

Semi-automatic landmark point annotation for geometric morphometrics: parameter optimisation

P.A. Bromiley¹, A.C. Schunke², H. Ragheb¹,
N.A. Thacker¹ and D. Tautz²

1. Imaging Science and Biomedical Engineering, School of Cancer and Imaging Sciences, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK.
2. Max Planck Institute for Evolutionary Biology, Department for Evolutionary Genetics, August-Thienemann-Str. 2, 24306 Plön, Germany.
`paul.bromiley@manchester.ac.uk`

The semi-automatic landmark point annotation software included in the TINA Geometric Morphometrics toolkit (www.tina-vision.net) featured a number of free parameters, such as image patch and smoothing kernel sizes for each stage of the patch-based registration, and additional smoothing kernel sizes for the global registration and the array-based voting process. The optimal values for these parameters were expected to vary depending on the size of image features, which was itself a function of the framing of the sample within the CT volume and the resolution of the images. Therefore, a set of optimisation experiments were performed in order to find approximately optimal values. This process constituted a search in an 8-dimensional parameter space, which would have been intractable in terms of processor time requirements if the full inter-dependency of the parameters were considered; therefore, a simpler approach was adopted. Initial, approximate values for each parameter were obtained from experiments performed during the development of the registration procedure. The values used were a global registration smoothing kernel size of 5 voxels, local registration image patch sizes of 40, 20 and 10 voxels and smoothing kernel sizes of 20, 10 and 5 voxels for the first, second and third stages of local registration respectively, and a smoothing kernel size of 10 voxels in the array-based voting procedure. A pattern search was then performed around these initial values, optimising each independently. Throughout this document, all Gaussian kernels were isotropic in voxel coordinates and all kernel sizes are given as standard deviations in voxels; all image patches were square in voxel coordinates and all patch sizes are given as the apothem i.e. a patch of size n centred on (c_x, c_y) spans coordinates $(c_x - n, c_y - n)$ to $(c_x + n, c_y + n)$ in voxels.

First, a set of experiments was performed in order to evaluate the dependence of the algorithm on the number of local registration stages. The default parameters given above were used, together with image patch and smoothing kernel sizes of 80 and 40 for a fourth local registration stage; these parameters were extrapolated from those for the other stages. Experiments were run using only the last stage of local registration (i.e. with a patch size of 5), the last two stages etc. A set of experiments was then performed in which the value of each parameter was varied independently around the initial value, whilst the other parameters were held constant. This constituted the first stage of a pattern search in the parameter space. The data used in these experiments consisted of eight micro-CT image volumes of *Mus* skulls, with expert manual annotation of 40 mandible landmarks for each (see main document for details). Leave-one-out experiments were performed for each set of parameter values, using seven of the image volumes to construct the database and automatically locating the landmark points on the left-out image volume, repeating for all eight volumes.

The results of the experiments on the number of local registration stages are shown in Figure 1. The results of the parameter optimisation experiments are shown in Figures 2 to 5. In each case, the accuracy of automatic point localisation is given as the Euclidean distance in voxels between the automatically and manually identified locations for that point. The data are presented as box-and whisker plots for points that passed the outlier test from all eight image volumes, showing the maximum, 75th percentile, median, 25th percentile and minimum error. The figures also show the percentage of points that passed the outlier test.

Interpretation of the results was complicated by the fact that the optimal performance should balance a high proportion of points passing the outlier test with a low error on those points. Since these were non-commensurate objectives, no combined cost function could be defined; this fact prevented an automatic optimisation of the parameters using a standard algorithm. The experiments on the number of local registration stages showed that median point localisation accuracy generally improved with more stages. However, the number of points passing the outlier test peaked at three stages; using a fourth stage resulted in a minimal improvement in point localisation accuracy but a significant reduction in the number of points passing the test. Therefore, three local registration stages were used in all further experiments. The existence of this optimum in the number of local registration

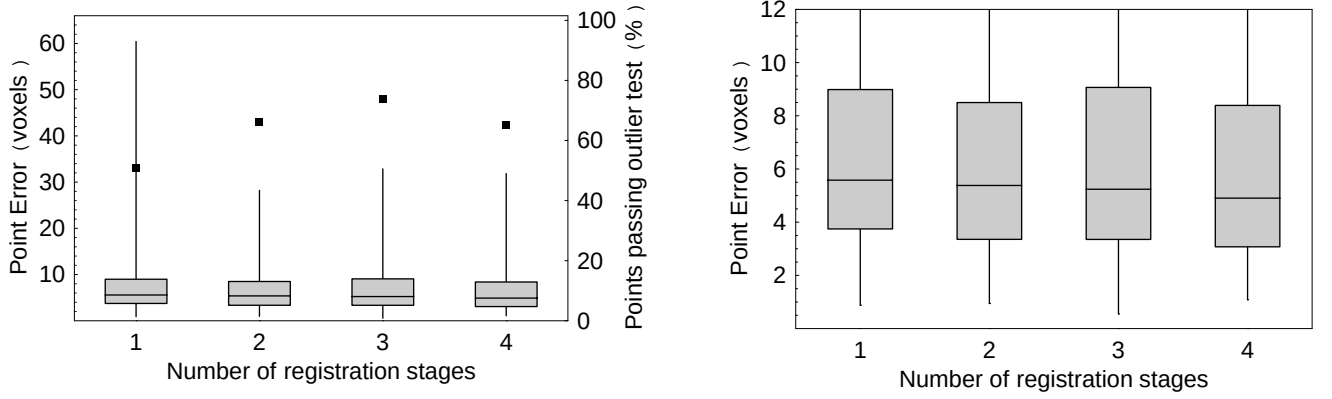


Figure 1: Optimisation of the number of local registration stages. The box-and-whisker plots show the error in voxels for the automatically located points that passed the outlier test (read against the left-hand scale); the black points show the percentage of the points that passed the outlier test (read against the right-hand scale). The right-hand plot shows a detail from the left-hand plot.

stages can be explained by the competing demands on the local registration. Utilising more image information by using a larger patch size will result in a more stable cost function; however, it will also increase the probability of systematic errors due to shape variation between the images.

The parameter optimisation results showed that the automatic point localisation procedure was relatively insensitive to the parameter values across the majority of the ranges tested, indicating that no further iterations of pattern search were required in order to produce approximately optimal values. This also indicated that three stages of patch-based registration were sufficient to attain the maximum achievable accuracy given the information available from the data, since low sensitivity to the patch-based registration parameters indicated a degree of redundancy between the various stages. Figure 5 shows a clear optimum for the global smoothing kernel size of 5 voxels; both lower and higher values resulted in a smaller proportion of the points passing the outlier test, whilst the accuracy of those points varied little. Figure 2 shows that there is little statistically significant variation in point localisation accuracy with the parameters of the first stage of patch based registration, although smaller smoothing kernel sizes resulted in a significant reduction in the percentage of points passing the outlier test. Therefore, values of 40 voxels for the patch and 20 voxels for the smoothing kernel sizes respectively were chosen, as these produced the highest proportion of points passing the outlier test. On the same basis, values of 15 voxels for the patch and 10 voxels for the smoothing kernel sizes respectively were chosen for the second stage of patch-based registration (see Figure 3). Values of 10 voxels for the patch and 5 voxels for the smoothing kernel sizes respectively were chosen for the third stage of patch-based registration (see Figure 4), since these resulted in a high proportion of points passing the outlier test, with slightly lower errors than were achieved with patch sizes of 15. Finally, a value of 15 voxels was selected for the smoothing kernel size in the array-based voting procedure, since this resulted in a high proportion of points passing the outlier test, with a slightly lower error than was achieved with a size of 20.

The outlier testing stage of the algorithm operated by comparing a threshold value to the distances between the main mode in the smoothed 3D voting array and the entries into that array with the lowest χ^2 . Optimal values for both the outlier test threshold and the number of comparisons were obtained empirically. This required calculation of the true and false positive and negative rates for the outlier test which, in turn, required an additional threshold on acceptable point localisation error. This error threshold was inevitably arbitrary to some degree. However, since two sets of manual annotations were available for the 12 consomic *Mus musculus* specimens used in the evaluation, it was possible to evaluate the manual annotation error (see main document). The accuracy of automatic annotations was estimated by comparing their locations to corresponding manual annotations; the result contained contributions from the errors on both the manual and automatic annotations. Therefore, setting the error threshold to the mean or median of the manual repeatability would not produce meaningful results. Instead, the error threshold was set to the mean of the worst outliers in the manual annotations for each of the twelve specimens. Therefore, it defined an incorrect automatic localisation as one that had an error greater than that which would be seen in typical manual annotations. The chosen threshold was 30 voxels, corresponding to 1.05mm.

The outlier test thresholds were optimised using the dataset of 40 mandible landmarks on 8 *Mus musculus* specimens (see main document for details). Figure 6 shows an ROC curve of the true and false positive rates as the outlier

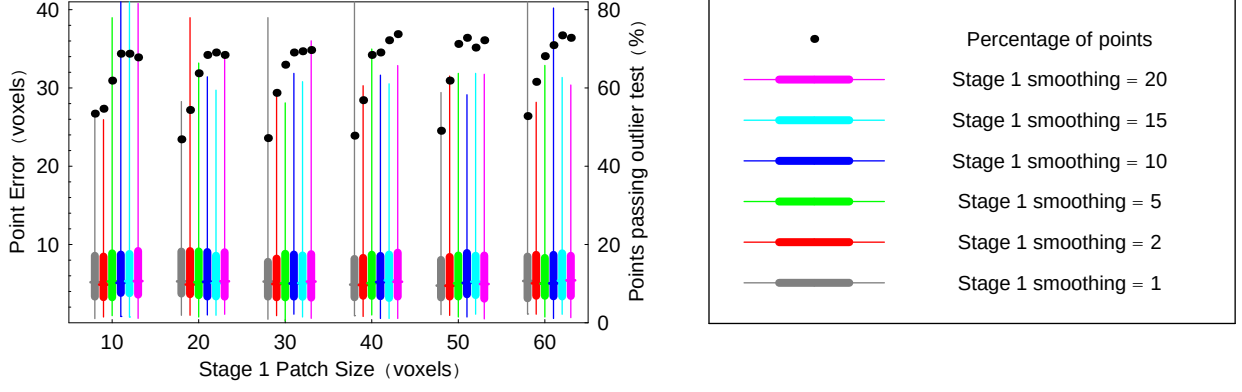


Figure 2: Optimisation of the first stage registration parameters. The box-and-whisker plots show the error in voxels for the automatically located points that passed the outlier test (read against the left-hand scale); the black points show the percentage of the points that passed the outlier test (read against the right-hand scale).

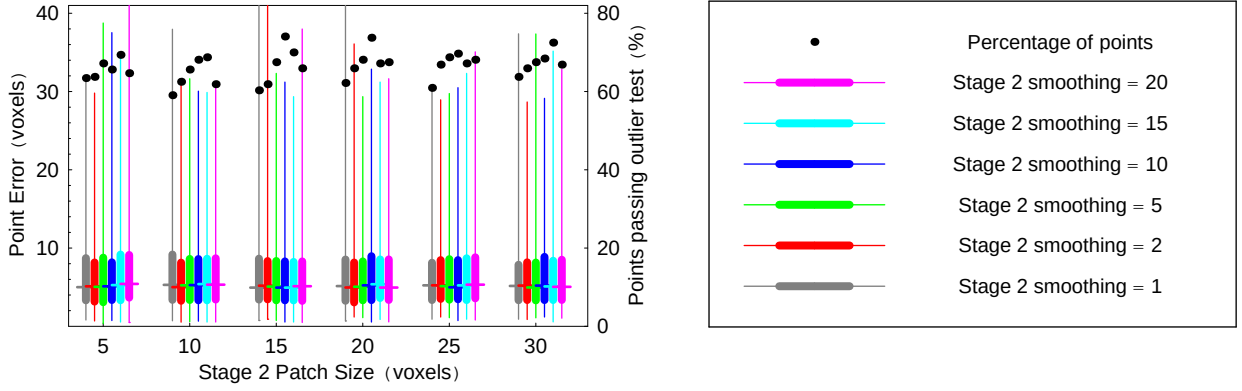


Figure 3: Optimisation of the second stage registration parameters. The box-and-whisker plots show the error in voxels for the automatically located points that passed the outlier test (read against the left-hand scale); the black points show the percentage of the points that passed the outlier test (read against the right-hand scale).

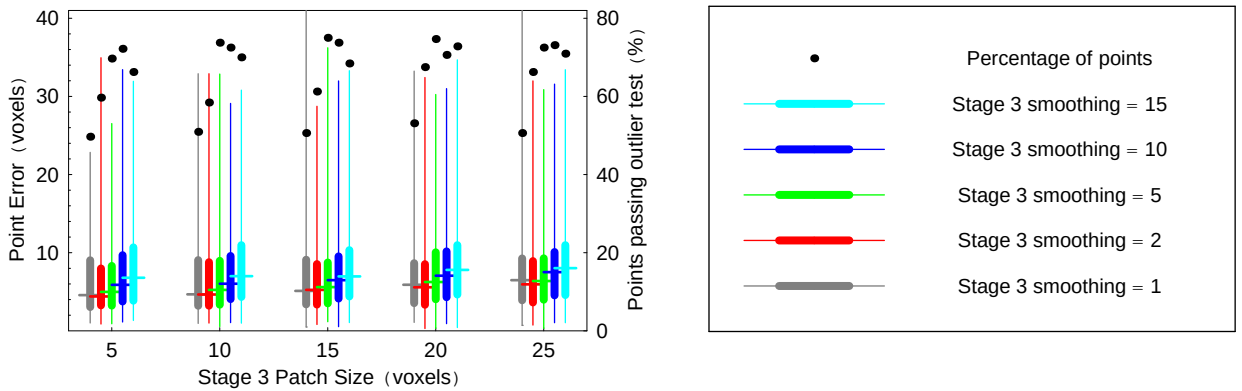


Figure 4: Optimisation of the third stage registration parameters. The box-and-whisker plots show the error in voxels for the automatically located points that passed the outlier test (read against the left-hand scale); the black points show the percentage of the points that passed the outlier test (read against the right-hand scale).

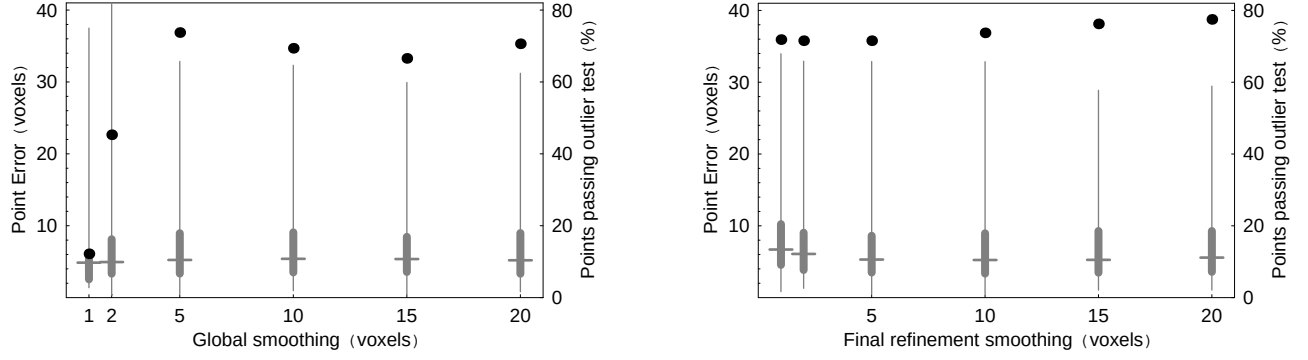


Figure 5: Optimisation of the smoothing parameters for the global registration (left) and the array-based voting (right). The box-and-whisker plots show the error in voxels for the automatically located points that passed the outlier test (read against the left-hand scale); the black points show the percentage of the points that passed the outlier test (read against the right-hand scale).

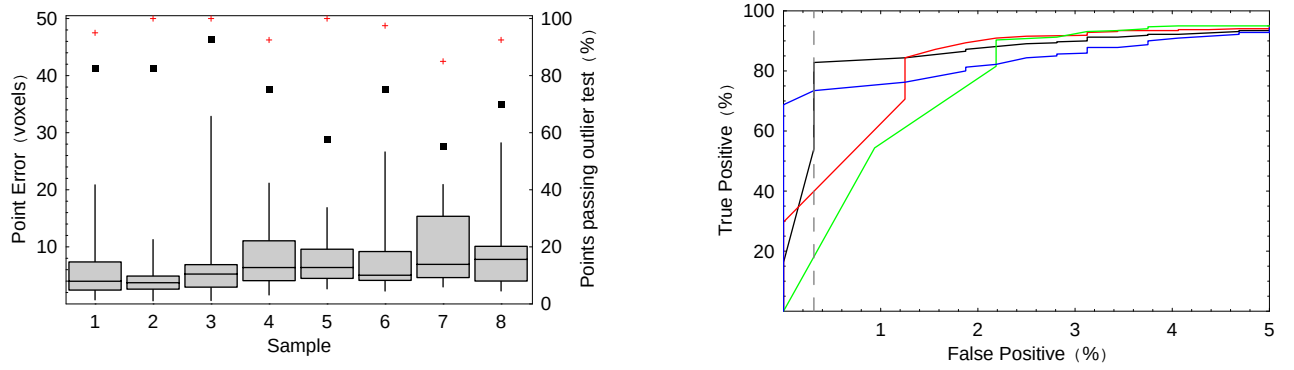


Figure 6: (Left) Box-and-whisker plots of the errors for automatic localisation of 40 mandible landmarks on the 8 *Mus* specimens (read against the left-hand scale); only points passing the outlier test were included. The black and red points show the percentage of points passing the outlier test and the percentage with errors lower than the threshold, respectively (read against the right-hand scale). (Right) ROC curve of the true and false positive rates of points passing the outlier test when comparing to 1 (green), 2 (red), 3 (black) and 4 (blue) points. The dashed line shows the optimised operating point for the outlier test.

test thresholds were varied; in order to limit the size of the parameter space, the same threshold was applied in all comparisons of the outlier test. A true positive was defined as a point that passed the outlier test and had an error lower than the error threshold of 30 voxels; conversely, a false positive was a point that passed the outlier test and had an error larger than the error threshold. The number of database points included in the comparison was varied from 1 to 4. The operating point of the algorithm was then chosen to achieve a false positive rate lower than 0.5%, reflecting the number of extreme outliers seen in manual annotations (see main document). The optimal performance i.e. the highest true positive rate for a false positive rate below 0.5%, was achieved with thresholds of 20 voxels and three database points included in the comparison.

Figure 6 also shows the performance of the automatic landmark annotation algorithm on the dataset of 40 mandible landmarks on 8 *Mus musculus* specimens with the optimised parameters. The median point localisation error for points passing outlier detection, across all 40 points in all 8 datasets, was 5.2 voxels; it should be noted that this included a contribution from the manual annotation error. The mean of the worst outliers, again in points passing the outlier test across all 8 datasets, was 22.3 voxels. The percentage of points passing the outlier test was 73.75%, and 95.3% of the points had an error lower than the threshold of 30 voxels. Therefore, the software indicated to the user that 16.25% of the points should be manually checked; however, only 4.7% of these points would require adjustment in order to achieve equivalent errors to those seen in a typical manual annotation process.