ASSESSING THE ACCURACY OF NON-RIGID REGISTRATION WITH AND WITHOUT GROUND TRUTH

R. Schestowitz, C. J. Twining, T. Cootes, V. Petrovic, and C. J. Taylor

Imaging Science and Biomedical Engineering, Stopford Building, University of Manchester, Oxford Road, Manchester M13 9PT, UK.

W. R. Crum

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. One approach, which has been described previously, utilizes a measure of overlap between data labels. Our new approach exploits the fact that, given a set of non-rigidly registered images, a generative statistical appearance model can be constructed. We observe that the quality of the model depends on the quality of the registration, and can be evaluated by comparing synthetic images sampled from the model with the original image set. We derive indices of model specificity and generalisation that can be used to assess model/registration quality. We show that both approaches detect the loss of registration as a set of correctly registered MR images of the brain is progressively perturbed. We compare the sensitivities of the different methods and show that, as well as requiring no ground truth, our new specificity measure provides the most sensitive approach to detecting misregistration.

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 [5]. 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, 16, 15]. 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 *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.

2. METHOD

The first of the proposed methods for assessing registration quality uses a generalisation of Tanimoto's spatial overlap measure [7]. 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]:

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

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) [3]. The

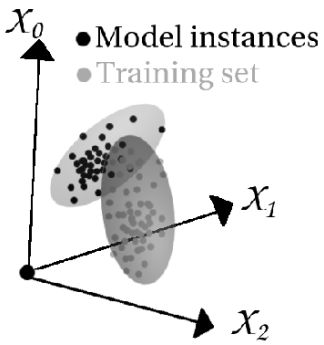

Fig. 1. Training set and model in hyperspace

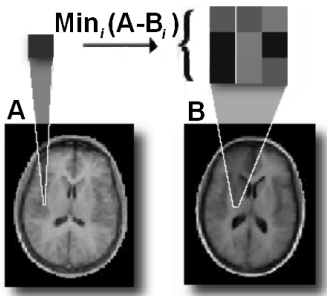

Fig. 2. The calculation of a shuffle difference image

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 [18]. 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:

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

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, as illustrated in Figure 2. 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 effect on the pixel-by-pixel contribution to shuffle distance as r is increased is illustrated in Figure 3.

3. EXPERIMENTAL VALIDATION

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]. 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.

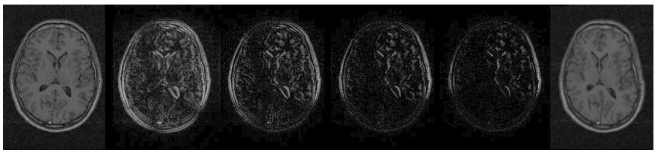

Fig. 3. Shuffle distance evaluation: **Left:** one image, **Right:** another image, **Centre, from left to right:** images showing contributions to shuffle distance, for $r = 0$ (abs. diff.), 1.5, 2.1 & 3.7 respectively.

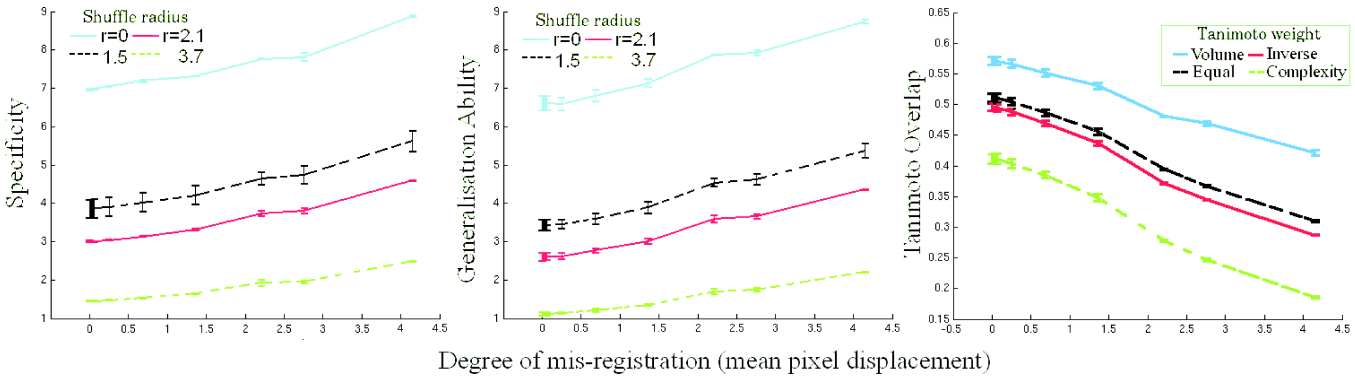

Fig. 4. **From Left:** Specificity, Generalisation & Tanimoto overlap as a function of registration perturbation.

4. RESULTS

The results of the validation experiment are shown in Figure 4. Note that 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.

These 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 of the range of perturbations shown in Figure 4 is plotted in Figure 6 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.

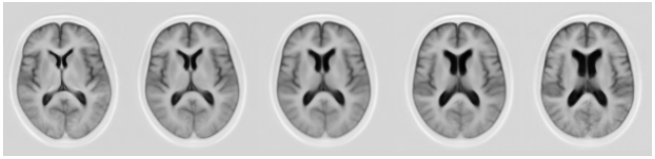

Fig. 5. Appearance model which was built automatically by group-wise registration. First mode is shown, ± 2.5 standard deviations.

5. CONCLUSIONS

We have introduced a model-based approach to assessing the accuracy of non-rigid registration, without the need for ground truth. The validation experiments, based on perturbing correspondences obtained using ground truth, show that we are able to detect increasing mis-registration using just the registered image data. The results obtained for different sizes of shuffle neighbourhood show that the use of shuffle distance rather than Euclidean or absolute distance improves the range of mis-registration over which we can detect significant changes in registration accuracy and improves the sensitivity of the approach.

More broadly, registration performance can be evaluated reliably both in the cases when ground truth information is available and when it is not. In particular, the methods based on generative statistical model evaluation are shown to be in agreement with the ground truth expressed through the true image region overlap metric based on the Tantimoto formulation. Proposed metrics are also shown to have sufficient sensitivity to detect very subtle changes in registration perfor-

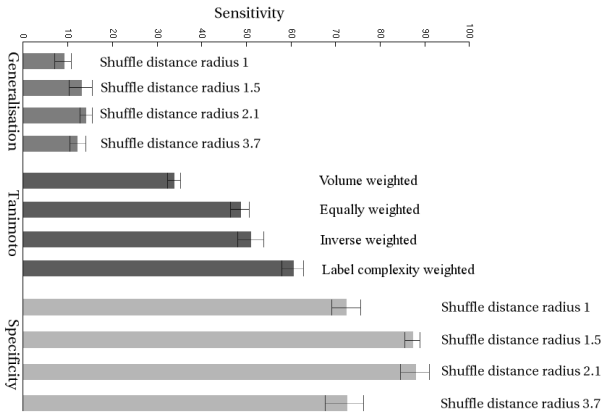

Fig. 6. The sensitivity of the different registration assessment methods.

mance, on the level of perturbations measured in fractions of a pixel.

We believe that this represents an important advance in the assessment of NRR, because it establishes an entirely objective basis for evaluating the reliability of NRR-based experiments, and for comparing the performance of different methods of NRR. The fact that no ground truth data is required means that the method can be applied routinely. Further work is needed to compare the results obtained using our new approach with those obtained using more sophisticated segmentation-based methods of evaluation.

Acknowledgement: The authors would like to thank MGH for 3-D segmented brains which they made publicly available. That hand-annotated data was assumed to be ground truth in the experiments described throughout this paper.

6. 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. In *Proceedings of MICCAI*, 3749:99-106, 2005.

[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, T. Hartkens, and D. L. G. Hill. Non-rigid image registration: theory and practice. *British Journal of Radiology*, 77:140-153, 2004.

[6] 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.

[7] 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.

[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] 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.

[14] 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.

[15] P. Rogelj, S. Kovacic, and J. C. Gee. Validation of a non-rigid registration algorithm for multimodal data. *Medical Imaging*, volume 4684, 2002.

[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.