Non-Rigid Registration Assessment Without Ground Truth

R. S. Schestowitz[a], C. J. Twining[a], T. F. Cootes[a], V. S. Petrović[a], W. R. Crum[b], and C. J. Taylor[a*]

[a]Imaging Science and Biomedical Engineering, Stopford Building,
University of Manchester, Oxford Road, Manchester M13 9PT, UK.
[b]Centre for Medical Image Computing, Department of Computer Science,
University College London,Gower Street, London WC1E 6BT, UK.

Abstract. We compare two methods for assessing the performance of groupwise non-rigid registration algorithms. The first approach, which has been described previously, utilizes a measure of overlap between ground-truth anatomical labels. The second, which is new, exploits the fact that, given a set of non-rigidly registered images, a generative statistical model of appearance can be constructed. We observe that the quality of this model depends on the quality of the registration, and define measures of model *specificity* and *generalisation* – based on comparing synthetic images sampled from the model, with those in the original image set – that can be used to assess model/registration quality. We show that both approaches detect the loss of registration accuracy as the alignment of a set of correctly registered MR images of the brain is progressively perturbed. We compare the sensitivities of the two approaches and show that, as well as requiring no ground truth, *specificity* provides the most sensitive measure of misregistration. Finally, we use *specificity* and *generalisation* to compare three NRR algorithms.

1 Introduction

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 [6]. 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 [4, 19].

Various methods have been proposed for assessing the results of NRR [9, 11, 13, 16]. Most of these require access to some form of ground truth. One approach involves the construction of artificial test data, which limits application to 'off-line' evaluation. Other methods can be applied directly to real data, but require that anatomical ground truth be provided, typically involving annotation by an expert. This makes validation expensive and prone to subjective error.

We present two methods for assessing the performance of non-rigid registration algorithms; one requires ground truth to be provided, whereas the other does not. We compare the performance of the two approaches by systematically varying the quality of registration of a set of MR images of the brain.

2 Method

The first of the proposed methods for assessing registration quality uses a generalisation of Tanimoto's spatial overlap measure [1]. 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 [5]:

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

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 β_l affects the relative weighting of different labels. With $\beta_l = 1$, label contributions are implicitly volume weighted with respect to one another. We have also considered the cases where β_l weights for the inverse label volume (which makes the relative weighting of different labels equal), where β_l weights for the inverse label volume squared (which gives labels of smaller volume higher weighting) and where β_l weights for a measure of label complexity (which we define arbitrarily as the mean absolute voxel intensity gradient in the label).

*E-mail addresses for correspondence: Roy.Schestowitz@stud.man.ac.uk; Carole.Twining, V.Petrovic, Tim.Cootes, Chris.Taylor@manchester.ac.uk; Bill.crum@cs.ucl.ac.uk

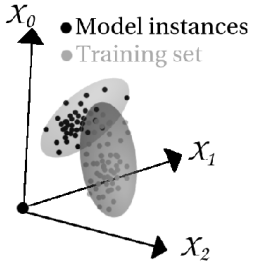

Figure 1. Training set and model in hyperspace

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) [3]. The idea is that a correct registration produces an anatomically meaningful dense correspondence between the images, resulting in a better appearance model of the anatomy. 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, as illustrated in Figure 1. 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 as:

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

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. Given a sufficiently large synthetic set, even registered image with differing brightness levels will be paired with a nearby match. Similarly, we can define Specificity S as:

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

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, as illustrated for a single pixel in Figure 2. Shuffle distance images are formed by taking 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 effect on the pixel-by-pixel contribution to the shuffle distance image as r is increased is illustrated in Figure 3.

3 Experimental Validation and Application

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 [18]. The resulting appearance model is shown in Figure 6. A test set of different mis-registrations was then created by applying smooth pseudo-random spatial warps to the registered images. These warps were based on biharmonic Clamped Plate Splines. 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 at each

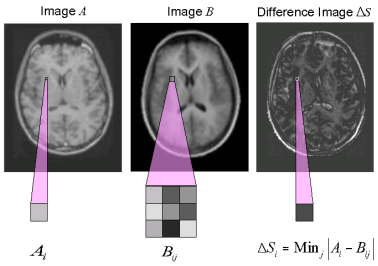

Figure 2. The calculation of a shuffle difference image

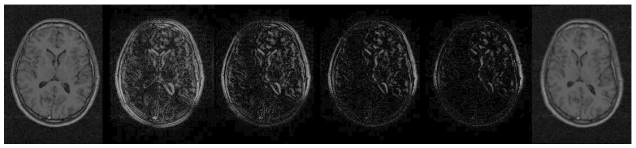

Figure 3. Shuffle difference images between the images on the extreme left and right for $r = 1$ (abs. diff.), 1.5, 2.1 & 3.7 from left to right.

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.

To illustrate practical application of the method, we use the brain images described above, and compare the results of three different registration algorithms: a pairwise registration of each training set image to a fixed reference image chosen from the training set, and two MDL groupwise methods, as described above, one with no explicit constraints on the spatial deformations during the registration process (Groupwise 1) and a second which uses a statistical shape model to constrain the allowed spatial deformations between the images during registration (Groupwise 2) [18].

4 Results

The results of the validation experiment are shown in Figure 4. 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 β_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 $(\frac{1}{\overline{\sigma}})\frac{M-M_0}{d}$, where M is the quality measured for a given degree of deformation d, M_0 is the measured quality at registration (no deformation) and $\overline{\sigma}$ is the mean error in the estimate of M over the range.

Sensitivities of the different methods, averaged over the range of perturbations shown in Figure 4, are summarised in Figure 5 for all the methods of assessment. Since sensitivity across the whole range is desired, this average shows that the Specificity measure with shuffle radius 1.5 or 2.1 is the most sensitive of the measures studied, and that this advantage is statistically significant. Exceeding this shuffle radius may lead to performance degradation, as deformations will be obscured by the shuffling.

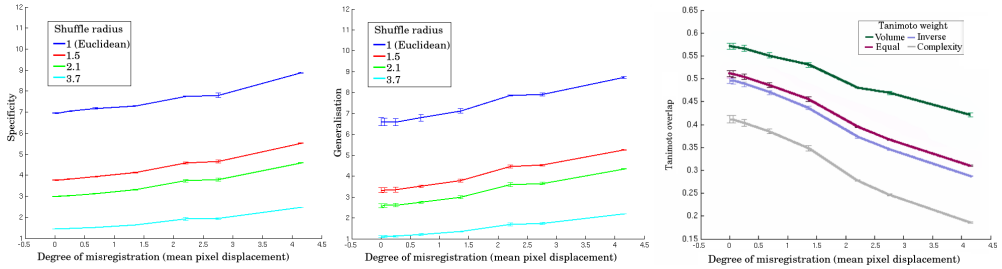

Figure 4. From Left: Specificity (S), Generalisation (G) & Tanimoto overlap ($\mathcal{O}$) as a function of misregistration.

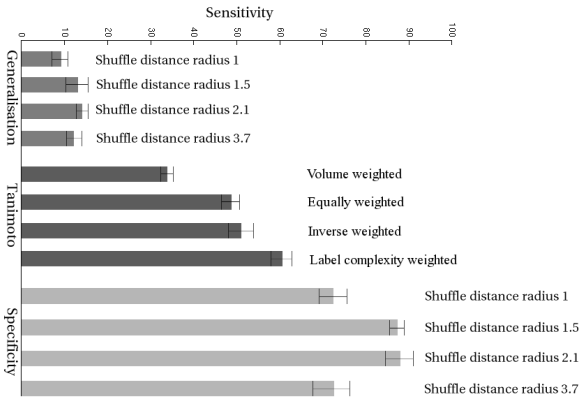

Figure 5. The sensitivities of the different registration assessment methods and their standard errors.

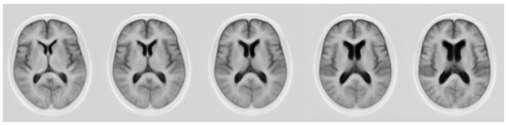

Figure 6. Appearance model constructed from groupwise registered images. First mode of variation is shown, ± 2.5 standard deviations.

The results for different registration algorithms are shown in Figure 7. The specificity obtained for the two groupwise methods is significantly better than that obtained using the pairwise approach, implying better registration, but it is not possible to distinguish between the two groupwise methods. By applying the same NRR algorithms to an annotated dataset, it becomes evident that generalised overlap measures agree with this assessment, for all possible assignments to β_l. As might be expected from the sensitivity results presented above, it is not possible to distinguish between any of the methods using generalisation.

5 Conclusions

We have introduced a model-based approach to assessing the accuracy of non-rigid registration, without the need for ground-truth. We have also described validation experiments where we progressively perturbed the initially good registration of a set of images, and found a monotonic relationship between our model-based measures and the degree of perturbation. We found that this behaviour was qualitatively identical to that obtained using a 'gold standard' method of assessment, based on the overlap of ground-truth anatomical labels associated with the images. A quantitative comparison of the two approaches demonstrated that one of the model-based measures, *specificity*, provides a more sensitive measure of misregistration than the overlap-based approach. This is not as surprising as it might seem at first sight, since the model-based approach uses the full intensity information in the registered images, whereas the overlap-based approach uses a more impoverished representation of image structure. We tested different variants of the two approaches, and found that the model-based approach worked best when shuffle distance was used to measure

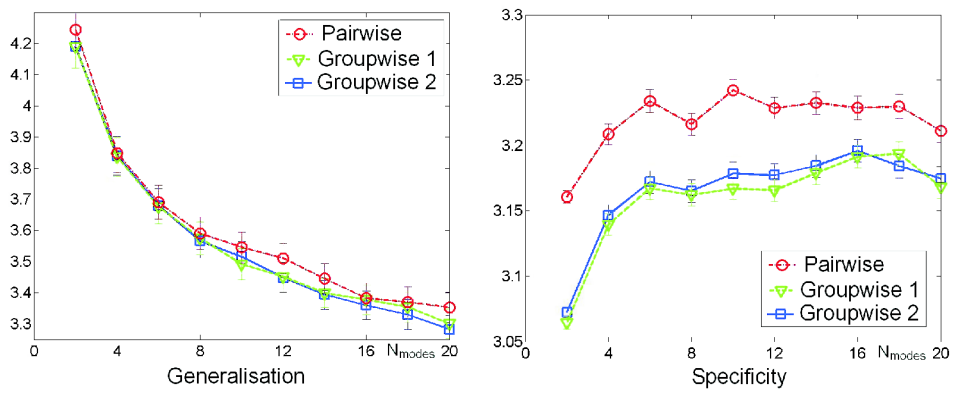

Figure 7. Generalisation and Specificity of the three registration methods as a function of the number of modes included in the appearance model.

separation in image space, whilst the overlap-based approach worked best when a label complexity weighting was applied. We also applied the model based approach to compare three different registration methods, and showed that a groupwise registration approach gave better results than a pairwise approach.

These results are important, because they suggest that the performance of NRR algorithms can be compared objectively, using just the registered images they produce, and that the quality of registration can be assessed in routine applications of NRR, without the need for any additional information. It is important to note that our approach does not depend on the specifics of the registration method used, or on the particular form of generative model constructed from the registered data. It can be applied to the results of registration, whatever the NRR algorithm used, and different forms of generative model could easily be substituted. We demonstrated this in recent experiments, which were omitted due to the limited scope of this paper.

Acknowledgements

The authors would like to thank David Kennedy of the Centre for Morphometric Analysis at MGH for segmented brain data. The work was supported by the EPSRC/MRC-funded Medical Image and Signal IRC (GR/N14248/01), Integrated Brain Image Modelling (EPSRC GR/S82503/01) and Modelling, Understanding and Predicting Structural Brain Change (EPSRC GR/S48844/01).

References

1. M. Beauchemin and K. P. B. Thomson. The evaluation of segmentation results and the overlapping area matrix. *International Journal of Remote Sensing*, 18(18):3895-3899, 1997.
2. T. F. Cootes, C. Beeston, G. J. Edwards, and C. J. Taylor. A unified framework for atlas matching using active appearance models. In *Information Processing in Medical Imaging*, 1613:322-333, 1999.
3. T.F. Cootes, G.J. Edwards and C.J.Taylor. Active appearance models. In *European Conference on Computer Vision*, 2:484-498, 1998.
4. T. F. Cootes, S. Marsland, C. J. Twining, K. Smith, and C. J. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *European Conference on Computer Vision*, 2034:316-27, 2004.
5. 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. In *Proceedings of MICCAI*, 3749:99-106, 2005.
6. W. R. Crum, T. Hartkens, and D. L. G. Hill. Non-rigid image registration: theory and practice. *British Journal of Radiology*, 77:140-153, 2004.
7. R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor. A minimum description length approach to statistical shape modeling. *IEEE Transactions on Medical Imaging*, 21(5):525-537, 2002.
8. G. J. Edwards, T. F. Cootes, and C. J. Taylor. Face recognition using active appearance models. In *European Conference on Computer Vision*, 2:581-595, 1998.
9. J. M. Fitzpatrick and J. B. West. The distribution of target registration error in rigid-body point-based registration. *IEEE Transaction Medical Imaging*, 20:917-27, 2001.
10. A. F. Frangi, D. Rueckert, J. A. Schnabel, and W. J. Niessen. Automatic construction of multiple-object three-dimensional statistical shape models: application to cardiac modelling. *IEEE Transactions on Medical Imaging*, 21:1151-66, 2002.
11. P. Hellier, C. Barillot, I. Corouge, B. Giraud, G. Le Goualher, L. Collins, A. Evans, G. Malandain, and N. Ayache. Retrospective evaluation of inter-subject brain registration. In *Medical Image Computing and Computer-Assisted Intervention*, 2208:258-265, 2001.
12. K. N. Kutulakos. Approximate N-view stereo. In *European Conference on Computer Vision*, 1:67-83, 2000.
13. P. Rogelj, S. Kovacic, and J. C. Gee. Validation of a nonrigid registration algorithm for multimodal data. *Medical Imaging*, volume 4684, 2002.
14. D. Rueckert, A. F. Frangi, and J. A. Schnabel. Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration. *IEEE Transactions on Medical Imaging*, 22(8)1014-1025, 2003.
15. D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, D. J. Hawkes. Non-rigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712-721, 1999.
16. J. A. Schnabel, C. Tanner, A. Castellano-Smith, M. O. Leach, C. Hayes, A. Degenhard, R Hose, D. L. G. Hill, and D. J. Hawkes. Validation of non-rigid registration using finite element methods. In *Information Processing in Medical Imaging*, 2082:344-357, 2001.
17. M. B. Stegmann, B. K. Ersboll, and R. Larsen. FAME - a flexible appearance modeling environment. *IEEE Transactions on Medical Imaging*, 22(10):1319-1331, 2003.
18. C. J. Twining, T.F. Cootes, S. Marsland, S. V. Petrovic, R. S. Schestowitz, and C. J. Taylor. A unified information-theoretic approach to groupwise non-rigid registration and model building. In *Information Processing in Medical Imaging*, 3565:1-14, 2005.
19. B. Zitova and J. Flusser. Image registration methods: a survey. *Image Vision Computing*, 21:977-1000, 2003.