Assessing the Accuracy of Non-Rigid Registration With and Without Ground Truth

R. S. Schestowitz[1], W. R. Crum[2], V. S. Petrovic[1],
C. J. Twining[1], T. F. Cootes[1] and C. J. Taylor[1]

[1]Imaging Science and Biomedical Engineering, University of Manchester
Stopford Building, Oxford Road, Manchester M13 9PT, United Kingdom

[2]Centre for Medical Image Computing, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom

Non-rigid registration (NRR) of both pairs and groups of images has been used increasingly in recent years, as a basis for medical image analysis. Applications include structural analysis, atlas matching and change analysis. The problem is highly under-constrained and the plethora of different algorithms that have been proposed generally produce different results for a given set of images. We present two methods for assessing the performance of non-rigid registration algorithms applied to groups of images; one requires ground truth to be provided *a priori*, whereas the other does not. We compare the two approaches by systematically varying the quality of registration of a set of MR images of the brain.

The first of the proposed methods for assessing registration quality uses a generalisation of Tanimoto's spatial overlap measure. We start with a manual mark-up of each image, providing an anatomical/tissue label for each voxel, and measure the overlap of corresponding labels following registration. Each label is represented using a binary image, but after warping and interpolation into a common reference frame, based on the results of NRR, we obtain a set of fuzzy label images. These are combined in a generalised overlap score [1]:

$$\mathcal{O} = \frac{\sum_{\text{pairs},k} \sum_{\text{labels},l} \alpha_l \sum_{\text{voxels},i} MIN(A_{kli}, B_{kli})}{\sum_{\text{pairs},k} \sum_{\text{labels},l} \alpha_l \sum_{\text{voxels},i} MAX(A_{kli}, B_{kli})} \tag{1}$$

where i indexes voxels in the registered images, l indexes the label and k indexes image pairs. A_{kli} and B_{kli} represent voxel label values in a pair of registered images and are in the range [0, 1]. The $MIN()$ and $MAX()$ operators are standard results for the intersection and union of fuzzy sets. The generalised overlap measures the consistency with which each set of labels partitions the image volume. The parameter α_l affects the relative weighting of different labels. With $\alpha_l = 1$, label contributions are implicitly volume weighted with respect to one another. We have also considered the cases where α_l weights for the inverse label volume (which makes the relative weighting of different labels equal), where α_l weights for the inverse label volume squared (which gives labels of smaller volume higher weighting) and where α_l weights for a measure of label complexity (which we define arbitrarily as the mean absolute voxel intensity gradient in the label).

The second method assesses registration in terms of the quality of a generative statistical appearance model, constructed from the registered images – for all the experiments reported here, this was an active appearance model (AAM). The idea is that a correct registration produces an anatomically meaningful dense correspondence between the set of images, resulting in a better appearance model. We define model quality using two measures – generalisation and specificity. Both are measures of overlap between the distribution of original images, and a distribution of images sampled from the model. If we use the generative property of the model to synthesise a large set of images, $\{I_\alpha : \alpha = 1, \ldots m\}$, we can define Generalisation G:

$$G = \frac{1}{n} \sum_{i=1}^{n} \min_{\alpha} |I_i - I_\alpha|, \tag{2}$$

where $|\cdot|$ is a measure of distance between images, I_i is the i^{th} training image, and $\min_\alpha$ is the minimum over α (the set of *synthetic* images). That is, Generalisation is the average distance from each training image to its nearest neighbour in the synthetic image set. A good model exhibits a low value of G, indicating that the model can generate images that cover the full range of appearances present in the original image set. Similarly, we can define Specificity S:

$$S = \frac{1}{m} \sum_{\alpha=1}^{m} \min_{i} |I_i - I_\alpha|. \tag{3}$$

That is, Specificity is the average distance of each synthetic image from its nearest neighbour in the original image set. A good model exhibits a low value of S, indicating that the model only generates synthetic images that are similar to those in the original image set. The uncertainty in estimating G and S can also be computed. In our experiments we have defined $|\cdot|$ as the shuffle distance between two images. Shuffle distance is the mean of the minimum absolute difference between each pixel/voxel in one image, and the pixels/voxels in a shuffle neighbourhood of radius r around the corresponding pixel/voxel in a second image. When $r \leq 1$, this is equivalent to the mean absolute difference between corresponding

pixels/voxels, but for larger values of r the distance increases more smoothly as the misalignment of structures in the two images increases.

The overlap-based and model-based approaches were validated and compared, using a dataset consisting of 36 transaxial mid-brain slices, extracted at equivalent levels from a set of T1-weighted 3D MR scans of different subjects. Eight manually annotated anatomical labels were used as the basis for the overlap method: L/R white matter, L/R grey matter, L/R lateral ventricle, and L/R caudate. The images were brought into alignment using an NRR algorithm based on MDL optimisation [2]. A test set of different mis-registrations was then created by applying smooth pseudo-random spatial warps (based on biharmonic Clamped Plate Splines) to the registered images. Each warp was controlled by 25 randomly placed knot-points, each displaced in a random direction by a distance drawn from a Gaussian distribution whose mean controlled the average magnitude of pixel displacement over the whole image. Ten different warp instantiations were generated for each image for each of seven progressively increasing values of average pixel displacement. Registration quality was measured, for each level of registration degradation, using several variants of each of the proposed assessment methods.

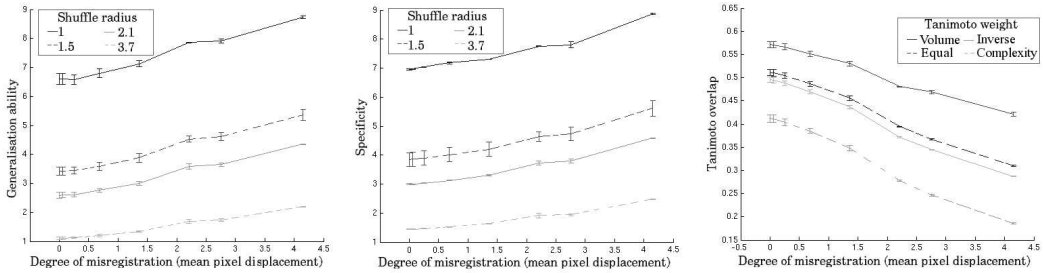

Figure 1. Behaviour of proposed metrics with increasing registration perturbation:
a) Generalisation, b) Specificity and c) Tantimoto overlap

The results of the validation experiment are shown in Figure 1. Note that $\mathcal{O}$ is expected to decrease with increasing perturbation of the registration, whilst G and S are expected to increase. All three metrics are generally well-behaved and show a monotonic response to increasing perturbation. This validates the model-based measures of registration quality, which are shown both to change monotonically with increasing perturbation of the registration and to correlate with the gold-standard approach based on manually annotated ground truth.

The results for different values of r (shuffle radius) and α_l all demonstrate monotonic behaviour with increasing perturbation, but the slopes and errors vary systematically. This affects the size of perturbation that can be detected. To make a quantitative comparison of the different methods, we define the sensitivity, as a function of perturbation as $(\frac{1}{\overline{\sigma}})\frac{m-m_0}{d}$, where m is the quality measured for a given value of displacement, m_0 is the measured quality at registration, d is the degree of deformation and $\overline{\sigma}$ is the mean error in the estimate of m over the range.

Sensitivity averaged over the range of perturbations shown in Figure 1 is plotted in Figure 2 for all the methods of assessment. This shows that the Specificity measure with shuffle radius 1.5 or 2.1 is the most sensitive of the measures studied, and that this difference is statistically significant.

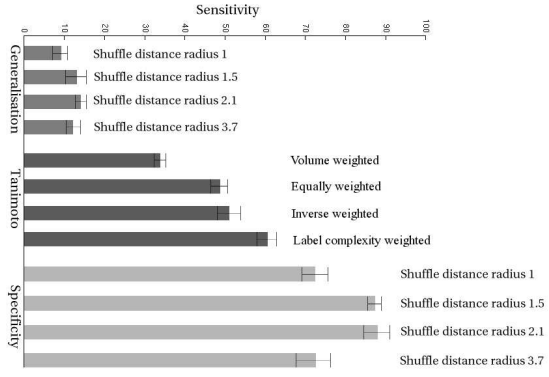

Figure 2. The sensitivity of registration assessment methods.

References

[1] W. R. Crum, O. Camara, D. Rueckert, K. Bhatia, M. Jenkinson and D. L. G. Hill. Generalised Overlap Measures for Assessment of Pairwise and Groupwise Image Registration and Segmentation. Proceedings of MICCAI 2005, LNCS 3749, pp 99-106.

[2] C. J. Twining, T. Cootes, S. Marsland, V. Petrovic, R. Schestowitz and C. J. Taylor. A Unified Information-Theoretic Approach to Groupwise Non-Rigid Registration and Model Building. Proceedings of IPMI 2005, LNCS 3565, pp 1-14.