Title
Effects of different segmentation methods on geometric morphometric data collection from primate skulls

Author(s)
Ito, Tsuyoshi

Citation

Issue Date
2019-11

URL
http://hdl.handle.net/2433/259829

Type
Journal Article

Textversion
author

Kyoto University
Effects of different segmentation methods on geometric morphometric data collection from primate skulls

Tsuyoshi Ito

Department of Evolution and Phylogeny, Primate Research Institute, Kyoto University, Inuyama, Aichi 484-8506, Japan

E-mail: ito.tsuyoshi.3a@kyoto-u.ac.jp

Running headline: Effects of segmentation in morphometrics
Abstract

1. Increasing numbers of studies are analysing the shapes of objects using geometric morphometrics with tomographic data, which are often segmented and transformed to three-dimensional (3D) surface models before measurement. The present study aimed to evaluate the effects of different image segmentation methods on geometric morphometric data collection using computed tomography data collected from non-human primate skulls.

2. Three segmentation methods based on a visually-selected threshold, a half-maximum height protocol and a gradient and watershed algorithm were compared. For each method, the efficiency of surface reconstruction, the accuracy of landmark placement and the level of variation in shape and size compared with various levels of biological variation were evaluated.

3. The visual-based method inflated the surface in high-density anatomical regions, whereas the half-maximum height protocol resulted in large numbers of artificial holes and erosion. However, the gradient-based method mitigated these issues and generated the most efficient surface model. The segmentation method used had a much smaller effect on shape and size variation than interspecific and inter-individual differences. However, this effect was statistically significant and not negligible when compared with intra-individual (fluctuating asymmetric) variation.

4. Although the gradient-based method is not widely used in geometric morphometric analyses, it may be one of promising options for reconstructing 3D surfaces. When evaluating small variations, such as fluctuating asymmetry, care should be taken around combining 3D data that were obtained using different segmentation methods.

Keywords

computed tomography; geometric morphometrics; measurement error; repeatability; segmentation
1. Introduction

Geometric morphometrics is a statistical method that uses Cartesian landmark coordinates to analyse the shapes of objects (Adams, Rohlf, & Slice, 2004, 2013; Slice, 2007; Mitteroecker & Gunz, 2009; Zelditch, Swiderski, & Sheets, 2012). This approach has the advantages of typically capturing a large number of shape variables and allowing changes in shape to be visualised (Mitteroecker & Gunz, 2009), which generally enables greater flexibility in data acquisition, more sensitive detection of shape variation and easier visual interpretation of the results compared with traditional morphometrics (Rohlf & Marcus, 1993; Parsons, Robinson, & Hrbek, 2003; Bernal, 2007; Maderbacher et al., 2008; van der Niet, Zollikofer, León, Johnson, & Linder, 2010; Breno, Leirs, & Van Dongen, 2011).

Geometric morphometric data are appropriate for many different types of studies that seek to explain patterns of shape variation and the covariation of shape with other variables, making them useful for answering a wide variety of ecological and evolutionary questions, e.g. modularity, allometry, evolutionary process, ecological, geographical and phylogenetic diversities (Table S1). Consequently, increasing numbers of studies are using geometric morphometrics (Lawing & Polly, 2010), particularly in the fields of evolutionary biology (1,382), zoology (833), anthropology (674), ecology (531), genetics & heredity (379) and the others (3,377) [the numbers in parentheses indicate the number of literatures (1990–2018) retrieved from Web of Science using the search term ‘geometric morphometric$’] (Fig. S1).

Three-dimensional (3D) landmark coordinates for geometric morphometric analyses are often captured using 3D surface models (Mitteroecker & Gunz, 2009; van der Niet et al., 2010; Adams & Otárola-Castillo, 2013). These models make it easier to obtain semi-landmarks on surfaces and thus to capture topography that lacks anatomical landmarks (Gunz, Mitteroecker, & Bookstein, 2005; Gunz & Mitteroecker, 2013) compared with the direct digitisation of an object using a contact-based 3D digitizing device such as MicroScribe (Immersion Corp., San Jose, USA). Furthermore, when sharing data, it is often preferable to deposit the original
3D data, such as surface models, rather than digitised coordinate data because the latter carries the risk of inter-observer error (Shearer et al., 2017). 3D surface models can be obtained with a 3D scanner (Friess, 2012), generated by photogrammetry (Evin et al., 2016; Muñoz-Muñoz, Quinto-Sánchez, & González-José, 2016) or extracted from tomographic data (two-dimensional image stacks) using procedures such as computed tomography (CT), micro-CT, magnetic resonance imaging (MRI), and micro-MRI (Vannier, Conroy, Marsh, & Knapp, 1985; Lorensen & Cline, 1987). However, 3D scanning and photogrammetry can capture only the surface of the object that is visible from the outside, while tomographic data can capture the entire body, including the internal structure (Conroy & Vannier, 1984), which is essential in some situations, such as where the shape of a precious fossil specimen in sediments or a rock matrix needs to be captured non-invasively. Thus, recent advances in information technology and enhancements in computer performance combined with the requirement for non-invasive measurements have led to 3D surface and tomographic data being increasingly recognised as important tools in biology (Cunningham, Rahman, Lautenschlager, Rayfield, & Donoghue, 2014; Baird & Taylor, 2017) (Fig. S1).

When using tomographic data, the surface is usually reconstructed once the region of interest (ROI) has been segmented from the remainder of the image. Several segmentation methods are available, each of which has its own advantages and disadvantages. Manual segmentation is still commonly used but is both labour-intensive and time-consuming and may carry the risk of human error (Stammberger, Eckstein, Michaelis, Englmeier, & Reiser, 1999). Threshold-based segmentation techniques are probably the most widely used, whereby a threshold is often determined by visual judgement (Heuzé et al., 2010; Toro-Ibacache, 2013; Noback & Harvati, 2015; Ito & Nishimura, 2016) or using the half-maximum height (HMH) protocol (Coleman & Colbert, 2007). While the visual-based method is probably the easiest to apply, its accuracy is unclear. By contrast, the HMH protocol calculates the threshold value as the mean of the maximum and minimum
grey values along a row of pixels that spans the boundary transition (Spoor,
Zonneveld, & Macho, 1993), which provides accurate measurements (Coleman &
Colbert, 2007). However, this approach often cannot efficiently segment the entire
ROI if it exhibits a heterogeneous grey-value distribution (Rathnayaka, Sahama,
Schuetz, & Schmutz, 2011), in which case multiple local thresholds can be
calculated for a single ROI (Kubo, Kono, Saso, Mizushima, & Suwa, 2008), but this
is a bit time-consuming. These issues seem to be solved by performing
segmentation based on the grey-value gradient (edge detection) (Scherf & Tilgner,
2009; Rathnayaka et al., 2011). However, this requires some programming skills,
as commonly used graphical user interface software often does not implement
functions that can efficiently perform this. Since different segmentation methods
may provide different results, care should be taken when using the reconstructed
surfaces to obtain measurements.

As with traditional morphometrics, the validity of geometric
morphometrics suffers from the effects of measurement error (Arnqvist,
Mårtensson, & Hungaricae, 1998). Researchers sometimes evaluate the
 repeatability of measurements before undertaking the main analyses (Klingenberg
& McIntyre, 1998; Willmore, Klingenberg, & Hallgrimsson, 2005; White & Searle,
2008; Viscosi & Cardini, 2011). Furthermore, systematic surveys have been
conducted to examine intra-observer, inter-observer and inter-device (MicroScribe
digitiser, 3D scanner, photogrammetry and/or CT) measurement errors and
device-induced random errors (von Cramon-Taubadel, Frazier, & Lahr, 2007;
Fruciano et al., 2017; Robinson & Terhune, 2017; Shearer et al., 2017; Marcy,
Fruciano, Phillips, Mardon, & Weisbecker, 2018). These studies have demonstrated
that although the inter-device error is much smaller than the inter-observer error
(Shearer et al., 2017), it is still significant (Fruciano et al., 2017; Robinson &
Terhune, 2017), with particularly large differences being observed between
MicroScribe and the other devices (Robinson & Terhune, 2017) and low-cost 3D
scanners tending to produce larger amounts of random error than high-resolution
micro-CT systems (Marcy et al., 2018). Thus, it is generally agreed that care should be taken when combining data collected by different devices, particularly when evaluating small amounts of biological variation, such as intra-individual and intraspecific variations.

While great effort has been made to evaluate the effects of measurement error on the collection of geometric morphometric data, little remains known about the effects of different segmentation methods. A small number of papers have reported on the effects of segmentation methods using a subset of samples prior to the main analyses. One significant study is that of Toro-Ibacache (2013), which showed that variation between the HMH- and visual-based segmentation methods was much smaller than intraspecific variation in human skulls. By contrast, Gunz et al (2012) reported that different threshold levels considerably affected surface measurements of the bony labyrinth, while Ranthnayaka et al (2011) detected significant differences in the thickness of the reconstructed long bone surface between HMH-based and grey-value gradient-based (Canny-edge detection) methods, although this was not specifically in relation to geometric morphometric data collection. However, while it is clear that the threshold choice will have a larger effect on the measurements when lower resolution images are used (i.e. where there is a larger voxel size relative to the object size) (Hassan, Souza, Jacobs, de Azambuja Berti, & van der Stelt, 2010; Gunz et al., 2012), it is not yet fully understood how and to what extent different segmentation methods affect measurements at a specific resolution. Therefore, to further expand our knowledge in this area, the effects of various segmentation methods on geometric morphometric data collection need to be systematically evaluated, alongside a comparison with various levels of biological variation.

The aim of the present study was to evaluate the effects of three segmentation methods (visual-, HMH- and grey-value gradient-based methods) on geometric morphometric data collection from non-human primate skulls and to compare inter-method variation in the shape data with interspecific,
inter-individual and intra-individual variations, as well as intra-observer error. The gradient-based method was conducted by implementing the code of the user-friendly cross-language programming interface SimpleITK (Lowekamp, Chen, Ibáñez, & Blezek, 2013; Yaniv, Lowekamp, Johnson, & Beare, 2018) in the open-source software Python programming language (Python Software Foundation, https://www.python.org/) for easy sharing, testing and further improvements.

2. Materials and Methods

2.1. Samples

The dried crania of 19 adult males belonging to six species of the genus *Macaca* were obtained from the Primate Research Institute of Kyoto University (Inuyama, Japan) (Table 1). The specimens were scanned using the Asteion Premium 4 helical CT scanner (Toshiba Medical Systems, Otawara, Japan) with a slice thickness of 0.5 mm and serial CT images were reconstructed from the original volumetric data using a pixel size ranging from 0.232 mm$^2$ to 0.348 mm$^2$ and an interslice interval of 0.2 mm.

2.2. Image segmentation

The images were processed using SimpleITK in Python unless otherwise stated. The CT images were first resampled using linear interpolation to make them cubic with a voxel size of 0.15 mm$^3$. The resampled images were then denoised while preserving the edges using a curvature flow filter with a time step of 0.01 and five iterations.

The resampled and denoised images were segmented using three alternative methods (Fig. 1). The segmentation was not manually edited for any of the methods.

The first method involved visually judging a global threshold for segmentation (Fig. 1c). Using the Amira 6 software (FEI Visualisation Sciences Group, Bordeaux, France), a global threshold was manually selected for each specimen to obtain an optimum value where no scanning artefacts could be seen
and as many bony regions as possible were visible (e.g. Noback & Harvati, 2015). This process was repeated twice for each specimen and the mean value was then calculated and used in the following analyses [note: no specimen exhibited an error of >10 Hounsfield units (HU)].

In the second method, a global threshold was calculated based on the HMH protocol (Coleman & Colbert, 2007) (Fig. 1d). For each specimen, the HMH was calculated from randomly drawn lines in randomly selected coronal slices using the ‘SCIKIT-IMAGE’ package (van der Walt et al., 2014) in PYTHON. The global threshold was then calculated as the mean HMH value of multiple lines. To target only those lines that passed through the bone-to-air transition, any lines for which the minimum HU was outside the range of −2,000 to −950 and the maximum HU was outside the range of −800 to 3,000 were removed from the calculation. While this approach does not necessarily guarantee that lines pass through bone-to-air transition and may have underestimated the threshold compared with manually drawing the lines, it is expected to be more stable where a larger number of lines are used. To evaluate the repeatability of this method, this process was repeated 10 times for each specimen using 10, 100, 1,000 and 10,000 lines and the intraclass correlation coefficient (ICC) was calculated using the ‘ICC’ package (Wolak, Fairbairn, & Paulsen, 2012) in R software (R Developmental Core Team, 2019). In the following analyses, the median value of 10 replicates with 10,000 lines was taken as the global threshold for segmentation.

In the third method, segmentation was performed based on the image gradient with watershed (GWS) algorithm (Withey & Koles, 2008; Aly, Bin Deris, & Zaki, 2011) (Fig. 1e). Sobel filter was first applied to each image as this is considered an efficient gradient detector (Senthilkumaran & Rajesh, 2009). The gradient magnitude was then calculated, which functions as the topographic surface in the watershed algorithm. Seeds for the watershed algorithm were selected based on threshold-based segmentation. Seeds for bone were obtained by selecting the bony region that was within an HU range of −500 to 10,000 and was
connected to a region with an HU > 0 and by eroding the selected region so that it did not exceed the bone–air boundary. Seeds for air were obtained by selecting the region that was within an HU range of $-10,000$ to $-1,000$. The watershed algorithm was then used to segment the bony regions based on the gradient magnitude and seeds. These fully automated procedures were carried out for all specimens using the same parameters.

The accuracy of each segmentation method was evaluated by comparing the results with a local HMH value as a reference. This evaluation was conducted in two anatomical regions: the anterior zygomatic arch, which is usually hard and has a high density and a high-HU value, and the posterior maxilla, which is often thin and has a low density and a low-HU value. In each region, the segmentation boundaries were detected on a line passing through the bone–air transition in a coronal slice using the ‘scikit-image’ package in Python. The locations of the segmentation boundaries of the three alternative models (the visual-, HMH- and GWS-based models) were then compared with that of the reference model (local HMH) along this line (Fig. 1b).

2.3. 3D surface reconstruction and landmark-based analyses

The segmented images were subjected to 3D surface reconstruction using the ‘generate surface’ module of Amira 6 with the options compactify and unconstrained smoothing (smoothing extent of five). The generated surfaces were saved in the ‘ply’ format.

The ply-format data were first duplicated and linkable anonymised so that an observer could take two measurements in a random order without knowing the segmentation method or specimen ID. In total, 40 3D landmarks were obtained from the duplicated and anonymised surface models by a single observer (TI) using the Stratovan Checkpoint software (Stratovan Cor., Sacramento, CA, USA) (Table S2; Fig. S2). If the observer made a different judgement about whether a landmark was missing or not for the two replicates (which occurred in four instances), both were treated as missing. The effect of segmentation method on the placement of
landmarks was examined by calculating the distances between homologous landmarks among the three segmentation models using the average values for the replicates.

Missing landmarks were estimated by mapping the weighted averages from the complete dataset onto the specimen with missing values using the 'MORPHO' package (Schlager, 2017) in R. This was undertaken separately for the three models. A generalised Procrustes analysis (GPA) was then performed to register the landmark configurations using the 'GEOMORPH' package (Adams & Otárola-Castillo, 2013) in R. The resulting GPA-registered shape data were subjected to principal component analysis to visualise shape variations, and a Procrustes analysis of variance (ANOVA) was performed to evaluate the effect of segmentation method compared with individual, species, asymmetry and random effects using the MORPHOJ software (Klingenberg, 2011).

3. Results

3.1. 3D surface reconstruction

The visual-based model was much more efficient in reconstructing the bony regions than the HMH-based model but resulted in slightly larger and more numerous artificial holes (i.e. holes that were generated in the surface models but were not actually present) than the GWS-based model (Fig. 2). The visual-based model agreed relatively well with the reference model at the posterior maxilla, where the bony density was low (Fig. 3). However, the results were often up to 2 pixels thicker than for the reference model in the high-density anterior zygomatic arch.

The HMH-based model resulted in many artificial holes, particularly at the posterior maxilla, occipital and sphenoid bones (Fig. 2). This model showed good agreement with the reference model in the anterior zygomatic arch region but was often thinner than the reference model or did not lead to segmentation at the posterior maxilla (Fig. 3). The use of 10 random lines was found to be insufficient to
efficiently replicate the results for segmenting skulls, likely due to the heterogeneous distribution of HU values (Table 2; Fig. S3). Consequently, when lines are randomly drawn, at least 100 or ideally thousands of lines should be used to select a global threshold.

The GWS-based model resulted in fewer artificial holes than both the visual- and HMH-based models, although the presence of artificial holes could not be completely ruled out (Fig. 2). This model also consistently showed good agreement with the reference model at both the anterior zygomatic arch and the posterior maxilla (Fig. 3).

3.2. Landmarks and geometric morphometrics

There were one or two missing landmarks (RHI and/or HOR) in three specimens for the visual-based model, 1–6 missing landmarks (RHI, PIF, PNS, HOR, ZMI and/or PGP) in 17 specimens for the HMH-based model and one or two missing landmarks (RHI and/or PIF) in two specimens for the GWS-based model. It should be noted that HOR was missing in most (17 out of 19) specimens for the HMH-based model.

The levels of inter-method and intra-observer error were evaluated for each landmark. The inter-method error between the GWS- and visual-based models was generally almost equal to or smaller than that between the GWS- and HMH-based models and between the HMH- and visual-based models (Table S3; Fig. 4). Furthermore, the inter-method error between the GWS- and visual-based models was not significantly larger than the intra-observer error (Table S3). By contrast, the inter-method error between the HMH-based model and each of the other two models exceeded the intra-observer error for some landmarks, including PIF and MM1. The level of intra-observer error was not significantly different among the three methods for any landmark except RHI (ANOVA, $P = 0.0349$) and HOR ($P = 0.0499$) (Fig. S4).

Shape variations were assessed in the context of geometric morphometrics. The first four principal components (PCs) accounted for 73.1% of the total shape
variance. The scatterplot of PC1 (37.7%) and PC2 (20.7%) reflected the level of interspecific variation well (Fig. 5a), while that of PC3 (8.1%) and PC4 (6.7%) explained inter-individual (intraspecific) variation (Fig. 5b). In both plots, the inter-method and intra-observer errors appeared to be much smaller than the interspecific and inter-individual variations. This was supported by the Procrustes ANOVA, which showed that species and individual explained much more of the variance than the method and residuals (Table 3), with the intra-individual variance (fluctuating asymmetry) being nearly four-fold larger than the inter-method variance. However, although the effect of method was small, it was significant for both size and shape.

4. Discussion

4.1. 3D surface reconstruction

Visually judging a global threshold is probably the easiest and potentially most commonly used method for image segmentation during 3D reconstruction (Heuzé et al., 2010; Toro-Ibacache, 2013; Noback & Harvati, 2015; Ito & Nishimura, 2016). With this method, users generally select a global threshold where no scanning artefacts can be seen and bony regions are as visible as possible (e.g. Noback & Harvati, 2015). Artificial holes usually become more of a problem than artefacts when observing scans of dry bones. Therefore, by its very nature, the visual-based method is predicted to provide a lower threshold and therefore a more inflated surface than is actually present on average. Supporting this, the present study demonstrated that although the visual-based method accurately segmented the low-density posterior maxilla, it inflated the surface of the high-density anterior zygomatic arch. However, this inflation was only by 1–2 pixels, indicating that, if using a high-resolution image, the visual-based method will give an accurate 3D reconstruction to a certain extent.

The HMH-based method is also widely used in geometric morphometrics due to its accuracy (Gröning, Fagan, & O’Higgins, 2011, 2012; Coquerelle et al.,
In the original paper describing this approach, the means of 10 HMH values along various randomly selected slices were used to obtain the global threshold (Coleman & Colbert, 2007). Many studies have used this method or some modified version of it, such as using a histogram of the grey values of a respective volume rather than crossing lines to calculate the HMH (Senck, Bookstein, Benazzi, Kastner, & Weber, 2015). However, although HMH is probably one of the most accurate and best approaches for reconstructing a 3D surface when the object is simple and has homogeneous grey-value distributions, it is often not suitable for a complex object with heterogeneous grey-value distributions, such as a skull, depending on the resolution of the image. Supporting this, the present study demonstrated that the HMH-based model could not efficiently reconstruct the 3D surface in the low-density posterior maxilla as it produced many artificial holes and erosion. To address this, users must manually edit the segmentation or set multiple local thresholds across an object, which is not only labour-intensive and time-consuming but may also introduce human error or arbitrariness. Another potential problem with the HMH protocol is the arbitrariness of selecting the lines or regions that are used to calculate the HMH, which affects its repeatability. This study suggested that at least 100 and ideally thousands of randomly drawn lines should be used when selecting a global threshold for skulls. Otherwise, lines should be drawn on pre-defined regions like as Evteev, Anikin, & Satanin (2018) instead of random selection. Thus, it might be difficult to apply this method to a complex object such as a skull unless to some extent high-resolution images are available.

The use of a gradient-based method appears to mitigate these issues mentioned above. Rathnayaka et al. (2011) demonstrated that the Cany filter-based method generated 3D models of long bones with a significantly higher accuracy than a visual-based method, and Scherf and Tilgner (2009) showed that the Ray casting algorithm had a higher efficiency and accuracy for the 3D reconstruction of
the trabecular bone than the HMH protocol. Similarly, the present study demonstrated that the watershed algorithm induced less artificial holes, erosion and inflation than both the visual- and HMH-based models in the 3D reconstruction of the cranium. While Rathnayaka et al (2011), Scherf and Tilgner (2009) and the present study all used approaches that were based on image gradients (edges) in the first step of the algorithm, they used different methods of ROI designation once the gradient had been calculated – Rathnayaka et al (2011) extracted ROI outlines by removing the branches of edges, filling gaps and implementing 3 × 3 neighborhood pixel operations, while Scherf and Tilgner (2009) cast rays along the surface-normals, resulting in the area outside the ROI becoming bright and the region inside the ROI becoming dark. While the application of these methods requires some script writing, this brings the added benefit of flexibility and automaticy, and both MATLAB (Mathworks, MA, USA) and Python, which were used in Rathnayaka et al (2011) and the present study, respectively, are user-friendly languages. Thus, the gradient-based method may represent one of promising options for the 3D surface reconstruction of skulls, although it is not yet widely used in the field of geometric morphometrics (but see Le Cabec, Kupczik, Gunz, Braga, & Hublin, 2012; Le Cabec, Gunz, Kupczik, Braga, & Hublin, 2013; Navarro & Maga, 2016; Hublin et al., 2017; Pan et al., 2017).

Segmenting images with lower contrast and/or more noise (e.g. when differentiating between soft tissues or extracting fossils from surrounding sediments/rocks) is more challenging. However, the application of automated operations, such as gradient (edge) detection, morphological operations and watershed algorithms, will probably eliminate or at least reduce the labor associated with manual processing. Other than these, there are also many kinds of segmentation methods (Pham, Xu, & Prince, 2000), some of which can potentially efficiently handle low contrast and noisy images; for example, when segmenting the same region across many individuals within a population, the atlas-based algorithm, which is based on volume registration, is likely useful (Cabezas, Oliver, Lladó,
Freixenet, & Cuadra, 2011). Segmentation methods should be considered depending on the image properties, the number of specimens and the required accuracy for downstream analyses.

4.2. Landmarks and geometric morphometrics

Insufficient segmentation induces artificial holes and erosion and thus increases the number of missing landmarks. The present study showed that the HMH-based method could not efficiently reconstruct low-density regions, such as those around HOR, so a global threshold based on this method is not necessarily recommended for the 3D reconstruction of skulls unless high-resolution images are available.

Analysis of the amount of inter-method error for each of the 40 landmarks showed that the positioning of the landmarks was most similar between the GWS- and visual-based methods, suggesting that the HMH-based method leads to erroneous positioning. In particular, the positions of PIF and MM1 were very different between the HMH-based model and the other two models, and these differences significantly exceeded the intra-observer error. In addition, the inter-method error between the HMH- and GWS-based methods also significantly exceeded the intra-observer error for PRS and RHI. This indicates that landmarks that are located on the tip (e.g. PIF, PRS, RHI) or in low-density regions (e.g. MM1) are more susceptible to improper thresholding.

Significant differences in the level of intra-observer error among methods were only observed for RHI and HOR. It is notable that the intra-observer error for RHI was high with the GWS-based method, although the reason for this remains unclear. Since the sample size of HOR was small for the HMH-based method, the result for this landmark should be interpreted with caution. For most other landmarks, the method that was used did not affect the likelihood of random error occurring, which suggests that repeatability is good assuming only one segmentation method is used.

Shape was assessed at various levels of biological variation. The image
segmentation method that was used had a much smaller effect on the variation in
shape than the species or individual, supporting the findings of a previous
evaluation of human skull variations (Toro-Ibacache, 2013). However, method still
had a significant effect. Furthermore, the effect of method cannot be considered
negligible when compared with the level of intra-individual variation (fluctuating
asymmetry). Therefore, when evaluating small biological variations, such as
fluctuating asymmetry, care should be taken if combining multiple sources of 3D
data that were obtained using different segmentation methods.

There is growing recognition of the importance of open data and data
sharing in terms of the verifiability of results, the reproducibility of studies and the
reuse of data (Rowe & Frank, 2011; Wicherts & Bakker, 2012; Davies et al., 2017;
Lowndes et al., 2017; Culina et al., 2018). Data sharing is particularly important for
morphological studies, as it not only saves financial and labour resources but also
helps protect specimens from damage by reducing the amount of repeated handling
and measurement by different researchers (Davies et al., 2017). However, it is also
ture that data sharing has been considered to require some cautions, because
measurement error is inevitable (Fruciano et al., 2017; Robinson & Terhune, 2017;
Shearer et al., 2017; this study). Researchers willing to use shared 3D surface
models which were reconstructed from tomographic data should consider the
properties of original tomographic images, segmentation methods, and their
potential impact on inference. The findings of the present study will form the basis
for such consideration. Ideally, original image stacks (with sufficient header
information) rather than (or in addition to) 3D models or landmarks are encouraged
to be shared (see also Davies et al., 2017).

4.3. Limitations and future directions

The present study had several limitations and leaves some challenges for
future studies. The effects of different segmentation methods on geometric
morphometric data collection were evaluated using images obtained from a single
helical CT scanner with similar parameter settings. However, the efficiency of 3D
Surface reconstruction is greatly affected by the resolution and quality of the images (Hassan et al., 2010), so higher resolution images are likely to generate a better reconstruction regardless of which segmentation method is used (Palacio-Mancheno, Larriera, Doty, Cardoso, & Fritton, 2014), while extremely low resolution images (e.g. medical CT images of the mammalian bony labyrinth) will be more susceptible to threshold selection (Gunz et al., 2012). At such an extremely low resolution, the difference between the HMH- and visual-based methods may not be negligible even at the inter-individual level, and it is also unclear how the gradient-based method would perform. Image filtering will also affect the segmentation, with too little smoothing likely retaining noise and thus causing artefacts and too much smoothing readily producing artificial holes. Therefore, these points should be kept in mind when applying the findings of the present study to different kinds of images. GWS-based method that was implemented in this study could not fully remove the artificial holes. This may have been partly due to the seeds not being correctly or sufficiently assigned or the gradient not being efficiently detected in low-intensity regions. It is expected that future research will further improve the segmentation method and evaluate the validity of combining images of various resolutions, which are processed with various filters.

Data accessibility

All of the CT data in dicom format are available at the Digital Morphology Museum, Kyoto University (dmm.pri.kyoto-u.ac.jp/dmm/WebGallery/): The PRCT Nos. are 325, 1257, 357, 322, 900, 1243, 920, 342, 989, 1001, 326, 1010, 1007, 387, 1005, 1245, 903, 323 and 1239. The other datasets and code used in this study are available at Dryad Digital Repository (https://doi.org/10.5061/dryad.qk58nt2).

Competing interests

We declare we have no competing interests.
Funding

This work was partly funded by the Keihanshin Consortium for Fostering the Next Generation of Global Leaders in Research (K-CONNEX), established by Human Resource Development Program for Science and Technology, MEXT and the JSPS Grants-in-Aid for Scientific Research (Grant 16H04848, 17K15195, 19H01002 and 19K16211).

Acknowledgements

I thank the study material committee of PRI and all the persons who were involved in collecting and preserving the skeletal remains of macaques. I am grateful to editors, an anonymous reviewer and Allison Bronson for their constructive comments on this manuscript, the laboratory members of evolutionary morphology section of PRI for their fruitful comments on this study and Enago (www.enago.jp) for English language review.

References

morphometrics for studying skull morphology during growth in Mastomys
doi:10.1644/10-mamm-a-331.1

atlas-based segmentation for magnetic resonance brain images. *Computer
Methods and Programs in Biomedicine, 104*(3), e158–e177.

for taking measurements on three-dimensional models. *American Journal of
Physical Anthropology, 133*(1), 723–725.

imaging of matrix-filled fossil skulls by high-resolution computed tomography.

Culina, A., Baglioni, M., Crowther, T. W., Visser, M. E., Woutersen-Windhouwer, S.,
& Manghi, P. (2018). Navigating the unfolding open data landscape in ecology
doi:10.1038/s41559-017-0458-2

Cunningham, J. A., Rahman, I. A., Lautenschlager, S., Rayfield, E. J., & Donoghue,
Evolution, 29*(6), 347–357. doi:10.1016/j.tree.2014.04.004

Davies, T. G., Rahman, I. A., Lautenschlager, S., Cunningham, J. A., Asher, R. J.,
*Proceedings of the Royal Society B: Biological Sciences, 284*(1852), 20170194.

Dobney, K. (2016). The use of close-range photogrammetry in zooarchaeology:
Creating accurate 3D models of wolf crania to study dog domestication.
*Journal of Archaeological Science: Reports, 9*, 87–93.
doi:10.1016/J.JASREP.2016.06.028

from newborn to 5 years old based on computed tomography. *American Journal
of Human Biology, 30*(4), e23132.

doi:10.4436/jass.90004

(2017). Sharing is caring? Measurement error and the issues arising from
combining 3D morphometric datasets. *Ecology and Evolution, 7*(17), 7034–
7046. doi:10.1002/ece3.3256

ligament on mandibular stiffness: a study combining finite element analysis
doi:10.1016/j.jbiomech.2011.01.008

under masticatory loads: Which input variables are important? *Anatomical


R Developmental Core Team. (2019). R: A language and environment for statistical


Pan, L., Thackeray, J. F., Dumoncel, J., Zanolli, C., Oettlé, A., de Beer, F., … Braga,


doi:10.1016/j.jhevol.2014.11.005

Palacio-Mancheno, P. E., Larriera, A. I., Doty, S. B., Cardoso, L., & Fritton, S. P.

(2014). 3D assessment of cortical bone porosity and tissue mineral density

using high-resolution µCT: Effects of resolution and threshold method. Journal


Pan, L., Thackeray, J. F., Dumoncel, J., Zanolli, C., Oettlé, A., de Beer, F., … Braga,

J. (2017). Intra-individual metameric variation expressed at the

eyamel-dentine junction of lower post-canine dentition of South African fossil

hominins and modern humans. American Journal of Physical Anthropology,

162(4), 806–815. doi:10.1002/ajpa.23240


empirical comparison of traditional truss-based morphometric methods with a

newer geometric method applied to New World cichlids. Environmental Biology

of Fishes, 67(4), 417–431. doi:10.1023/a:1025895317253


segmentation. Annual Review of Biomedical Engineering, 2(1), 315–337.

doi:10.1146/annurev.bioeng.2.1.315

R Developmental Core Team. (2019). R: A language and environment for statistical


image segmentation methods on the accuracy of long bone 3D reconstructions.

Medical Engineering & Physics, 33(2), 226–233.


etiology for cribra orbitalia and porotic hyperostosis. American Journal of

Physical Anthropology, 164(1), 76–96. doi:10.1002/ajpa.23258


collection: Combining data from multiple sources. American Journal of

Physical Anthropology, 164(1), 62–75. doi:10.1002/ajpa.23257


https://doi.org/10.1016/0169-5347(93)90024-j


**Figure legends**

Figure 1. Schematic diagram of the three segmentation methods that were compared. (a) Distribution of the Hounsfield unit (HU) values in two bone regions with high and low values, respectively. (b) Segmentation based on local half-maximum height (HMH) (reference model). (c) Segmentation based on the visual-based method. That the high-HU region is erroneously inflated compared with the reference model, when the low-HU region is correctly segmented. (d) Segmentation based on the HMH-based method. If a global threshold is calculated to be higher than the local HMH of low-HU region, the low-HU region could not be segmented or is erroneously eroded. (e) Segmentation based on the image gradient with watershed (GWS)-based method. With this approach, watersheds are searched for based on the gradient magnitude by placing seeds within the catchment based on different thresholds. Although some regions lacked seeds, this is often less of an issue.
because there are usually seeds within the same catchment. The boundary of segmentation is often close to local HMH in both the high- and low-HU regions.

Figure 2. (a–c) Three-dimensional surface reconstructions obtained using the visual-based method (grey) (a), the half-maximum height (HMH)-based method (orange) (b) and the image gradient with watershed (GWS)-based method (sky blue) (c). (d, e) Superimposition of the visual- and GWS-based models (d) and the GWS- and HMH-based models (e). (f) Pictures of the original cranium showing the left-lateral (upper) and occlusal (bottom) views. The specimen is PRICT-1257 (PRISK-5866).

Figure 3. The segmentation of the visual-based method (grey), the half-maximum height (HMH)-based method (orange) and the image gradient with watershed (GWS)-based methods (sky blue) in comparison to a reference model (a local HMH). Positive values indicate that the model was too thick, while negative values indicate too thin, compared to the reference (local HMH). Zero indicates that it is identical to the reference model.

Figure 4. Inter-method error: g_v, between the image gradient with watershed (GWS)-based method and the visual-based method; g_h, between the GWS- and the half-maximum height (HMH)-based methods; h_v, between the HMH- and the visual-based methods. Herein, the mean values of replicates are used. Horizontal dashed lines denote the median of intra-observer errors. Outliers are not shown. For bilateral landmarks, the left- and right-side data are pooled. Asterisk indicates whether inter-method error is significantly larger/smaller than intra-observer error (*, P = 0.05; **, P = 0.01).

Figure 5. The scatterplots of PC scores: (a) PC1 vs. PC2, (b) PC3 vs. PC4. Colour denotes segmentation methods; symbol denotes species. The data of same individual are circled by 95% confidence ellipse.
### Tables

Table 1. Samples used in this study.

<table>
<thead>
<tr>
<th>PRICT No&lt;sup&gt;a&lt;/sup&gt;</th>
<th>PRISK ID&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>322</td>
<td>528</td>
<td><em>M. cyclopis</em></td>
</tr>
<tr>
<td>323</td>
<td>1358</td>
<td><em>M. cyclopis</em></td>
</tr>
<tr>
<td>342</td>
<td>3046</td>
<td><em>M. fascicularis</em></td>
</tr>
<tr>
<td>357</td>
<td>4477</td>
<td><em>M. fascicularis</em></td>
</tr>
<tr>
<td>989</td>
<td>8644</td>
<td><em>M. fuscata</em></td>
</tr>
<tr>
<td>1001</td>
<td>8658</td>
<td><em>M. fuscata</em></td>
</tr>
<tr>
<td>1005</td>
<td>9332</td>
<td><em>M. fuscata</em></td>
</tr>
<tr>
<td>1007</td>
<td>9340</td>
<td><em>M. fuscata</em></td>
</tr>
<tr>
<td>1010</td>
<td>9361</td>
<td><em>M. fuscata</em></td>
</tr>
<tr>
<td>1239</td>
<td>6162</td>
<td><em>M. fuscata</em></td>
</tr>
<tr>
<td>1243</td>
<td>6470</td>
<td><em>M. fuscata</em></td>
</tr>
<tr>
<td>1245</td>
<td>6474</td>
<td><em>M. fuscata</em></td>
</tr>
<tr>
<td>1257</td>
<td>5866</td>
<td><em>M. fuscata</em></td>
</tr>
<tr>
<td>325</td>
<td>218</td>
<td><em>M. mulatta</em></td>
</tr>
<tr>
<td>326</td>
<td>223</td>
<td><em>M. mulatta</em></td>
</tr>
<tr>
<td>900</td>
<td>1849</td>
<td><em>M. nemestrina</em></td>
</tr>
<tr>
<td>903</td>
<td>2299</td>
<td><em>M. nemestrina</em></td>
</tr>
<tr>
<td>387</td>
<td>3052</td>
<td><em>M. radiata</em></td>
</tr>
<tr>
<td>920</td>
<td>9532</td>
<td><em>M. radiata</em></td>
</tr>
</tbody>
</table>

<sup>a</sup> Digital Morphology Museum, KUPRI
(dmm3.pri.kyoto-u.ac.jp/dmm/WebGallery/)

<sup>b</sup> Skeletal collection, KUPRI
(pri.kyoto-u.ac.jp/databases/matedb/)
Table 2. The repeatability of HMH calculation.

<table>
<thead>
<tr>
<th>Number of lines</th>
<th>ICC</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.075</td>
<td>-0.001</td>
<td>0.234</td>
</tr>
<tr>
<td>100</td>
<td>0.692</td>
<td>0.541</td>
<td>0.837</td>
</tr>
<tr>
<td>1,000</td>
<td>0.946</td>
<td>0.906</td>
<td>0.975</td>
</tr>
<tr>
<td>10,000</td>
<td>0.995</td>
<td>0.991</td>
<td>0.998</td>
</tr>
</tbody>
</table>

ICC: intraclass correlation coefficient.
CI (confidence interval limit) was estimated with the alpha level of 0.05.

Table 3. Procrustes ANOVA for centroid size and shape.

<table>
<thead>
<tr>
<th>Effect</th>
<th>SS</th>
<th>MS</th>
<th>df</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Centroid size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extra 1 (species)</td>
<td>50445</td>
<td>10088.98</td>
<td>5</td>
<td>21.7</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Individual</td>
<td>6048</td>
<td>465.23</td>
<td>13</td>
<td>531.3</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error 1 (method)</td>
<td>33</td>
<td>0.88</td>
<td>38</td>
<td>9.9</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Residual</td>
<td>5</td>
<td>0.09</td>
<td>57</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Shape</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extra 1 (species)</td>
<td>0.34898692</td>
<td>0.0011442</td>
<td>305</td>
<td>4.9</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Individual</td>
<td>0.18569714</td>
<td>0.0002342</td>
<td>793</td>
<td>17.9</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Side (DA)</td>
<td>0.00501611</td>
<td>0.0000965</td>
<td>52</td>
<td>7.4</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Ind * Side (FA)</td>
<td>0.01222904</td>
<td>0.0000131</td>
<td>936</td>
<td>4.1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error 1 (method)</td>
<td>0.01375109</td>
<td>0.0000032</td>
<td>4294</td>
<td>2.1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Residual</td>
<td>0.00993181</td>
<td>0.0000015</td>
<td>6441</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DA: directional asymmetry, FA: fluctuating asymmetry.
Figures

Fig. 1
Fig. 2
Fig. 3

Anatomical region

Segmentation • vis • hmh • gws
The pairs of segmentation methods

Fig. 4
Fig. 5

(a) Segmentation
- vis
- hmh
- gws

Species
- M. cyclops
- M. fascicularis
- M. fuscata
- M. mulatta
- M. nemestrina
- M. radiata

807 Fig. 5
Table S1. Ecological and evolutionary insights from geometric morphometric data.

<table>
<thead>
<tr>
<th>Typical questions</th>
<th>Example references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integration and modularity, response (direction) to selection</td>
<td>(Mitteroecker &amp; Bookstein, 2008; Klingenberg, Debat, &amp; Roff, 2010; Martínez-Abadías et al., 2012; Klingenberg &amp; Marugán-Lobón, 2013)</td>
</tr>
<tr>
<td>developmental (in)stability, canalization</td>
<td>(Klingenberg &amp; McIntyre, 1998; Debat, Alibert, David, Paradis, &amp; Auffray, 2000; Santos, Iriarte, &amp; Céspedes, 2005)</td>
</tr>
<tr>
<td>Ontogenetic and allometric trajectories and their relation to evolution</td>
<td>(de León &amp; Zollikofer, 2001; Mitteroecker, Gunz, &amp; Bookstein, 2005; Mitteroecker &amp; Bookstein, 2009)</td>
</tr>
<tr>
<td>Evolutionary process (e.g., genetic drift vs. natural selection), tempo and mode in evolution</td>
<td>(Perez &amp; Monteiro, 2009; Chira et al., 2018; Ponce de León et al., 2018; Weaver &amp; Gunz, 2018)</td>
</tr>
<tr>
<td>Ecological, geographical, chronological and phylogenetic diversity</td>
<td>(Adams &amp; Rohlf, 2000; Claude, Pritchard, Tong, Paradis, &amp; Auffray, 2004; Cardini, Jansson, &amp; Elton, 2007; Drake &amp; Klingenberg, 2010; Esquerré &amp; Keogh, 2016)</td>
</tr>
<tr>
<td>Phylogenetic inference, phylogenetic signal, taxonomic classification</td>
<td>(Cardini &amp; Elton, 2008; González-José, Escapa, Neves, Cúneo, &amp; Pucciarelli, 2008; Catalano, Goloboff, &amp; Giannini, 2010; Catalano, Ercoli, &amp; Prevosti, 2014; Klingenberg &amp; Gidaszewski, 2010; Goloboff &amp; Catalano, 2011; Détroit et al., 2019)</td>
</tr>
<tr>
<td>Genotype–phenotype mapping</td>
<td>(Liu et al., 2012; Pallares, Harr, Turner, &amp; Tautz, 2014; Mitteroecker, Cheverud, &amp; Pavlicev, 2016; Gunz et al., 2019)</td>
</tr>
</tbody>
</table>
References


<table>
<thead>
<tr>
<th>No</th>
<th>Abbreviation</th>
<th>Definition</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PRS</td>
<td>Prosthion: antero-inferior point on projection of premaxilla between central</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>2</td>
<td>NSP</td>
<td>Nasospinale: inferior-most midline point of piriform aperture.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>3</td>
<td>RHI</td>
<td>Rhinion: most anterior midline point on nasals.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>4</td>
<td>NAS</td>
<td>Nasion: midline point on fronto-nasal suture.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>5</td>
<td>PIF</td>
<td>Posterior-most point of incisive foramen.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>6</td>
<td>PNS</td>
<td>Tip of posterior nasal spine.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>7</td>
<td>HOR</td>
<td>Hormion: midpoint of the posterosuperior border of the vomer.</td>
<td>Cobb and O'Higgins (2007)</td>
</tr>
<tr>
<td>8</td>
<td>BAS</td>
<td>Basion: anterior most point of foramen magnum.</td>
<td>Frost et al. (2003)</td>
</tr>
<tr>
<td>9</td>
<td>OPS</td>
<td>Opisthion: posterior most point of foramen magnum.</td>
<td>Frost et al. (2003)</td>
</tr>
<tr>
<td>10</td>
<td>INI</td>
<td>Inion: most posterior point of cranium, when viewed in the Frankfurt horizontal, be it on sagittal/nuchal crest or not.</td>
<td>Frost et al. (2003)</td>
</tr>
<tr>
<td>11</td>
<td>LIA</td>
<td>Posterior-most point of lateral incisor alveolus.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>12</td>
<td>MM1</td>
<td>Mesial M1: contact points between P4 and M1, projected labially onto alveolar</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>14</td>
<td>NPM</td>
<td>Meeting point of nasal and pre-maxilla on margin of piriform aperture.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>15</td>
<td>DCR</td>
<td>Dacryon: most superior point of the lacrimomaxillary suture (intersection with frontal bone).</td>
<td>Cobb and O'Higgins (2007)</td>
</tr>
<tr>
<td>17</td>
<td>ZMI</td>
<td>Zygo-max inferior: antero-inferior point of zygomaticomaxillary suture.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>18</td>
<td>FRO</td>
<td>Frontomalare orbitale: where frontozygomatic suture crosses inner orbital rim.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>19</td>
<td>FRT</td>
<td>Frontomalare temporale: where frontozygomatic suture crosses lateral edge of zygomatic arch.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>20</td>
<td>ZTS</td>
<td>Zygo-temp superior: superior point of zygomatico-temporal suture on lateral face of zygomatic arch.</td>
<td>Frost et al. (2003)</td>
</tr>
<tr>
<td>21</td>
<td>ZTI</td>
<td>Zygo-temp inferior: inferior point of zygomatico-temporal suture on lateral face of zygomatic arch.</td>
<td>Frost et al. (2003)</td>
</tr>
<tr>
<td>22</td>
<td>ZAP</td>
<td>Meeting point of zygomatic arch and alisphenoid on superior margin of nterveomaxillary fissure.</td>
<td>Cardini et al. (2007)</td>
</tr>
<tr>
<td>23</td>
<td>PMA</td>
<td>Most posterior point of maxillary alveolus on the maxilla-palatine.</td>
<td>Ito et al. (2014)</td>
</tr>
<tr>
<td>24</td>
<td>PGP</td>
<td>Most inferior point on the postglenoid process.</td>
<td>Lockwood et al. (2004)</td>
</tr>
<tr>
<td></td>
<td>visual–HMH</td>
<td>HMH–GWS</td>
<td>GWS–visual</td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>Mean</td>
<td>0.32</td>
<td>0.47</td>
<td>0.37</td>
</tr>
<tr>
<td>Median</td>
<td>0.30</td>
<td>0.29</td>
<td>0.21</td>
</tr>
<tr>
<td>SD</td>
<td>0.18</td>
<td>0.28</td>
<td>0.21</td>
</tr>
<tr>
<td>P</td>
<td>0.15 +</td>
<td>0.55</td>
<td>0.41</td>
</tr>
<tr>
<td>*</td>
<td>0.36</td>
<td>0.45</td>
<td>0.34</td>
</tr>
<tr>
<td>SD</td>
<td>0.14</td>
<td>0.36</td>
<td>0.18</td>
</tr>
<tr>
<td>P</td>
<td>0.01 +</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>*</td>
<td>0.32</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>SD</td>
<td>0.09</td>
<td>0.18</td>
<td>0.12</td>
</tr>
<tr>
<td>P</td>
<td>0.78 -</td>
<td>0.32</td>
<td>0.26</td>
</tr>
<tr>
<td>*</td>
<td>0.26</td>
<td>0.29</td>
<td>0.27</td>
</tr>
<tr>
<td>SD</td>
<td>0.21</td>
<td>0.18</td>
<td>0.15</td>
</tr>
<tr>
<td>P</td>
<td>0.15</td>
<td>0.34</td>
<td>0.27</td>
</tr>
</tbody>
</table>

* Plus/minus indicates that the mean of inter-method error is larger/smaller than the mean of intra-observer error.
Figure S1. Literature published between 1990 and 2018 that is related to geometric morphometrics. Publications were retrieved from Web of Science using the search term ‘geometric morphometric$’.

The proportion of papers in each research field is shown on the left based on the tags provided in Web of Science, while the number of publications per year is shown on the right. The purple bar denotes the number of publications retrieved when ‘computed tomography’ was added to the search term. Note that the Web of Science topic search only scans the title, abstract, author keywords and keywords plus. By contrast, Google Scholar detected approximately 12,900 publications over the same period (or 1,640 when ‘computed tomography’ was added).
Figure S2. Landmarks used in this study. (a) frontal, (b) left-lateral, and (c) occlusal views of cranium. See also Table S2.
Figure S3. The repeatability of the half-maximum height (HMH)-based methods relying on the 10, 100, 1,000, and 10,000 randomly drawn lines. Horizontal axis indicates samples: from the left, PRICT-325, 1257, 357, 322, 900, 1243, 920, 342, 989, 1001, 326, 1010, 1007, 387, 1005, 1245, 903, 323 and 1239.
Figure S4. Intra-observer error: vis, the visual-based model; hmh, the half-maximum height (HMH)-based method; gws, the image gradient with watershed (GWS)-based model. For bilateral landmarks, the left- and right-side data are pooled. Outliers are not shown.