines represent the upper and lower 95% confidence limits.

Figure 7. Comparison of inter-observer errors. Standard deviation values calculated using both manual landmarks and after replacing each observer's set iteratively with their automatic landmarks. All but the labiale superius landmark had significantly smaller variances in the automatic landmark indication comparison (A_{Auto} vs. B_{Auto}).

Figure 8. Facial template registration. The template (left), built as the average of more than 8000 admixed facial scans, can easily wrap onto any face (three example faces on the right), accurately representing its particular traits. This allows for the explanation of any face in the template's coordinates, enabling a spatially-dense analysis between any registered surfaces.

Figure 9. Comparison of rigid and non-rigid registration algorithms. Sample averages using the 41 validation faces and 100 mandible scans. Scans were registered using rigid registration only (left) and then simply mapped exactly to their closest point on the target surfaces or mapped using rigid plus non-rigid (visco-elastic) registration (right).