FRONTIERS IN ZOOLOGY

METHODOLOGY Open Access

Semi-automatic landmark point annotation for geometric morphometrics

Paul A Bromiley[1*], Anja C Schunke[2], Hossein Ragheb[1], Neil A Thacker[1] and Diethard Tautz[2]

Abstract

Background: In previous work, the authors described a software package for the digitisation of 3D landmarks for use in geometric morphometrics. In this paper, we describe extensions to this software that allow semi-automatic localisation of 3D landmarks, given a database of manually annotated training images. Multi-stage registration was applied to align image patches from the database to a query image, and the results from multiple database images were combined using an array-based voting scheme. The software automatically highlights points that have been located with low confidence, allowing manual correction.

Results: Evaluation was performed on micro-CT images of rodent skulls for which two independent sets of manual landmark annotations had been performed. This allowed assessment of landmark accuracy in terms of both the distance between manual and automatic annotations, and the repeatability of manual and automatic annotation. Automatic annotation attained accuracies equivalent to those achievable through manual annotation by an expert for 87.5% of the points, with significantly higher repeatability.

Conclusions: Whilst user input was required to produce the training data and in a final error correction stage, the software was capable of reducing the number of manual annotations required in a typical landmark identification process using 3D data by a factor of ten, potentially allowing much larger data sets to be annotated and thus increasing the statistical power of the results from subsequent processing e.g. Procrustes/principal component analysis. The software is freely available, under the GNU General Public Licence, from our web-site (www.tina-vision.net).

Background

Anatomical point landmarks are useful features for a wide range of tasks in medical image analysis and machine vision, and are of particular relevance to morphometrics. Traditional approaches to morphometrics focused on the measurement and analysis of specific lengths, angles, areas etc., and were limited to a relatively small number of such features. Since the pioneering work of Bookstein [1], methods based on the application of statistical shape analysis to large numbers of point landmarks have become increasingly popular. One such approach to landmark-based shape analysis is to perform Procrustes superimposition of a set of annotated specimens, in order to remove non-shape variation (translation, rotation and scale) according to Kendall's definition [2]. A principal component analysis (PCA) can then be performed on the superimposed landmarks in order to identify the main modes of shape variation. Such methods are supported by modern data acquisition methodologies, mainly high resolution CT scans, which provide a multitude of characters, and so potential landmark locations, on outer and inner surfaces. Landmark-based geometric morphometrics can provide a quantitative measurement of the shape of an entire structure or organism. The results can provide a more thorough understanding of forms, e.g. through functional morphology or shape spaces, than could be achieved through traditional morphometrics, and provide a route to phylogeny reconstruction (e.g. [3,4]). In combination with other data, they can also be used to establish correlations between ecological factors and shape (e.g. [5,6]), or to quantify genetic parameters of shape. In all cases, quantitative measurement of multiple shape parameters allows powerful, statistical tests of morphometric hypotheses. However, landmark-based morphometric methods have a significant drawback, in that they require

*Correspondence: paul.bromiley@manchester.ac.uk
[1]Centre for Imaging Sciences, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK
Full list of author information is available at the end of the article

the annotation of large numbers of landmarks across multiple specimens, a task that is both difficult and time consuming when performed manually.

Bookstein [1] divided landmarks into three classes according to their relationship to local features. Type one are anatomical points that are defined locally through the juxtaposition of distinct tissues, for example the intersection of cranial sutures, or of veins in insect wings. Type two are intermediate, for example points of locally maximal curvature. Type three are defined by distant, rather than local, features, for example the centre of a circle that is tangent to a structure at more than one point. In a limited number of cases, for example insect wings (e.g. [7]), annotation can be performed on a 2D image of a 2D structure, such that the entire image can be viewed simultaneously. Manual annotation of type one landmarks is then relatively straightforward, although still time consuming if large numbers of landmarks are involved. However, the majority of anatomical structures are three dimensional, and modern tomographic imaging methodologies can provide 3D data. Landmark annotation then becomes more challenging, for two reasons. First, common display technologies are limited to two dimensions, such that only a sub-set of the 3D image e.g. a 2D slice or a projection such as a surface rendering, can be viewed at any time. The display must therefore be repeatedly manipulated during annotation in order to view the location of each landmark. Second, 2D images of specimens allow for intersections of a structure, e.g. a suture, with the background of the image, while 3D objects have more degrees of freedom in rotation, so the same points would need to be specified e.g. as the anterior-most point of a suture. The result is that the process of manual landmark annotation is more difficult, and so more time consuming, in 3D than in 2D data. This has significant implications for subsequent analysis of the data, since the statistical power of any analysis technique, and so the confidence limits on the conclusions, will be dictated in part by the number of data sets that are used.

In [8] the authors described a software package designed to support the process of manual landmark annotation on 3D medical image volumes, with particular reference to micro-CT images. The software presents both 2D and 3D renderings of the image volume, the latter using a fast volume rendering algorithm [9,10] in order to provide the most informative view of the data possible whilst not imposing requirements for specific graphics hardware, and provides numerous functions specifically designed to accelerate the process of landmark annotation for geometric morphometrics. However, the manual input required to annotate a significant number of landmark points is still considerable. In this work, the authors describe an extension to the software package that supports semi-automatic localisation of morphological landmarks in a query image. As described below, the algorithm described here was specifically designed for use in geometric morphometrics, avoiding techniques that could introduce shape-dependent biases into the results.

The problem addressed here was to find the locations of landmarks in a 3D query image volume given a database of example image volumes containing similar structures in which the required landmarks had been manually annotated i.e. to find the mappings from the landmarks in the database images to the corresponding positions in the query image. The literature includes landmark detection techniques, e.g. [11,12], in which a query image is analysed to locate points of maximal surface curvature, maximal intensity gradient etc., that would constitute potential landmarks. Correspondences between landmarks in different images can then be established either manually or automatically. Such methods typically require a surface segmentation, and multiple rules regarding which points constitute potential landmarks. A potentially simpler alternative when manually annotated training data is available, and the approach adopted here, is to consider the mappings between landmark points as sparse transformations from the coordinate systems of the database images to that of the query image, such that the problem falls into the general domain of registration. Registration, the estimation of a transformation that maps one image (the source) into the coordinate system of another image (the target) is a core problem in machine vision and medical image analysis with a correspondingly extensive literature; general reviews of medical image registration are provided by [13-15] and [16], whilst [17,18] and [19] provide recent reviews of surface registration algorithms, with particular reference to surfaces represented by point clouds or meshes.

Image registration is performed by optimising the parameters of a transformation model using a cost function that quantifies the similarity of the transformed source and target images. This can be viewed as a model fitting process, in which the transformed source image constitutes a model, and the target image the data to which the model is fitted. Transformation models can be divided into two classes. The first are global, such as rigid, similarity or affine transformations, where a single set of parameters specifies the transformation. The second are deformable, where the transformation can be characterised as a vector field that varies across the source image. By Kendall's definition [2], shape variation between the structures in the source and target images cannot be modelled using a global transformation model. By definition, the specimens included in a landmark-based morphometric analysis will have differences in shape, requiring a deformable transformation. Significant effort has been applied to the problem of deformable registration of medical images; [20] provides a recent review.

However, deformable registration is an ill-posed problem [20]. Therefore, such methods frequently use a cost function based on two terms; a data term based on the comparison of image intensities or derived features, and a regularisation term that constrains the deformation using an assumed physical model such as viscous fluid flow, elasticity, diffusion etc. in order to make the problem well-posed.

Methods based on free-form deformations of image patches or sub-regions, which attempt to minimise or eliminate the influence of assumed models by using a piecewise rigid or affine transformation, have also been investigated. Lau et al. [21] described a hierarchical approach, in which overlapping sub-regions on a regular grid were independently registered using cost functions based on mutual information (MI; [22-24]), normalised mutual information (NMI; [25]) and the correlation coefficient (CC; [13]), without a regularisation term. A dense deformation field was then estimated from the sparse field of displacement vectors at the region centres by median filtering and Gaussian interpolation, introducing an assumption of smoothness. Malsch et al. [26] described a method in which only sub-regions with high information content were used. Irregularly spaced sub-regions with high local variance were selected and registered using the CC; again, a dense deformation field was estimated from the sparse field of displacement vectors at the region centres by interpolation with thin-plate splines (TPS; [27]), introducing a smoothness assumption based on minimising the bending energy. Söhn et al. [28] extended this approach by analysing the quality of the optimum alignment found for each sub-region using the second derivative of the cost function. Sub-regions on a regular grid were independently aligned using a NMI cost function, and an elastic relaxation was applied depending on the alignment quality: when the cost function exhibited a clear optimum, no relaxation was applied; a combination of data and elastic terms were applied when the optimum was degenerate; and the relaxation was performed with no data term when the optimum was indistinct or absent. B-spline interpolation was then applied to estimate a dense deformation field. Erdt et al. [29] adopted a similar approach, and provided a mathematical framework for the use of eigenvectors of the Hessian matrix of the cost function for estimation of alignment quality, in terms of a Taylor series expansion of the shape of the cost function at the optimum. This method can also be derived within a statistical framework in terms of the minimum variance bound [30,31], and such error information has also been utilised within regularised registration techniques [32]. Finally, [33] described a patch-based registration method inspired by patch-based, multi-atlas segmentation algorithms. The method assumed that the deformation fields between a set of training images and a template were known; a dictionary of patches from the training images, and their deformations, was then constructed. A query image could then be registered to the template by selecting patches around points of high information (high Canny edge detector [34] responses were used), finding the most similar patches in the dictionary, and constructing a weighted combination of the deformations of those dictionary patches. A dense deformation field was then constructed using TPS interpolation, again introducing a smoothness assumption. The assumption of smoothness was the only model-based constraint on the allowable deformation in these free-form, non-rigid registration methods, and was required only to interpolate a dense deformation field. Therefore, for applications requiring only the identification of landmarks, such methods require no assumed model of the deformation.

Significant effort has also been applied to the problem of automatic landmark point localisation in 2D images within the computer vision field. One popular approach has been the use of statistical shape models, in algorithms such as the Active Shape Model (ASM) [35,36] and related work. In the original work, a set of training images were aligned into a common coordinate system using Procrustes analysis, and the coordinates of the landmarks from each training image, in this reference frame, were concatenated to form a single, high-dimensional vector, such that the complete set of training images defined a point cloud in a high-dimensional space. A principal component analysis (PCA) was applied to extract the major modes of variation of this point cloud. The shape model then consisted of two components; a linear combination of these modes, weighted by a set of shape parameters, and a global rigid transformation that located the model in an image. The model was fitted to a query image by optimising the weights and transformation model parameters in order to maximise the image intensity gradient at the landmark point positions i.e. assuming that landmarks would be located on edges in the image. The Active Appearance Model (AAM) [37] extended the same approach to include image intensities, thus producing a model of both shape and appearance. Later developments, for example the Constrained Local Model (CLM) [38], replaced the global appearance model with a set of texture patches around the landmark points. Fitting consisted of a registration of image patches learned from the training data, or reconstructed from the appearance model, to the query image, with the shape model used as a constraint during optimisation in order to ensure that the relative locations of the patches represented a high-probability shape given the training data. AAMs and related algorithms learn a model directly from training data with no other input. However alternative approaches that incorporate a model of an object as a collection of

interconnected parts, e.g. the Pictorial Structure Model (PSM) [39], have also been developed. Whilst most effort has been focused on the application of these techniques to 2D images of faces or objects in natural scenes, they have also been applied to 3D medical image data for purposes such as segmentation; see [40] for a recent review.

The methods described above all perform a registration by optimising a cost function that measures the similarity between the intensities of two images, or image patches, regularised using a model that describes the probability of a given deformation. They exist on a spectrum of model complexity, from very limited models assuming only that the deformation field is smooth and continuous, through full physical (e.g. elastic) models of the allowable deformations, to AAMs that are bootstrapped from training data. However, a model-based regularisation could not be used in the work described here, for the following reasons. Most importantly, the aim was to produce landmarks for geometric morphometrics, which would be analysed with the standard techniques used in that area of research, including Procrustes analysis followed by PCA and interpolation between landmarks using thin-plate splines. The same techniques are used to build the shape models used in AAMs and related algorithms. Therefore, the subsequent analysis would be capable of regenerating the shape model used in automatic annotation i.e. any mode of shape variation present in the query image, but not included in the annotation model, would not be found in PCA analysis of the results. Since the aim of many experiments in landmark-based geometric morphometrics is to quantify the modes of shape variation, this form of bias would be unacceptable. The training data for an AAM would need to exhibit all possible shape variation within the relevant shapes in order to guarantee that such biases were not present; this would make the training set prohibitively large. Similarly, a smoothness assumption would not allow points of infinite curvature in the deformation field. However, such points could be present at the interface between sliding surfaces e.g. the points of contact between the upper and lower molar rows, which would be relevant landmark locations for many studies comparing morphology with ecology. Furthermore, the bootstrapped models used in AAMs and related algorithms require extensive offline training; more manually annotated images would be used to train the model than would typically be included in landmark-based geometric morphometrics experiments. Since the aim here was to maximally accelerate the landmark annotation process, a method without this requirement was needed.

The work described here adopted the hierarchical, patch-based registration used in methods based on free-form deformations, as described above. Patches of image data around each landmark in each database image were registered, using an affine transformation, to the query image. This avoided the use of any model-based shape constraint. Furthermore, since there was no need to interpolate a dense deformation field, no assumption of smoothness was required. A similar approach proved successful in earlier work on automatic landmark annotation for morphometric analysis of microscope images of fly wings, i.e. 2D images of planar objects with no out-of-plane rotations [41,42]. The software described here required a small database of images similar to the query image, in which the required landmarks had been manually annotated. In the absence of regularisation, a multi-stage registration approach was developed in order to compensate for shape variations between the database and query images, which would necessarily be present. The initial stages operated on the whole image, and so were affected by shape variations. However, the results from the initial stages were used only to initialise later stages operating on texture patches around each landmark. Multiple patch-based stages with reducing patch sizes were used, with the intention that the effect of global shape variation would be reduced as the patches became smaller. Automatic image registration algorithms are typically incapable of dealing with gross misalignments e.g. where the images are rotated by 180° with respect to one another, and so manual intervention was required to provide an initialisation. In the work described here, four non-coplanar landmark points in each volume were used for this purpose (see the Conclusions for further comments), and this stage of the algorithm can be omitted completely if care is taken during sample preparation, such that the specimens are in approximately the same orientation in all image volumes used. The point-based stage of the registration minimised the RMS distances between the registration points. All image-based stages minimised the χ^2 of the scaled intensity gradients in the database and query images; the scaling provided some independence to variations in scanner parameters and average bone density, and was estimated using maximum likelihood after the point-based stage of registration.

Each database image produced one estimate of the location of each landmark in the query image; these were combined to generate the final estimated landmark location. A robust alternative to simple approaches such as averaging was implemented, allowing sub-selection of only the most reliable estimates i.e. those for which the database image provided the best model of the query image. Furthermore, this did not require a shape model and could operate with a small number of examples. Since the introduction of the generalised Hough transform [43], array-based voting schemes have been shown to be effective for locating structures in images. In particular, several

recent papers (e.g. [44,45]) have used Random Forests [46] to locate structures in images, combining the estimates from each tree in a voting array. A similar approach was adopted here; the estimated positions of a given landmark from each database entry cast votes into a 3D array, which was then convolved with a Gaussian kernel to approximate the random error on the estimates. Outliers resulting from failed registrations formed a broad background distribution, whilst points from the signal distribution (i.e. successful registrations) were randomly scattered around the true location of the landmark point in the query image according to the random error on the estimation process, and so contributed to a single, dominant peak in the voting array. The most significant mode in the smoothed array was taken as the estimated location for the landmark. This provided a degree of robustness to outliers; however, situations might still occur in which no element of the database provided a good estimate of the landmark location. Therefore, a robust outlier detection method was applied to the final estimated locations, based on testing for consistency between the result from the array-based voting and the estimated χ^2 per degree of freedom on the points that contributed to the array.

Methods

The automatic landmark annotation process was based on a hierarchical, free-form registration of image patches from the database to the query image. The sequence of processes involved is shown in Figure 1. An initial alignment was derived from four landmark points, by minimising the root-mean-squared (RMS) distances between corresponding points over a nine-parameter affine transformation (i.e. 3D translation, rotation and scale) using the simplex algorithm [47]. Figure 2 shows a typical image volume for experiments in landmark-base geometric morphometrics, a 3D micro-CT image of a *Mus musculus* specimen, together with manual annotations of a typical set of landmarks, described in detail in Table 1. Four typical global registration points are shown in Figure 2b. An automated check was implemented to ensure that the points were not coplanar. This point-based registration could have been decomposed into individual transformations, deriving the translation parameters by aligning the centroids of the four points and the scaling parameters from the standard deviations of the points, leaving only the rotation to be obtained through optimisation. However, in practice all nine parameters of the transformation model were obtained via the optimisation, so that the distance between the centroids of the points in each registered image volume could be used as a semi-independent, automated check.

Once the initial manual alignment had been obtained, it was used to initialise a multi-stage automatic registration process. The first stage was performed on the entire image volume, and optimised a nine-parameter affine transformation using the simplex algorithm [47]. The latter stages operated on patches of image data around each landmark point. As described above, the intention was to terminate the process with patches small enough that the effects of shape variation between the database and query images were minimised. However, registration of small patches had a correspondingly small capture range: in preliminary work, a single stage of patch-based registration proved to be insufficient to attain accuracies equivalent to manual annotation. Therefore, additional stages of patch-based registration were included, with the patch size reduced between each stage; an empirical evaluation of accuracy and time versus the number of registration stages was performed, and the optimal number of patch-based stages was shown to be three, for a total of five stages of registration including the point-based initialisation and the global, image-based stage (see Additional file 1). Some experiments in geometric morphometrics

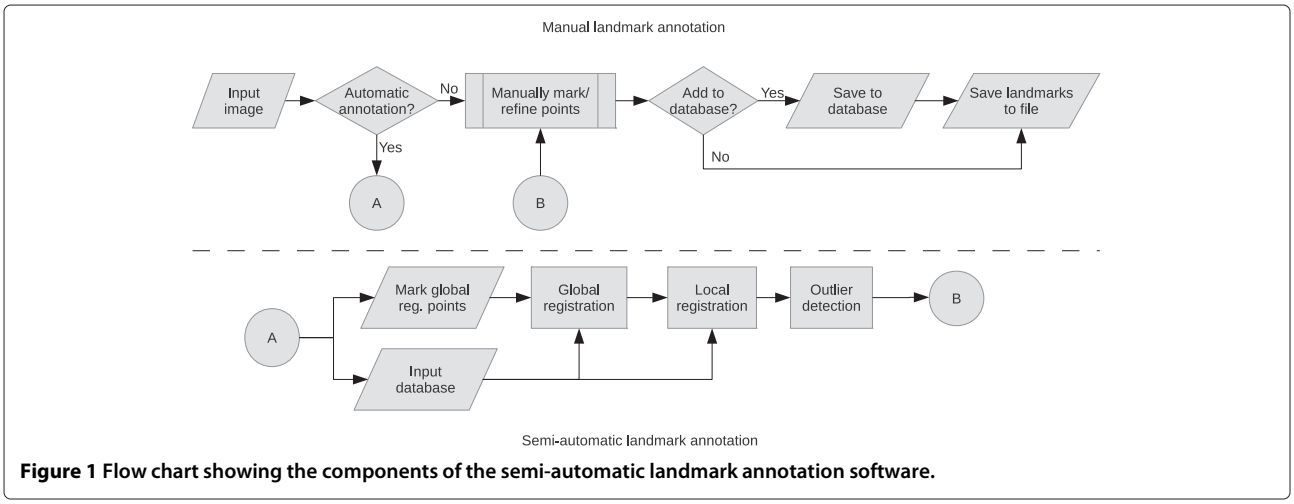

Figure 1 Flow chart showing the components of the semi-automatic landmark annotation software.

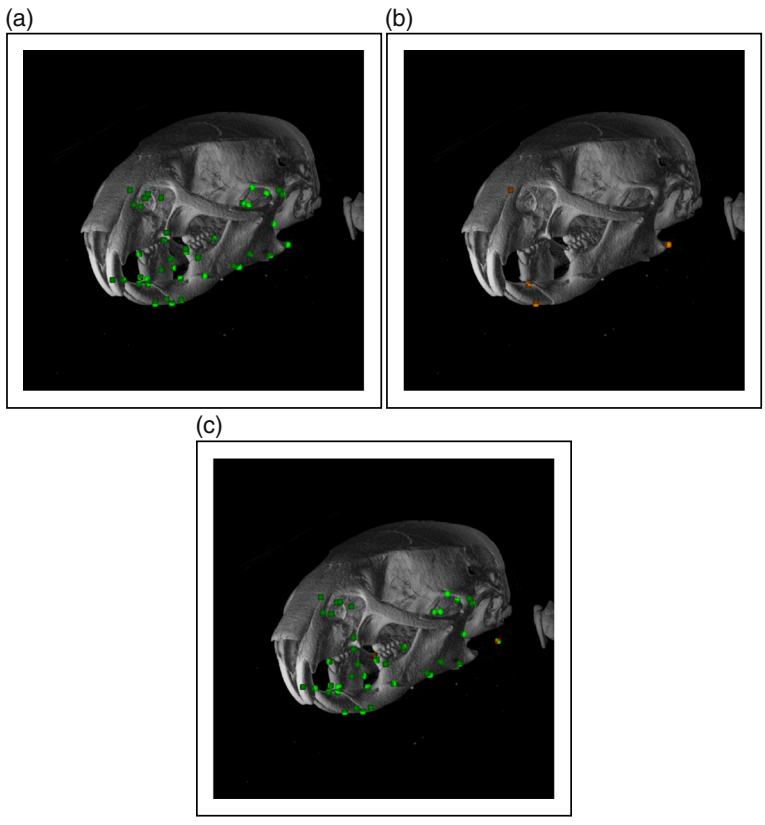

Figure 2 3D renderings of the automatic landmark point annotation process on a *Mus musculus* specimen. (a) Manual annotation of the 40 mandible landmarks. **(b)** Four of the manual landmarks used in the global registration. **(c)** Automatically annotated landmarks, derived using a database of seven *Mus* specimens. Landmarks passing the outlier test are shown in green: those failing are shown in chequered green and red.

may include landmarks specified as the extremal points on curved surfaces from certain view directions. The inclusion of rotation in the patch-based registration would destabilise the registration in such cases (e.g. registration of an arc of a circle to that circle is degenerate over rotation, but not translation or scaling). Therefore, the transformation model used in the patch-based stages of registration consisted only of 3D translation and scaling. Each stage of registration was initialised using the concatenation of the transformation matrices from all previous stages. However, a simple check was included to prevent problems with fit failures; if the result after any stage of registration generated a projected location that lay outside the boundaries of the query image volume, then the parameters of that stage were set to an identity transformation.

Since a nine-parameter affine transformation model was used, the problem was over-determined with realistic patch sizes. Therefore, rather than using cuboids from the data as image patches, 2D slices aligned to the major axes of the image volume and centred on the landmark points were used; in the first stage, operating on the entire image volumes, the slices were taken through the centre points of the volumes. This considerably reduced the amount of data included in the cost function calculation, and so reduced the processor time required whilst still achieving sufficient levels of accuracy.

The cost function for all automatic registration stages was the χ^2 per degree of freedom of the scaled images or image patches (a derivation is provided in Appendix A)

$$\chi^2 = \frac{1}{(N-D)} \sum_{v} \frac{(J_v - \gamma I_v)^2}{\sigma_J^2 + \sigma_I^2 \gamma^2} \quad where \quad \gamma = \frac{|J|}{|I|} \quad (1)$$

where, assuming without loss of generality that I is the source (database) image and J the target (query) image, J_v is the value of voxel v in the target image, I_v is the value of the corresponding voxel in the transformed source image after re-sampling onto the voxel grid of the target image using an interpolation algorithm, N is the number of voxels in the images or image patches, σ_I and σ_J are the standard deviations of the noise on the scaled images or patches, D is the number of transformation model parameters being optimised, and γ is a scaling factor, providing some degree of independence to differences in scanner

Table 1 The 50 skull landmark set

Landmark no.	Description
1	Anterior end of nasal suture.
2	Posterior end of nasal suture.
3	Posterior end of frontal suture.
4	Posterior end of parietal suture.
5	Posterior-most point of occipital.
6	Dorsal-most point of foramen magnum.
7	Anterior-most point of premaxilla behind incisivi.
8	Posterior-most point of palatal suture.
9	Anterior-most medial point of occipital.
10	Anterior-most point of foramen magnum.
11	Anterior end of molar row.
12	Posterior end of molar row.
13	Tip of incisor.
14	Anterior tip of premaxilla.
15	Anterior end of incisive foramen.
16	Posterior end of incisive foramen.
17	Anterior-most dorsal point of infraorbital foramen.
18	Anterior-most lateral point of infraorbital foramen.
19	Anterior-most ventral point of infraorbital foramen.
20	Anterior-most dorsal point of orbita.
21	Anterior-most ventral point of orbita.
22	Dorsal-most point of lateral-most point of zygomatic arch.
23	Ventral-most point of lateral-most point of zygomatic arch.
24	Posterior end of orbita.
25	Anterior-most point of bulla.
26	Anterior-most point of acoustic meatus.
27	Dorsal-most point of bulla.
28	Ventral-most point of bulla.
29	Posterior-most point of bulla.
30	Dorsal end of condyle.

These landmarks were identified on the 12-element data set described in Table 4. Points 11 to 30 were identified on the left-hand-side of the specimen; points 31 to 50 were the equivalent points on the right-hand side of the specimen.

parameters and average bone density. Trilinear interpolation was used in the resampling. The scaling factor was obtained through a maximum-likelihood based approach (see Appendix A). Inevitably, the scaling could be computed only from aligned images. Therefore, it was computed after the point-based stage of registration, and then remained fixed through all of the image-based registration stages. Preliminary work on this topic was described in [48].

In order to reduce the noise on the images and thus provide a smoother cost function, reducing the probability that the optimisation would become trapped in a local minimum, Gaussian smoothing was applied to all image patches prior to registration. The kernel was truncated at three standard deviations from the mean and, in order to ensure that no edge effects were present, a boundary region equal to three standard deviations of the smoothing kernel was added around all stored image patches and included in the smoothing, but excluded from the χ^2 calculation, thus ensuring that all smoothed voxels contributing to the χ^2 were calculated from equal numbers of un-smoothed voxels and avoiding truncation effects.

Rather than operating directly on the image intensities, the cost function was applied to image intensity gradients in order to provide further robustness to differences in average bone density or scanner parameters. This strategy has been found useful in other applications that require matching of similar but non-identical images e.g. stereo pairs of natural scenes [49]. Images showing the gradient components in the x and y directions of each patch were calculated, for each of the three orthogonal patches passing through each landmark point, using finite differencing i.e. taking the intensity difference between neighbouring voxels in the relevant direction. This gave a total of six gradient patches for each landmark. The cost function was applied to each gradient patch separately and the results summed to produce a single χ^2 per degree of freedom. Note that N in Eq. 1 refers to the total number of voxels included in the three orthogonal patches, not the total number in the six gradient patches, since the x- and y-gradients of each patch were obtained from the same original data.

Explicit inclusion of the noise term in the cost function allowed the χ^2 per degree of freedom at the end of the registration to be used as a check on registration error, through comparison to the χ^2 distribution. The noise on the original images was estimated from the width of zero crossings in horizontal and vertical gradient histograms [50]. Smoothing reduced the noise by a factor of $4\pi\sigma_K^2$ where σ_K was the standard deviation of the smoothing kernel [51] (see Appendix B). The calculation of image gradients by finite differencing introduced a further factor of $\sqrt{2}$ into the noise calculation.

Particular care was required in cases where part of an image patch lay outside the volume. Simply ignoring such regions would bias the registration towards moving the patches almost completely outside the query image volume in cases where there was any shape difference. Therefore, any regions lying outside the image volumes were zero-padded and included in the χ^2 calculation. Furthermore, image masks were generated for all patches, recording the zero-padded regions. The cost function was then calculated as three separate terms i.e. one for voxels lying inside both volumes, one for voxels where the I image voxel was zero padded, and one for voxels where the

J image voxel was zero padded; the fourth combination, where both voxels were zero-padded, was identically zero. The correct noise term was used in each case ($\sigma_f^2 + \gamma^2\sigma_I^2$, $\gamma^2\sigma_I^2$, and σ_f^2 respectively, divided by the correction factors for smoothing and differentiation). These three χ^2 terms were then summed prior to division by the number of degrees of freedom, the total number of voxels included in all three terms minus the number of parameters in the transformation model.

Array-based voting

The result from the various stages of registration was a projected location, in the coordinate system of the query image, for each landmark point in each database entry. These constituted multiple estimates of the position of each landmark in the query image, with the number of estimates equal to the number of database entries. The multiple estimates for each landmark were then combined in order to generate a single, final estimate of the location of that landmark in the query image. However, it could not be guaranteed that all estimates were accurate; if one of the database images exhibited significant shape difference, compared to the query image, in the region around one of the landmark points, then it would provide a poor model for that landmark. The corresponding registration would then be likely to fail, resulting in an outlier amongst the multiple estimates for the landmark. Any simple method of combination, such as taking the mean of all estimates, would be affected by the presence of such outliers.

Instead, a method that was robust to outliers was implemented, based on the array-based voting used in methods such as the Hough transform; the same approach has proven reliable in shape-model based approaches to the annotation of landmark points on both clinical and non-clinical images (e.g. [45,52,53]). The voting method was applied to each landmark independently. The multiple estimates for the position of a landmark were analysed to find the maximum and minimum values of their x, y and z coordinates; this specified a 3D volume large enough to contain the estimates. An array of this size was created, and a value of unity was entered into the array at the position of each estimate. The array was then smoothed using a Gaussian kernel, approximating the random error on the registration results. The kernel size was a free parameter of the algorithm (see Additional file 1 and see the Conclusions for further comments). Assuming that the majority of the database images provided good models of the query image, the entries in the voting array would include a compact distribution located close to the correct position for the landmark, representing successful registrations. The width of this distribution would be dictated by the random error on the registration, which would in turn be dictated by the noise on the original images. In contrast, outliers would by definition be scattered far from the correct position. Therefore, smoothing the array with a kernel corresponding to the random error on successful registrations would result in a single, main peak at the position of the compact group of accurate estimates, and a number of smaller, secondary peaks at the positions of outliers. The position of the highest peak in the smoothed array was taken as the final estimate of the landmark location, producing a method that combined the multiple estimates of each landmark location whilst being robust to the presence of outliers.

Prior to creation of the voting array for each landmark point, the projected locations of the corresponding database points in the coordinate system of the query image were compared to the boundaries of that image; any point lying outside the boundary plus a border of three times the standard deviation of the smoothing kernel was omitted from the voting process, in order to prevent problems with severe outliers, on the basis that these points could not contribute to a valid final estimate in any case. The size of the array was then determined by finding the range of the projected locations for the remaining points.

Outlier detection

In order for the final algorithm to have any utility, it was essential that it should provide a reliable indication of the accuracy of automatic annotations; otherwise, much of the manual interaction that it was intended to replace would be re-introduced through a requirement for manual inspection of the results. The array-based voting described in the previous section required that the majority of the multiple estimates of the position of each landmark were located close to the correct position; if this were not the case, due to significant differences in shape between the query image and all of the database images, then the voting would produce an incorrect result. An outlier detection method with an extremely low false positive rate, i.e. an extremely low number of outliers flagged as accurate annotations, was therefore required.

If a given database image patch formed a perfect model of the corresponding query image patch, then the only source of errors on the optimised transformation model parameters would be the random noise on the images. Since a maximum likelihood estimator was used as the registration cost function, the minimum variance bound (MVB; [30,31]) could be applied to estimate this distribution; error propagation could then be applied to find the distribution of the estimated landmark locations. However, in practice there will always be small shape differences between the database and query images, introducing a systematic error onto each patch registration. Assuming these systematic errors to be uncorrelated across the database images, they introduce a secondary

distribution, i.e. the multiple estimated landmark locations generated from the database images will be randomly scattered around the true landmark location in the query image according to a distribution dependent on random shape differences (which could be termed "shape noise"). This will add to the distribution due to random image noise, and so the MVB will provide an underestimate of the true distribution of the estimated landmark locations.

As stated above, [28] and [29] developed a method that measured the shape of the cost function around the optimum using the eigenvectors of the Hessian matrix, allowing the rejection of results where shape differences destabilised the registration to the point where there was no clear optimum. However, this method was limited to analysis of the cost functions for individual registrations. In the work described here, the multiple registration results for each landmark supported an analysis of the distribution of registration parameters induced by shape variation between the database images. The number of samples available was too small to allow a full characterisation of the distribution; therefore, unreliable results were identified through an analysis of the available point estimates. The individual registration results were first sorted into order based on the χ^2 per degree of freedom at the end of the registration process. The distances between the results and the final location generated by the array-based voting were then compared to a threshold, treating each distance as an estimate of the standard deviation of the distribution. The threshold and the number of comparisons performed were free parameters of the algorithm (see Additional file 1); in practice, comparison to three points proved optimal. Any final location for which any one of the three registration results with the lowest χ^2 per degree of freedom was more distant than the threshold was flagged as a potential outlier. The technique therefore imposed a requirement that the patches which were estimated to provide the best models of the query image patch, based on their χ^2 per degree of freedom, formed a distribution within the voting array no broader than the threshold.

Software

The semi-automatic landmark point localisation algorithm was implemented within the TINA Geometric Morphometrics toolkit, which also includes the TINA Manual Landmarking tool [8], and algorithms that perform quantitative shape analysis with weighted covariance estimates for increased statistical efficiency [54]. This package has been made available as free and open source software (FOSS) under the GNU General Public Licence (www.gnu.org), and can be obtained via the TINA web-site (www.tina-vision.net). The User Manuals for the TINA Geometric Morphometrics toolkit and the TINA software are included Additional files 2 and 3, respectively. Figure 1 shows the sequence of operations involved in semi-automatic landmarking, and how the algorithm interfaces with the manual annotation software. Figure 3 shows a screen-shot of the software in operation. The 3D renderings and associated landmark annotations are shown in more detail in Figure 2.

Results and discussion

Evaluation of the algorithm was performed using micro-CT images of rodent skulls with an in-plane resolution of 635×635 voxels and between 1000 and 1500 slices, with voxel dimensions of 0.035 mm along all axes. A detailed description of all image volumes and landmark points used is provided in Tables 1, 2, 3 and 4. The algorithm included a number of free parameters, such as image patch and smoothing kernel sizes for each stage of registration. As described in Additional file 1, these were optimised using a data set of 8 *Mus* specimens of varying species (see Table 2) with expert manual annotation of 40 mandible landmarks (see Table 3). All manual annotations were performed using the TINA software [8].

In order to avoid any bias, the evaluation of automatic point localisation accuracy was performed on a data set of 12 *Mus musculus* specimens from consomic strains, independent of those used in the parameter optimisation experiments (see Table 4). Two sets of expert manual annotations of 50 skull landmarks were performed on each (see Table 1), allowing estimation of manual annotation repeatability; the repeat annotation was performed after an interval of one week, in order to minimise bias due to training effects. Separate experiments were performed for each set of manual annotations in sets of leave-one-out experiments, using 11 specimens to construct the training database and the 12th as the query image, repeating for all 12 image volumes. Four manually annotated points in each volume were used to provide the initial, global alignment; however, the locations of these points were re-estimated by the automatic annotation algorithm, such that results were not contaminated with manual annotations.

A number of extreme outliers (Euclidean distances of over 100 voxels between automatic and manual annotations) were observed in the initial results; detailed investigation revealed that, in all cases, these were due to errors in the manual landmark annotations, consisting of transpositions of equivalent points on either side of the plane of bilateral symmetry. Since these were systematic rather than random errors, they were reported to, and corrected by, the expert; the correction process was limited to a single pass in order to prevent the possibility of experimenter effect. After this, one point transposition error remained in the first set of manual landmarks, and none remained in the second. All results reported here were generated

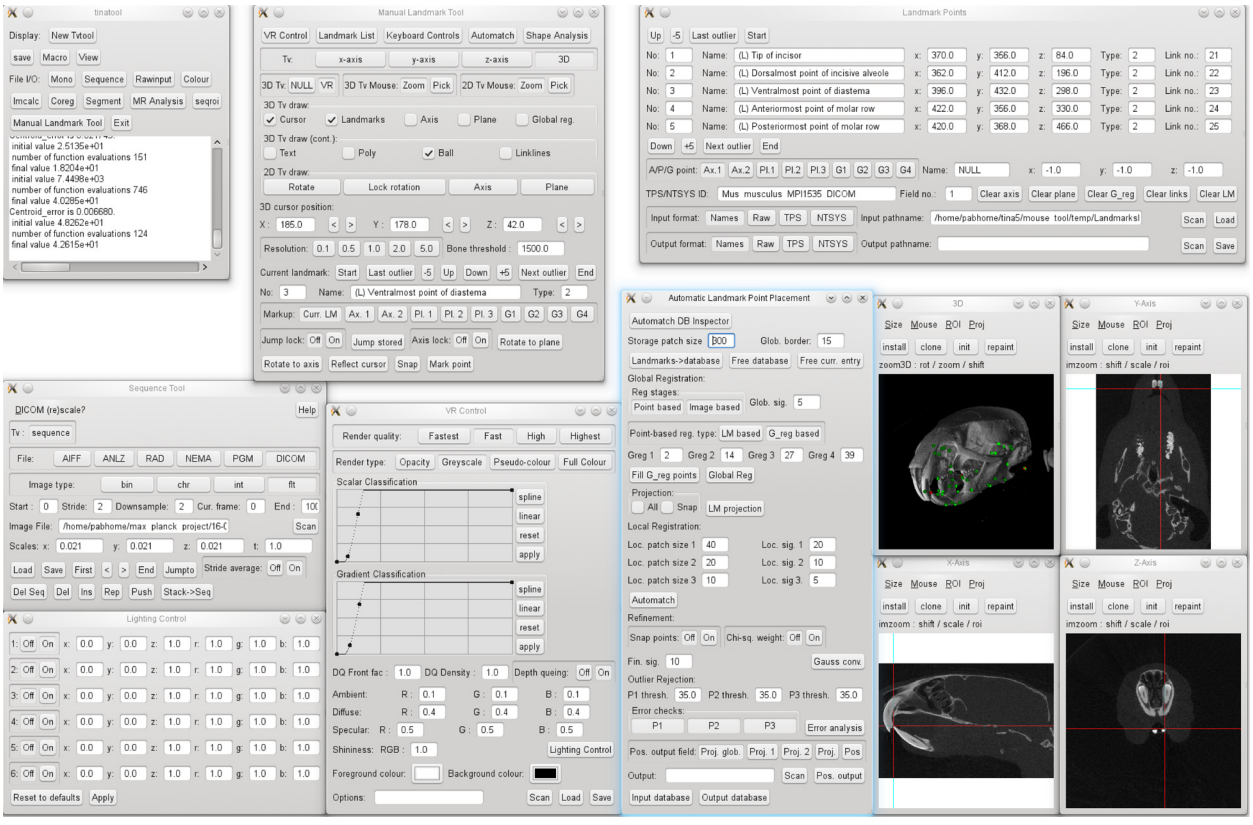

Figure 3 Screen-shot of the semi-automatic landmark annotation software.

Table 2 The 8- and 14-element data sets

Specimen	Description
1	*Mus macedonicus* (Macedonian mouse)
2	*Mus musculus domesticus* (House mouse)
3	*Mus musculus domesticus* (House mouse)
4	*Mus musculus domesticus* (House mouse)
5	*Mus musculus musculus* (House mouse)
6	*Mus musculus musculus* (House mouse)
7	*Mus musculus musculus* (House mouse)
8	*Mus musculus musculus* (House mouse)
9	*Apodemus flavicollis* (Yellow-necked mouse)
10	*Apodemus sylvaticus* (Wood mouse)
11	*Apodemus sylvaticus* (Wood mouse)
12	*Meriones unguiculatus* (Mongolian gerbil)
13	*Microtus fortis* (Reed vole)
14	*Phodopus sungorus* (Djungarian hamster)

The 8-element data set consisted of the first 8 specimens, all of which were various species from the genus *Mus*. The 14-element data set also included specimens from other genera. Expert manual annotation of 40 mandible landmarks was performed for each specimen; see Table 3 for details.

from the corrected landmarks. However, the number of such errors, 4% of the points in the first set of manual landmarks and 0.5% of the points in the second set, was recorded, and the lower value used as a target for the false positive rate of the error detection stage of the automatic point localisation algorithm.

In order to calculate true and false positive and negative rates for the outlier test, a threshold value on point localisation error in voxels was required. This is referred to in the following sections as the error threshold, and was determined using the repeatability of the manual annotations on the 12 consomic *Mus musculus* specimens. As described in Additional file 1, the worst outlier amongst the 50 points in each of the 12 specimens was found by comparing the two sets of manual annotations (after the correction process described in the previous paragraph), and the mean of those values used as the threshold. This gave a numerical value of 30 voxels, equivalent to 1.05 mm at the resolution of the image volumes used here. An automatic annotation was therefore defined as erroneous if its displacement from the corresponding manual annotation was greater than the largest displacement, on average, that would be seen in repeated manual annotations.

Table 3 The 40 mandible landmark set

Landmark no.	Description
1	Tip of incisor.
2	Dorsal-most point of incisive alveole.
3	Ventral-most point of diastema.
4	Anterior-most point of molar row.
5	Posterior-most point of molar row.
6	Lateral point of mandibular foramen.
7	Tip of coronoid process.
8	Anterior-most point of curve between coronoid and articular process.
9	Ventral-most point of curve between coronoid and articular process.
10	Anterior-most point of condyle.
11	Lateral-most point of condyle.
12	Posterior-most point of condyle.
13	Anterior-most point of curve between articular and angular process.
14	Tip of angular process.
15	Medial-most point of angular process.
16	Lateral point of masseteric crista at dorsal-most ventral point.
17	Lateral-most inner point of ventral border.
18	Anterior end of attachment area of transverse mandibular muscle.
19	Posterior-most point of incisive alveole.
20	Anterior end of masseteric crista.

These landmarks were identified on the 8- and 14-element data sets described in Table 2. Landmarks 1–20 consisted of these points on the left-hand side of the specimen; landmarks 21–40 consisted of the same points on the right-hand side of the specimen.

Table 4 The consomic *Mus musculus* specimens included in the 12-element data set

Specimen	Description
1	PWD/Ph (wild-derived *Mus musculus musculus* strain)
2	PWD/Ph (wild-derived *Mus musculus musculus* strain)
3	PWD/Ph (wild-derived *Mus musculus musculus* strain)
4	PWD/Ph (wild-derived *Mus musculus musculus* strain)
5	C57BL/6J-10d$^{PWD/Ph}$
6	C57BL/6J-10d$^{PWD/Ph}$
7	C57BL/6J-7$^{PWD/Ph}$
8	C57BL/6J-7$^{PWD/Ph}$
9	C57BL/6J-7$^{PWD/Ph}$
10	C57BL/6J (*Mus musculus domesticus* background)
11	C57BL/6J (*Mus musculus domesticus* background)
12	C57BL/6J (*Mus musculus domesticus* background)

This data set included several specimens of each pure strain (C57BL/6J and PWD/Ph), together with specimens in which chromosomes 7 or 10 from the *Mus musculus musculus* strain PWD/Ph had been substituted into the *Mus musculus domesticus* strain C57BL/6J. Expert manual annotation of 50 skull landmarks was performed for each specimen; see Table 1 for details.

Figure 4 shows the results of the leave-one-out experiments performed on the data set of 12 image volumes of consomic *Mus musculus* specimens using the second set of manual landmarks to produce the database and using four manual annotations to perform the initial, manual stage of alignment. The results are presented as box-and-whisker plots of the Euclidean distance in voxels between the automatic and manual annotations for each point, showing the median, minimum, maximum, 25th and 75th percentiles. The ROC curve, generated by varying the outlier test threshold, shows that the threshold and number of points included, derived from an independent data set, were approximately optimal for these data; there were no false positives at the operating point used, i.e. all 525 points passing the test were within the error threshold of 30 voxels. Across all twelve volumes, 87.5% of the automatic annotations passed the outlier test, and 98.3% were within the error threshold. In a hypothetical annotation process with a pre-built database, this level of performance equates to a potential user of the software having to manually inspect 12.5% of the points, but having to correct the positions of fewer than 2% of the points, i.e. fewer than 12 points. Therefore, combined with the four points per specimen required for the initial, point-based stage of registration, the user would be required to manually annotate an average of 5 points per specimen out of a total of 50 landmark points i.e. one tenth of the total number of points.

The results show that the automatic landmark localisation algorithm placed points to within a median of 3.60 voxels of the manual annotations. However, this value contains contributions from the errors on both the manual and automatic annotations, i.e. it does not represent the random error on the automatic landmark annotations. In order to quantify this random error, the repeatability of the automatic annotation process was evaluated by comparing the automatic landmarks generated from databases constructed from the two sets of manual annotations. Figure 5 shows the result of this comparison, together with the repeatability of manual annotation. The median Euclidean distance between the two sets of manual annotations across all points in all image volumes was 3.34 voxels: assuming that the error distributions on both sets of manual annotations were the same, this implies that the median manual annotation error on a single point was approximately 2.4 voxels. The mean of the worst outliers in all 12 volumes was 29.3 voxels. The median distance between the two sets of automatically located landmarks across all points in all volumes was 1.4 voxels;

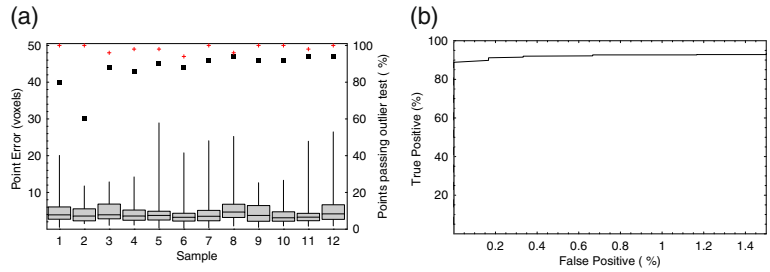

Figure 4 Automatic landmark annotation accuracy using an initial, point-based registration. (a) Box-and-whisker plots of the point localisation errors for automatic annotation of 50 skull landmarks on the 12 consomic *Mus musculus* specimens (read against the left-hand scale), using the second set of manual landmarks and an initial, point-based registration stage. The black squares and red crosses show, respectively, the percentage of points passing the outlier test and the percentage within the error threshold (read against the right-hand scale); only points passing the outlier test have been included in the box-and-whisker plots. **(b)** ROC curve of the true and false positive rates of points passing the outlier test; the operating point of the test is coincident with the y-axis.

again assuming equal errors on the two sets of results, the median of the automatic annotation error on a single point was approximately 1.0 voxels. The mean of the worst outliers in all twelve volumes was 7.4 voxels. Only automatically located points passing the outlier test were included in the results; this was 81.8% of the points across all 12 volumes. Due to the non-Gaussian distribution of the annotation errors, the manual and automatic repeatabilities were compared using a Mann-Whitney U test, which conclusively demonstrated ($U = 72116$, $U_\mu = 147300$, $U_\sigma = 5187$, $Z = -14.52$, $p \approx 0$) that the automatic repeatability was significantly better than the manual repeatability.

Evaluating the repeatability of the automatic annotation did not evaluate its accuracy; such an evaluation was not possible without a set of gold-standard (i.e. error-free) landmark locations. However, comparison of the median Euclidean distance in voxels between the two sets of manual annotations (3.34 voxels) and the median Euclidean distance between the second set of manual annotations and the automatic annotations derived from them (3.6 voxels) indicated no statistically significant difference (Mann-Whitney $U = 149405$, $U_\mu = 157200$, $U_\sigma = 5429$, $Z = -1.44$, $p = 0.15$). This indicated that automatic annotation was not significantly less accurate than manual annotation.

Double iteration without point-based registration

The reliability of the outlier test, demonstrated in the experiments described above, allows an alternative mode of operation for the algorithm that can potentially eliminate the need for manual annotation of the four points used in the initial stage of global registration. The gross misalignments between image volumes, for which this stage of registration was designed, can be avoided if care is taken during the preparation of specimens, such that they are all scanned in approximately the same orientation. The algorithm can then be applied in two stages; a first pass, with no point-based registration, generates an intermediate set of automatic annotations. Fewer points will pass the outlier test; however, the points that do can be used to perform the point-based stage of registration in a second pass of the algorithm. A second set of experiments, identical to those described above but using this

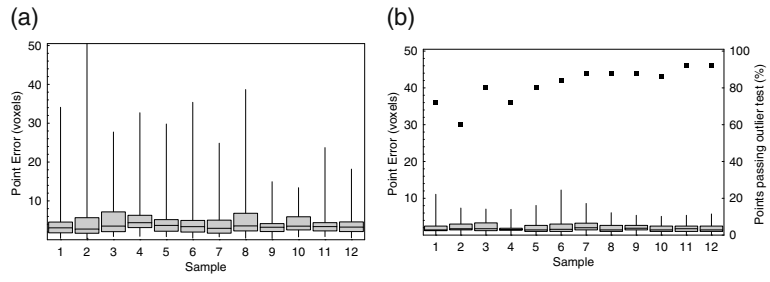

Figure 5 Repeatability of manual and automatic landmark annotation. Box-and-whisker plots of the repeatability of manual **(a)** and automatic **(b)** (read against the left-hand scale) localisation of 50 skull landmarks on the 12 consomic *Mus musculus* specimens. The automatic landmarks were generated using an initial, point-based stage of global registration. Only points passing the outlier test were included in the automatic annotation results; the black squares show the percentage of points passing (read against the right-hand scale).

"double-pass" mode of operation, was performed in order to test this approach. The results for automatic annotation using a database built from the second set of manual annotations are shown in Figure 6. The median Euclidean distance between the manual and automatic annotations across all twelve volumes was 3.49 voxels, compared to 3.60 voxels for the equivalent experiment in single-pass mode, but the difference was not statistically significant (Mann-Whitney $U = 134236$, $U_\mu = 137550$, $U_\sigma = 4906$, $Z = -0.68$, $p = 0.50$). Similarly, there was no significant difference between the number of points passing the outlier test (87.5% in single-pass and 87.3% in double-pass mode) or the percentage of points with errors lower than the error threshold (98.3% in single-pass and 97.8% in double-pass mode). The ROC curve shows that the outlier test parameters, derived from an independent data set, were approximately optimal for these data; the false positive rate of the outlier test was 0.17%, i.e. across the 600 points in the 12 volumes, i.e. only one point with an error larger than the threshold was not flagged as an outlier. Figure 7 shows a comparison of the repeatability of manual and automatic landmark annotation when the software was used in double-pass mode; as with the single-pass mode, the automatic repeatability was significantly better (Mann-Whitney $U = 77551$, $U_\mu = 146700$, $U_\sigma = 5162$, $Z = -13.39$, $p \approx 0$). However, it should be noted that the double-pass mode of the algorithm is dependent on the alignment of the specimens within the scanner, and is likely to fail if significant misalignments are present.

Multiple-genera database

In order to evaluate the robustness of the algorithm to databases containing specimens with significant shape differences, a further data set consisting of 14 micro-CT image volumes of rodent skulls from multiple genera was used (see Table 2), with manual annotations of 40 mandible landmarks on each (see Table 3). This was a superset of the 8 *Mus* skull data set used in the parameter optimisation experiments, and was therefore not completely independent (i.e. the free parameters of the algorithm were partially derived from this data set), although the evaluation of the free parameters showed a high degree of independence between performance and parameter values (see Additional file 1 for details). In addition to the 8 *Mus* specimens, the data set included one *Apodemus flavicollis*, two *Apodemus sylvaticus*, one *Meriones unguiculatus*, one *Microtus fortis* and one *Phodopus sungorus* specimen. This combination of specimens was chosen to exhibit a range of shape variation, i.e. the *Apodemus* specimens were more similar in shape to the *Mus* specimens than were the *Microtus*, *Phodopus* or *Meriones* specimens. The skull constituted the majority of the bone surfaces in the images and so dominated the registration result; the mandible is not rigidly fixed to the skull, and so the use of mandible landmarks provided a more significant challenge to the algorithm and was more suitable to illustrate failure modes. As above, the automatic point localisation algorithm was applied to the data in a set of leave-one-out experiments, using 13 image volumes to construct the database and predict landmark locations in the 14th volume, repeating for all 14 volumes. A second set of experiments was conducted using only the eight *Mus* specimens from the data set. The results are shown in Figures 8 and 9.

The effects of building the database from multiple genera can clearly be seen in the breakdown of the results by genus. During annotation of the *Microtus*, *Phodopus* and *Meriones* specimens, which varied significantly in shape from the *Mus* and *Apodemus* specimens, there were no similar specimens in the database and consequently the outlier test rejected all points. The median landmark error across the *Mus* specimens was slightly lower using a mixed-genera database than using a *Mus*-only database

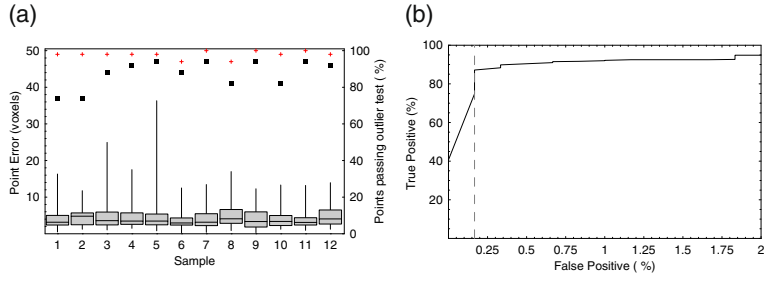

Figure 6 Automatic landmark annotation accuracy in "double-pass" mode. (a) Box-and-whisker plots of the point localisation errors for automatic annotation of 50 skull landmarks on the 12 consomic *Mus musculus* specimens (read against the left-hand scale), using the second set of manual landmarks. The algorithm was applied in "double-pass" mode with no initial, point-based stage of registration. The black squares and red crosses show, respectively, the percentage of points passing the outlier test and the percentage within the error threshold (read against the right-hand scale); only points passing the outlier test have been included in the box-and-whisker plots. **(b)** ROC curve of the true and false positive rates of points passing the outlier test; the dashed line shows the operating point.

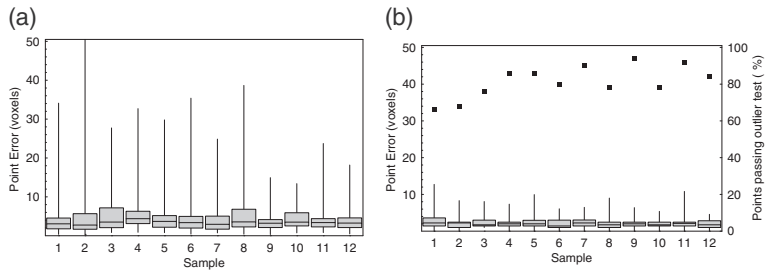

Figure 7 Repeatability of manual and "double-pass" automatic landmark annotation. Box-and-whisker plots of the repeatability of manual **(a)** and automatic **(b)** (read against the left-hand scale) localisation of 50 skull landmarks on the 12 consomic *Mus musculus* specimens. The automatic landmarks were generated using the "double-pass" mode with no initial, point-based stage of global registration. Only points passing the outlier test were included in the automatic annotation results; the black squares show the percentage of points passing (read against the right-hand scale).

(5.11 voxels compared to 5.24 voxels); however, the difference was not statistically significant (Mann-Whitney $U = 21464$, $U_\mu = 21830$, $U_\sigma = 1239$, $Z = -0.30$, $p = 0.77$), indicating that the additional specimens added little information. However, their presence did lead to a marked reduction in the number of points passing the outlier test (57.8% compared to 73.8%), although the percentage of points with errors lower than the error threshold was not significantly different (95.0% vs. 95.3%). The median annotation errors on the *Apodemus* specimens were larger than those on the *Mus* specimens (6.83 voxels vs. 5.11 voxels), but the difference was on the borderline of statistical significance (Mann-Whitney $U = 4545$, $U_\mu = 5428$, $U_\sigma = 506$, $Z = -1.75$, $p = 0.08$).

These results serve to indicate the robustness of the algorithm to variations in the database. When specimens from multiple genera with significant variation in shape were entered into the database, only those with similar shape to the query image contributed information to the final landmark location estimate. Conversely, for those specimens where the database provided no usable information, the algorithm successfully indicated that the automatic annotation was not reliable and rejected all points. Contamination of the database with multiple genera did not result in a significant decrease in landmark annotation accuracy, but did result in a large reduction in the number of points passing the outlier test, reflecting the bias of the outlier test towards low false positive rates.

Database size and processor time requirements

The dependence of algorithmic performance on the number of image volumes in the database was evaluated by repeating the leave-one-out experiments on the 12 consomic *Mus musculus* specimens with fewer database entries. In order to avoid confounding the results by varying several features of the experimental procedure simultaneously, databases with fewer than three entries (the smallest number required to perform the outlier test) were not considered, and the image volumes included were selected randomly. The results are shown in Figure 10; each box-and-whisker shows the point localisation errors in voxels across all 12 volumes. Only points passing the outlier test were included, and the percentage of such points is also shown. The results demonstrate that, as would be expected, the point localisation errors decreased

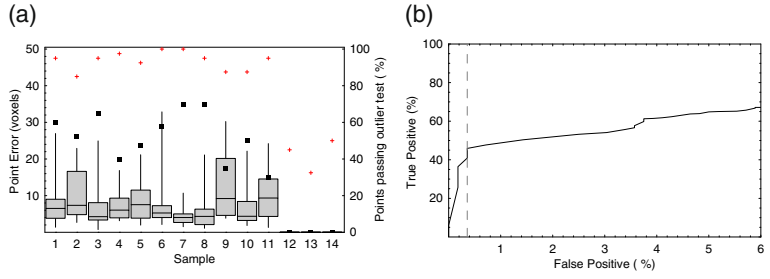

Figure 8 Automatic landmark annotation accuracy using a mixed-genera database. (a) Box-and-whisker plots of the point localisation errors for automatic annotation of 40 mandible landmarks on the data set of 14 rodent specimens (read against the left-hand scale). The black squares and red crosses show, respectively, the percentage of points passing the outlier test and the percentage within the error threshold (read against the right-hand scale); only points passing the outlier test have been included in the box-and-whisker plots. **(b)** ROC curve of the true and false positive rates of points passing the outlier test; the dashed line shows the operating point.

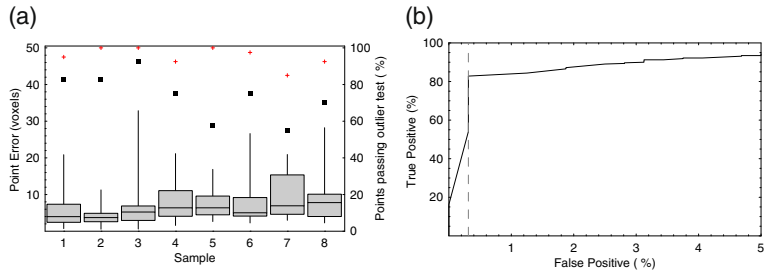

Figure 9 Automatic landmark annotation accuracy using 8 *Mus* specimens from the mixed-genera database. (a) Box-and-whisker plots of the point localisation errors for automatic annotation of 40 mandible landmarks on the data set of 8 *Mus* specimens (read against the left-hand scale). The black squares and red crosses show, respectively, the percentage of points passing the outlier test and the percentage within the error threshold (read against the right-hand scale); only points passing the outlier test have been included in the box-and-whisker plots. **(b)** ROC curve of the true and false positive rates of points passing the outlier test; the dashed line shows the operating point.

with increasing database size and the number of points passing the outlier test increased. However, both dependencies were relatively weak with this data set; there was little improvement in performance with database sizes larger than eight entries. This is significantly fewer images than would be required for the construction of an appearance model.

The processor time requirements of the algorithm were also evaluated during the tests on database depth. Experiments were performed on a Dell Precision workstation with 2 Intel Xeon 5670 processors and 24 Gb of main memory, running OpenSuse 11.3×64 (Linux kernel 2.6.34). It was anticipated that, in practical use, a single database would be constructed and then used to annotate multiple image volumes. Therefore, the database and image loading times were ignored and the wall-clock time required to perform the registration, array-based voting and outlier test was measured. Figure 11 shows the results, averaged over all experiments performed at each database size, as the number of seconds required to annotate a single point. The vast majority of the time taken to run the algorithm was accounted for by the registrations and so, since the registrations were performed independently for each database entry, there was a linear dependence on the number of entries. The gradient of the linear fit, i.e. the time taken per point per database entry, was 1.64 seconds. In [8], the time required to manually annotate 10 landmarks on the mandible in micro-CT images of three rodent specimens (*Microtus*, *Mus* and *Pachyuromys*) was evaluated in both TINA and AMIRA (www.vsg3d.com/amira), both by a non-expert and an expert AMIRA user. Average timings were 13s per point using AMIRA and 30s per point using TINA for the expert, and 54s per point using AMIRA and 65s per point using TINA for the non-expert, although strong training effects were observed with TINA, as would be expected of users handling unfamiliar software. Therefore, assuming a reasonable database size of around 8 entries and thus approximately 13s per point for automatic annotation, the time required to perform automatic annotation was comparable to the time required to perform manual annotation with AMIRA. However, unlike manual annotation with AMIRA, automatic annotation with TINA does not require continuous user input. Comparison between TINA and AMIRA in terms of manual annotation accuracy is provided by [8]; note that AMIRA does not provide automatic annotation.

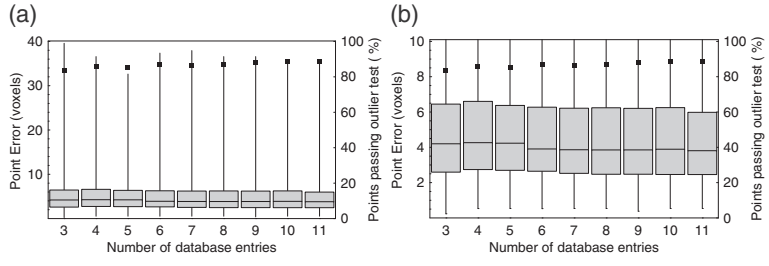

Figure 10 Dependence of automatic annotation accuracy on database size. Box-and-whisker plots of the point localisation errors for automatic annotation of 50 skull landmarks in the 12 consomic *Mus musculus* specimens, (read against the left-hand scale), against the number of image volumes in the database. The black squares show the percentage of points that passed the outlier test (read against the right-hand scale); only points passing the outlier test have been included in the box-and-whisker plots. Both graphs show the same data, plotted on different ranges.

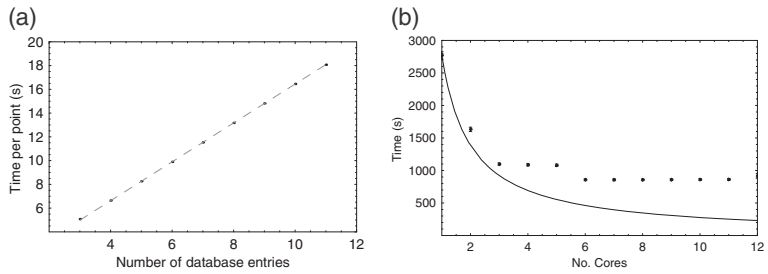

Figure 11 Dependence of run time on database size and the number of cores in use. (a) The average wall-clock time required to perform all registration stages, array-based voting and the outlier test on the 50 skull landmarks in each of the 12 consomic *Mus musculus* specimens, plotted against the number of image volumes in the database. The dashed line shows a linear fit to the data. **(b)** The average wall-clock time required to perform all registration stages, array-based voting and outlier test on 50 skull landmarks in each of the the 12 consomic *Mus musculus* specimens with 11 database entries, plotted against the number of processor cores used. The 1/cores curve that would be achieved with 100% parallelisation efficiency is also shown.

The software made extensive use of parallelisation, and so the dependence of the run time on the number of processor cores in use was also evaluated. Figure 11 shows the time taken to perform the registration, array-based voting and outlier test stages of the algorithm in leave-one-out tests on the 12 consomic *Mus musculus* specimens (i.e. with a database size of 11), averaged over the 12 experiments, against the number of processor cores, together with the 1/cores dependency that would be expected if the parallelisation efficiency was 100%. The results showed good parallelisation efficiency up to three cores, little further reduction in the time taken until six cores were in use (allowing full parallelisation over the six image patches for each landmark point; see the Methods section for details), and then no further reduction. The loss of parallelisation efficiency above three cores indicates that memory bandwidth was the main limiting factor, due to the algorithm performing large numbers of relatively simple operations on small blocks of data. However, these results indicate that the timings described above should be achievable on most relatively modern hardware.

Conclusions

Geometric morphometric analyses, consisting of manual annotations of landmark points followed by Procrustes analysis, are a popular way to quantify biological shapes for comparison to genetic, phylogenetic or ecological factors. The requirement for extensive and time-consuming manual annotation places practical limitations on the number of specimens that can be included in such analyses, in turn limiting their statistical power both in terms of the magnitude of shape variation that can be measured and the confidence intervals on any conclusions. In this paper, a semi-automatic landmark annotation algorithm designed to accelerate the landmark annotation process has been presented.

The aim was to produce landmarks for use in shape analysis, and so no constraint based on a shape model, or assumptions about shape, could be used in the algorithm. These models or assumptions would be recovered by the subsequent analysis and therefore potentially bias the results if they were not a perfect fit to the data. Instead, a free-form, intensity-based registration approach was adopted. Multiple example images, each with manually annotated landmarks, were registered to the query image using a multi-scale approach. The final stages of registration operated on small patches of image data around each landmark in order to minimise the effects of global shape variation. This resulted in one set of estimated landmark locations for each database image; these were combined using a Hough-like voting array in order to introduce robustness to biases and outliers whilst minimising assumptions about their distribution.

Since the algorithm was designed to replace manual annotation, it would have little utility if it required extensive manual checking of the results. Therefore, it had to provide some indication of the reliability of the automatic annotation in order to guide manual checking and, where necessary, correction of the results. Furthermore, the operating point on the ROC curve of the outlier test had to be biased such that the false positive rate was minimal, i.e. it was essential to flag all incorrect annotations as errors, even at the expense of flagging a significant proportion of correct annotations, in order that the user could trust the outlier test and thus avoid having to check all automatic annotations. The presence of outliers indicated either convergence to a local minimum or a database image that provided a poor model of the query image; the latter possibility rendered all statistical measures based on the assumption that the model fits the data, unreliable. Therefore, a method based on consistency between the multiple estimates for each landmark was developed. This treated each database image as an independent

model of the query image, and used a minimal assumption that good models should produce a compact peak in the Hough-like voting array, whilst poorly fitting models should produce a broad outlier distribution.

In order to have utility within the target user community, the algorithm had to perform landmark annotation as quickly and as accurately as manual annotation; therefore, evaluation focused on these two properties. Comparisons were performed between two independent sets of manual annotations of 50 skull landmarks in 12 *Mus musculus* specimens, and two sets of automatic annotations generated from them. Automatically annotated points failing the outlier test were not included in the comparisons. For the 87.5% of the points that were included, the results showed that automatic annotation was at least as accurate as, and more repeatable than, manual annotation, i.e. had a lower random error and no significant systematic error. The repeatability of the automatic annotations indicated that voxel-level accuracy was achieved. The outlier test proved extremely reliable, rejecting 12.5% of the automatic annotations to achieve a false positive rate of less than 0.5%, due to the care taken in the estimation of noise distributions.

The evaluation on a wider variety of rodent specimens, including multiple genera, demonstrated the reliability of the outlier test. In cases where the database contained no specimen that provided a good model of the query image, the algorithm correctly flagged all points as outliers. Conversely, the presence of database specimens that provided poor models of the query image did not significantly affect the accuracy of the automatic annotation process, even where they formed the majority of the database. This suggests an iterative mode of operation; any query image that generates large numbers of outliers is not well modelled by the existing database. Therefore, after manual correction of the outliers, it can be added to the database in order to expand the number of specimens that can be annotated accurately.

Timing tests demonstrated that the automatic landmark localisation process required approximately the same wall-clock time as manual annotation on reasonably modern hardware. However, no user input was required for this stage of the process. Therefore, actual manual annotation required for each image volume in a hypothetical landmark annotation process would be limited to the four registration points, visual check of the detected outliers, the majority of which would be in the correct location due to the low false-positive bias of the outlier detector, and correction of the true outliers, which averaged around 1 to 2 points per volume with a single-genera image database. For realistic landmark list sizes of 40 to 50 points, the automatic landmark localisation software could therefore reduce the required number of manual landmark annotations by a factor of ten. Further improvements were achieved using the double-pass mode of operation, although this would require care during sample preparation to ensure a reasonably good alignment of the samples within the image volumes. Annotation of the training images constitutes the majority of the manual annotation required when using the proposed technique. However, the evaluation of database size in the experiments on consomic *Mus musculus* specimens indicated that only eight images were required, far fewer than would be required by alternative techniques based on statistical shape models.

Several potential routes exist for future improvement of the software. For instance, it may be possible to automatically estimate patch sizes using techniques such as those described in [12], reducing the need to optimise the free parameters of the algorithm prior to application to new image types. Furthermore, annotation of the four points used to initialise the global registration requires significant user interaction. This could be simplified using the 3D rendering of the data provided by the software, by prompting the user to rotate the rendering into a standard orientation, which would then provide the initialisation. A similar orientation would have to be stored for each of the images in the database.

The software described in this paper has been made available as free and open source (FOSS) software under the GNU General Public Licence (www.gnu.org), and can be obtained via our web-site (www.tina-vision.net).

Appendix A: derivation of the cost function and scaling factor

Let I_ν and J_ν be corresponding voxels in two identical, noise-free images or image patches I and J. Scale the intensities of one of the images by a factor γ; without loss of generality, assume that this is image J. Add Gaussian random noise with standard deviations of σ_I and σ_J to the two images. Find an estimator for γ.

The derivation given here follows that given in [55]. An estimator of γ can be obtained as the gradient of a linear fit to J_ν vs. I_ν over all ν with an intercept of zero, as shown in Figure 12. Since there is noise on both I and J, the noise-free intensities of a point $A = I_\nu, J_\nu$ could correspond to any point C along the linear fit. However, if $\sigma_I = \sigma_J = \sigma$, the probability that a datum at A was generated from C is given by

$$\begin{aligned} P(C \rightarrow A) &= \frac{1}{2\pi\sigma^2} exp\left[-\frac{(I_C - I_A)^2}{2\sigma^2}\right] exp\left[-\frac{(J_C - J_A)^2}{2\sigma^2}\right] \\ &= \frac{1}{2\pi\sigma^2} exp\left[-\frac{r^2}{2\sigma^2}\right] \\ &= \frac{1}{2\pi\sigma^2} exp\left[-\frac{h^2}{2\sigma^2}\right] exp\left[-\frac{u^2}{2\sigma^2}\right] \end{aligned} \quad (2)$$

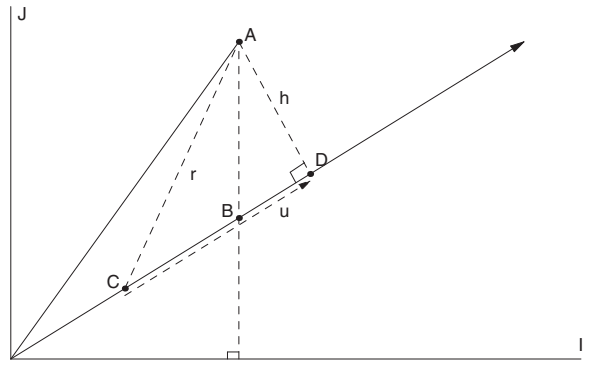

Figure 12 The construction of the χ^2 between two scaled image patches. A linear fit to the joint intensity histogram of a pair of images *I* and *J*. Since both images contain noise, a datum at point A could be generated from any point C along the linear fit. However, if the standard deviations of the noise on *I* and *J* are equal, then the probability of generating data at A depends only on the perpendicular distance *h* to the linear fit, not on *u*.

Integrating over u gives a constant $\sigma\sqrt{2\pi}$, so the probability depends only on h i.e. only on the perpendicular distance to the line. In the event that $\sigma_I \neq \sigma_J$, the images can be scaled to $I' = I/\sigma_I$ and $J' = J/\sigma_J$.

The distance h can be obtained from the vectors $A = (I_v, J_v)$ and $B = (I_v, \gamma I_v)$ using

$$A.B = |A||B|\cos\theta \quad , \quad \frac{|h|}{|A|} = \sin\theta \quad and \quad \cos^2\theta + \sin^2\theta = 1$$

where θ is the angle between the vectors A and B. After some manipulation, these give

$$h = \frac{J_v - \gamma I_v}{\sqrt{1+\gamma^2}}$$

Substituting this into Eq. 2 gives

$$P(J_v|I_v, \sigma, \gamma) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(J_v - \gamma I_v)^2}{2\sigma^2\left(1+\gamma^2\right)}\right] \qquad (3)$$

or equivalently

$$\chi^2 = \sum_v \frac{(J_v - \gamma I_v)^2}{\sigma^2\left(1+\gamma^2\right)} \propto \sum_v \frac{(J_v - \gamma I_v)^2}{1+\gamma^2}$$

In order to obtain an estimator for γ, differentiate the χ^2 w.r.t. γ and set the result equal to zero. After some manipulation, this gives

$$\sum_v \frac{\gamma^2 I_v J_v + \gamma\left(I_v^2 - J_v^2\right) - I_v J_v}{\left(1+\gamma^2\right)^2} = 0$$

The numerator in this fraction must be equal to zero, since the denominator cannot be regardless of the value of γ. Therefore, substituting

$$|I|^2 = \sum_v I_v^2 \quad , \quad |J|^2 = \sum_v J_v^2 \quad and \quad I.J = \sum_v I_v J_v$$

gives

$$\gamma^2 I.J + \gamma\left(|I|^2 - |J|^2\right) - I.J = 0$$

This quadratic can be solved in the usual way to give $\hat{\gamma}$, an estimator for γ

$$\hat{\gamma} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad where \quad a = I.J \quad , \quad b = |I|^2 - |J|^2$$
$$and \quad c = -I.J$$

This gives

$$\hat{\gamma} = \frac{|J|^2 - |I|^2 \pm \sqrt{\left(|I|^2 - |J|^2\right)^2 + 4\,(I.J)^2}}{2I.J}$$

Now, $I.J = |I||J|cos\phi$ where, if the intensities from I and J were concatenated to form two vectors in a v-dimensional space, ϕ would be the angle between those two vectors. This angle will be small if the signal-to-noise ratio is high, and so assuming that $\phi \approx 0$,

$$\hat{\gamma} = \frac{|J|^2 - |I|^2 \pm \sqrt{\left(|I|^2 + |J|^2\right)^2}}{2|I||J|}$$

which gives

$$\hat{\gamma} = \frac{|J|}{|I|} \quad or \quad \hat{\gamma} = -\frac{|I|}{|J|}$$

The two solutions are perpendicular and give the best and worst linear fit to the data. The positive estimator is used when the correlation between I and J is positive, and the negative estimator when the correlation is negative. A negative correlation may be seen with some image modalities, such as MR images acquired with different pulse sequences; however, in the work presented here micro-CT images were used, and so only the positive solution is relevant. The result is both a maximum likelihood and minimum χ^2 estimator.

The above derivation also provides the cost function for registration of I and J used in the work described here as Equation 3

$$\chi^2 = \sum_v \frac{(J_v - \gamma I_v)^2}{\sigma^2\left(1+\gamma^2\right)}$$

This is valid only if the noise on I and J are equal. In the more general case where they are not equal, the images can be scaled to to $I' = I/\sigma_I$ and $J' = J/\sigma_J$; let γ' represent γ estimated in the I', J' space, so that the cost function becomes

$$\chi^2 = \sum_v \frac{\left(J_v/\sigma_J - \gamma' I_v/\sigma_I\right)^2}{\left(1+\gamma'^2\right)}$$

Rearranging gives

$$\chi^2 = \sum_v \frac{\left(J_v - \gamma'\frac{\sigma_J}{\sigma_I} I_v\right)^2}{\sigma_J^2\left(1+\gamma'^2\right)}$$

and, since

$$\gamma' = \frac{|J'|}{|I'|} = \frac{\sigma_I}{\sigma_J}\frac{|J|}{|I|}$$

letting $\gamma = |J|/|I|$, the cost function becomes

$$\chi^2 = \sum_v \frac{(J_v - \gamma I_v)^2}{\sigma_J^2 + \sigma_I^2 \gamma^2}$$

The requirement for the images to be similar before the scaling factor can be estimated is not an issue when the manual, point-based stage of registration is included, since this will achieve the required approximate alignment independently of the voxel intensities. Less obviously, it is also not a problem when operating the algorithm in double-pass mode, due to a feature of registration using images consisting primarily of step edges between uniform regions. Since the cost function being optimised is

$$\chi^2 = \sum_v \frac{(J_v - \gamma I_v)^2}{\sigma^2\left(1 + \gamma^2\right)}$$

and assuming without loss of generality that J is the source image, the solution will be obtained where

$$\frac{\partial \chi^2}{\partial \mathbf{T}} = 0 = \frac{\partial \chi^2}{\partial J_v}\frac{\partial J_v}{\partial \mathbf{T}} \propto \sum_{i=1}^{N}\left(J_v - \gamma I_v\right)\frac{\partial J_v}{\partial \mathbf{T}}$$

where $\mathbf{T}$ represents the parameter vector of the transformation model. The $\partial J_v/\partial \mathbf{T}$ term ensures that only regions with a significant image gradient contribute to the alignment process. In images with large regions of smooth gradients, errors in the estimate of γ therefore have a significant effect on the accuracy of the registration result. However, micro-CT images consist primarily of significant step edges between regions that are comparatively uniform except for the effects of noise, and so the optimisation will attain a (possibly local) minimum that aligns the edges regardless of the estimate of γ used. Therefore, the first, image-based stage of global registration can be performed with $\gamma = 1$, in order to attain an approximate alignment. The equations given above can then be used to estimate γ, prior to the later, patch-based stages of registration.

Appendix B: the effects of smoothing on noise

If an image $I(x, y)$ with a uniform Gaussian noise field of standard deviation σ_I is smoothed through convolution with a Gaussian kernel of standard deviation σ_k, what is the standard deviation of the noise field on the smoothed image?

The full form for the 2D Gaussian distribution is

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp\left[\frac{-1}{2\left(1-\rho^2\right)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{\left(y-\mu_y\right)^2}{\sigma_y^2} - \frac{2\rho\left(x-\mu_x\right)\left(y-\mu_y\right)}{\sigma_x\sigma_y}\right)\right]$$

where σ_x and σ_y are the standard deviations in the x and y directions, ρ is the correlation coefficient, and μ_x and μ_y give the position of the mean. Assuming that the smoothing kernel will be an isotropic Gaussian, $\sigma_x = \sigma_y = \sigma_k$ and $\rho = 0$, so

$$G(x,y) = \frac{1}{2\pi\sigma_k^2} exp\left[\frac{-(x-\mu_x)^2 - \left(y-\mu_y\right)^2}{2\sigma_k^2}\right]$$

The smoothed image $I_G(x, y)$ is given by

$$I_G(x,y) = \sum_{x,y} I(x,y).G(x,y)$$

Error propagation [55] gives the standard deviation σ_f of the noise on a function of several random variables n_i,

$$\sigma_f^2 = \sum_i \frac{\partial f}{\partial n_i}\sigma_{n_i}^2$$

and so the standard deviation σ_G of the noise on $I_G(x, y)$ is given by

$$\sigma_G^2 = \sum_{x,y}\left[\frac{\partial}{\partial I}I(x,y)G(x,y)\right]^2\sigma_I^2 = \sigma_I^2\sum_{x,y}G^2(x,y) \approx \sigma_I^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G^2(x,y)dxdy$$

The definite integral of the Gaussian can be performed by converting to polar coordinates; let

$$A = \int_{-\infty}^{\infty} e^{-ax^2}dx \Rightarrow A^2 = \int_{-\infty}^{\infty} e^{-ax^2}dx\int_{-\infty}^{\infty} e^{-ay^2}dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-a\left(x^2+y^2\right)}dxdy$$

Transform to polar coordinates $r^2 = x^2 + y^2$, $dxdy = rdrd\theta$

$$A^2 = \int_0^{\infty}\int_0^{2\pi} e^{-ar^2}rdrd\theta = 2\pi\int_0^{\infty} e^{-ar^2}rdr = \pi\int_0^{\infty} e^{-ar^2}d\left(r^2\right) = \pi\left[\frac{-1}{a}e^{-ar^2}\right]_0^{\infty} = \frac{\pi}{a}$$

so

$$A = \sqrt{\frac{\pi}{a}} \tag{4}$$

Therefore, the more general form

$$B = \int_{-\infty}^{\infty} ke^{\frac{-(x+b)^2}{c^2}}dx$$

can be written, putting $y = x + b$, as

$$B = k \int_{-\infty}^{\infty} e^{\frac{-y^2}{c^2}} dx$$

or, putting $z = y/|c|$, as

$$B = k|c| \int_{-\infty}^{\infty} e^{-z^2} dz$$

Using Eq. 4 gives

$$B = \int_{-\infty}^{\infty} k e^{\frac{-(x+b)^2}{c^2}} dx = k|c|\sqrt{\pi}$$

Returning to the original problem,

$$G^2(x,y) = \left(\frac{1}{2\pi\sigma_k^2}\right)^2 \left(exp\left[\frac{-(x-mu_x)^2-\left(y-\mu_y\right)^2}{2\sigma_k^2}\right]\right)^2$$

and, since $(e^a)^2 = e^{2a}$

$$G^2(x,y) = \frac{1}{4\pi^2\sigma_k^4} exp\left[\frac{-\left(x-mu_x\right)^2-\left(y-\mu_y\right)^2}{\sigma_k^2}\right]$$

so

$$\sigma_G^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G^2(x,y)dxdy = \frac{1}{4\pi\sigma_k^2}\sigma_I^2$$

Note that smoothing will introduce correlations between neighbouring voxels over a range dependent on the standard deviation of the smoothing kernel. Therefore, the standard deviation of the noise after smoothing will no longer be a reliable indication of noise-induced intensity differences between neighbouring voxels. However, in the present case the result was used as an overall scaling for the cost function, and so remained valid.

Additional files

Additional file 1: Semi-automatic landmark point annotation for geometric morphometrics: parameter optimisation. This file provides a full description of, and results from, the experiments performed to optimise the free parameters of the automatic landmark point annotation algorithm.

Additional file 2: The TINA geometric morphometrics tool. This file is the manual for the TINA Geometric Morphometrics Tool, which includes the TINA Manual Landmarking Tool and the automatic landmark point annotation software described in this paper. It contains detailed information on the download, installation and use of the software. Updates are available at http://www.tina-vision.net/docs/memos.php, file 2010-007.

Additional file 3: Tina 5.0 user's guide. The file gives additional information on the TINA software in general, including features not used for the landmarking procedure.Updates are available at http://www.tina-vision.net/docs/memos.php, file 2005-002.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PAB developed and tested the manual and automatic landmark localisation software and drafted the manuscript. ACS provided all data sets and corresponding manual landmark annotations, and provided specifications and feedback on software requirements and performance both before and during development. HR developed a preliminary version of the image registration functions used in automatic landmark annotation. DT and NAT initiated the project and provided technical supervision, respectively. All authors read and approved the final manuscript.

Acknowledgements

The project was funded by the Max Planck Society, Project "Automatic Identification of 3D Landmarks in Micro-CT Mouse Skull Data".

Author details

[1]Centre for Imaging Sciences, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK. [2]Department for Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany.

Received: 4 December 2013 Accepted: 29 July 2014
Published: 27 August 2014

References

1. Bookstein F: *Morphometric Tools For Landmark Data*. Cambridge: Cambridge University Press; 1991.
2. Kendall DG: **Shape manifolds, procrustean metrics, and complex projective spaces.** *Bull Lond Math Soc* 1984, **16**(2):81–121.
3. Boell L, Tautz D: **Micro-evolutionary divergence patterns of mandible shapes in wild house mouse (Mus musculus) populations.** *BMC Evol Biol* 2011, **11:**306.
4. Cardini A, Elton S: **Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons.** *Biol J Linn Soc* 2008, **93:**813–834.
5. von Cramon-Taubadel N: **Global human mandibular variation reflects differences in agricultural and hunter-gatherer subsistance strategies.** *PNAS* 2011, **108**(49):19546–19551.
6. Meloro C, O'Higgins P: **Ecological adaptations of mandibular form in fissiped carnivora.** *J Mamm Evol* 2011, **18:**185–200.
7. Breuker CJ, Patterson JS, Klingenberg CP: **A single basis for developmental buffering of Drosophila wing shape.** *PLoS ONE* 2006, **1:**e7.
8. Schunke AC, Bromiley PA, Tautz D, Thacker NA: **TINA manual landmarking tool: software for the precise digitization of 3D landmarks.** *Front Zool* 2012, **9**(6):1-5.
9. Lacroute P, Levoy M: **Fast volume rendering using a shear-warp factorization of the viewing transform.** In *Proc. SIGGRAPH '94, July 24–29 Orlando, Florida*. New York USA: ACM; 1994:451–458.
10. Lacroute P, Levoy M: **The Volpack Volume Rendering Library.** [http://graphics.stanford.edu/software/volpack/]
11. Frantz S, Rohr K, Stiehl HS: **Development and validation of a multi-step approach to improved detection of 3D point landmarks in tomographic images.** *Image Vis Comput* 2005, **233**(11):956–971.
12. Liu J, Gao W, Huang S, Nowinski WL: **A model-based, semi-global segmentation approach for automatic 3-D point landmark localization in neuroimages.** *IEEE Trans Med Imaging* 2008, **27**(8):1034–1044.
13. Hill DLG, Batchelor PG, Holden M, Hawkes DJ: **Medical image registration.** *Phys Med Biol* 2001, **46:**R1–R45.
14. Wyawahare MV, Patil PM, Abhyankar HK: **Image registration techniques: an overview.** *Int J Signal Process Image Process Pattern Recogn* 2009, **2**(3):11–28.
15. Oliveira FPM, Tavares JMRS: **Medical image registration: a review.** *Comput Methods Biomech Biomed Engin* 2012, **17**(2):1–21.
16. Mani VRS, Arivazhagan S: **Survey of medical image registration.** *J Biomed Eng Tech* 2013, **1**(2):8–25.
17. Audette MA, Ferrie FP, Peters TM: **An algorithmic overview of surface registration techniques for medical imaging.** *Med Image Anal* 2000, **4:**201–217.
18. van Kaick O, Zhang H, Hamarneh G, Cohen-Or D: **A survey on shape correspondence.** *Comput Graph Forum* 2011, **30**(6):1681–1707.

19. Tam GKL, Cheng ZQ, Lai YK, Langbein FC, Liu Y, Marshall D, Martin RR, Sun XF, Rosin PL: **Registration of 3D point clouds and meshes: a survey from rigid to nonrigid.** *IEEE Trans Vis Comput Graph* 2013, **19**(7):1199–1217.
20. Sotiras A, Davatzikos C, Paragios N: **Deformable medical image registration: a survey.** *IEEE Trans Med Imaging* 2013, **32**(7):1153–1190.
21. Lau YH, Braun M, Hutton BF: **Non-rigid image registration using a median-filtered coarse-to-fine displacement field and a symmetric correlation ratio.** *Phys Med Biol* 2001, **46:**1297–1319.
22. Collignon A, Maes F, Delaere D, Vandermeulan D, Suetens P, Marchal G: **Automated multi-modality image registration based on information theory.** In *Information Processing in Medical Imaging.* Edited by Bizais Y, Barillot C, Paola RD. Dordrecht: Kulwer, Academic; 1995:263–274.
23. Viola P, Wells WM: **Alignment by maximisation of mutual information.** In *Proceedings ICCV'95*. Cambridge, MA, USA: IEEE Computer Society Press; 1995:16.
24. Viola P, Wells WM: **Alignment by maximisation of mutual information.** *Int J Comput Vis* 1997, **24**(2):137–154.
25. Studholme C, Hill DLG, Hawkes DJ: **An overlap invariant entropy measure of 3D medical image alignment.** *Pattern Recogn* 1999, **32:**71–86.
26. Malsch U, Thieke C, Huber PE, Bendl R: **An enhanced block matching algorithm for fast elastic registration in adaptive radiotherapy.** *Phys Med Biol* 2006, **51**(19):4789–4806.
27. Bookstein F: **Principal warps: thin-plate splines and the decomposition of deformations.** *IEEE Trans Pattern Anal Mach Intell* 1989, **11:**567–585.
28. Söhn M, Birkner M, Chi Y, Wang J, Yan D, Berger B, Alber M: **Model-independent, multimodality deformable image registration by local matching of anatomical features and minimization of elastic energy.** *Med Phys* 2008, **35**(3):866–878.
29. Erdt M, Steger S, Wesarg S: **Deformable registration of MR images using a hierarchical patch based approach with a normalized metric quality measure.** In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*. Barcelona, Spain: IEEE Computer Society Press; 2012:1347–1350.
30. Bromiley PA, Pokrić M, Thacker NA: **Computing covariances for mutual information coregistration.** In *Proc. MIUA'04*. UK: BMVA Press; 2004:77–80.
31. Bromiley PA, Pokrić M, Thacker NA: **Emprical evaluation of covariance estimates for mutual information coregistration.** In *Proc. MICCAI'04*. Saint-Malo, France: Springer-Verlag; 2004:607–614.
32. Rohr K, Stiehl HS, Sprengel R, Buzug TM, Weese J, Kuhn MH: **Landmark-based elastic registration using approximating thin-plate splines.** *IEEE Trans Med Imaging* 2001, **20**(6):526–534.
33. Kim M, Wu G, Shen D: **Sparse patch-guided deformation estimation for improved image registration.** In *Machine Learning in Medical Imaging, Volume 7588 of Lecture Notes in Computer Science*. Edited by Wang F, Shen D, Yan P, Suzuki K. Berlin, Heidelberg, Germany: Springer-Verlag; 2012:54–62.
34. Canny J: **A computational approach to edge detection.** *IEEE Trans Pattern Anal Mach Intell* 1986, **8**(6):679–698.
35. Cootes TF, Taylor CJ: **Active shape models - 'smart snakes'.** In *Proc. British Machine Vision Conference (BMVC'92)*. London: Springer-Verlag; 1992:266–275.
36. Cootes TF, Taylor CJ, Cooper D, Graham J: **Active shape models - their training and application.** *Comput Vis Image Understand* 1995, **61:**38–59.
37. Cootes TF, Edwards GJ, Taylor CJ: **Active appearance models.** *IEEE Trans Pattern Anal Mach Intell* 2001, **23:**681–685.
38. Cristinacce D, Cootes TF: **Automatic feature localisation with constrained local models.** *J Comput Vis* 2005, **61:**55–79.
39. Felzenswalb P, Hutenocher D: **Pictorial structures for object recognition.** *Int J Comput Vis* 2005, **61:**55–79.
40. Heimann T, Meinzer H: **Statistical shape models for 3D medical image segmentation: a review.** *Med Image Anal* 2009, **13**(4):543–563.
41. Palaniswamy S, Thacker NA, Klingenberg CP: **Automatic identification of morphometric landmarks in digital images.** In *Proc. BMVC'07, 10–13 September, Warwick, U.K.* UK: BMVA Press; 2007:112.
42. Palaniswamy S, Thacker NA, Klingenberg CP: **Automated landmark extraction in digital images - performance evaluation.** In *ProcVIE'08, July 19 - Aug 1, Xi'an, China*. UK: IET; 2008.
43. Ballard DH: **Generalizing the hough transform to detect arbitrary shapes.** *Pattern Recogn* 1981, **13:**111–122.
44. Gall J, Lempitsky V: **Class-specific hough forests for object detection.** In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*. New York USA: IEEE Computer Society Press; 2009:1022–1029.
45. Cootes TF, Ionita MC, Lindner C, Sauer P: **Robust and accurate shape model fitting using random forest regression voting.** In *Proc. ECCV'12*; 2012:278–291.
46. Breiman L: **Random forests.** *Mach Learn* 2001, **45:**5–32.
47. Nelder JA, Meade R: **A simplex method for function minimisation.** *Comput J* 1965, **7:**308–313.
48. Ragheb H, Thacker NA: **Quantitative localisation of manually defined landmarks.** In *Proc. MIUA'11, 14–15 July, London, U.K.* UK: BMVA Press; 2011:221–225.
49. Lane RA, Thacker NA, Seed NL: **Stretch-correlation as a real-time alternative to feature-based stereo matching algorithms.** *Image Vis Comput* 1994, **12**(4):203–212.
50. Olsen SI: **Estimation of noise in images: an evaluation.** *CVGIP: Graph Models Image Process* 1993, **55:**319–323.
51. Condon JJ: **Errors in elliptical Gaussian fits.** *Publ Astron Soc Pacs* 1997, **109**(732):166–172.
52. Valstar MF, Martinez B, Binefa X, Pantic M: **Facial point detection using boosted regression and graph models.** In *Proc. CVPR*. New York USA: IEEE Computer Society Press; 2010:2729–2736.
53. Lindner C, Thiagarajah S, Wilkinson JM, arcOGEN Consortium T, Wallis GA, Cootes TF: **Fully automatic segmentation of the proximal femur using random forest regression voting.** *IEEE Trans Med Imag* 2013, **32**(8):1462–1472.
54. Ragheb H, Thacker NA, Bromiley PA, Tautz D, Schunke AC: **Quantitative shape analysis with weighted covariance estimates for increased statistical efficiency.** *Front Zool* 2013, **10**(16):1–23.
55. Barlow R: *Statistics: A Guide to the use of Statistical Methods in the Physical Sciences, 1st edition*. Chichester: John Wiley and Sons; 1989.

doi:10.1186/s12983-014-0061-1
Cite this article as: Bromiley *et al.*: Semi-automatic landmark point annotation for geometric morphometrics. *Frontiers in Zoology* 2014 **11**:61.