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

*Some people*

Some places

### ABSTRACT

We present two methods for assessing the performance of non-rigid registration algorithms. We also show that assessment can be carried out with or without the need for some form of ground truth. One method utilizes a measure of overlap among data labels. The other method exploits the fact that, given a set of non-rigidly registered images, a generative statistical appearance model can be constructed. The quality of the model depends on the quality of the registration, and can be evaluated by comparing images sampled from it with the original image set. We derive indices of model specificity and generalisation, as well as introduce a formulation for overlap among anatomical labels. We show that all of them demonstrate the loss of registration as a set of correctly registered images is progressively perturbed. Finally, we compare the sensitivities of these methods.

## 1. INTRODUCTION

Non-rigid registration (NRR) of both pairs and groups of images has in recent years increasingly been used as a basis for medical image analysis. Applications include structural analysis, atlas matching and change analysis [5]. The problem is highly under-constrained and a host of algorithms [4, 18] that have become available will, given a set of images to be registered, in general produce different results.

Various methods have been proposed for assessing the results of NRR [8, 10, 15, 14]. 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 of the methods requires ground truth to be provided *a priori*, whereas the other does not. We compare both methods on a registration of a set of 38 MR brain images and show them to provide a robust evaluation of registration success. Moreover, we demonstrate that both methods are in fact closely correlated if not interchangeable.

## 2. METHOD

The first of the proposed methods assesses registration as the spatial overlap, defined using Tanimoto's formulation of corresponding regions in the registered images. The correspondence is defined by binary labels of distinct image regions (in this case brain tissue classes), produced by manual mark-up of the original images (ground-truth labels). When labels are registered in tandem with corresponding images, these labels become fuzzy. A correctly registered image set will exhibit high relative overlap between corresponding brain structures in different images and, in the opposite case, low overlap with non-corresponding structures. A generalised overlap measure [1] is used to compute a single figure of merit for the overall overlap of all labels over all subjects.

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

where $i$ indexes voxels in the registered images, $l$ indexes the label and $k$ indexes the two images under consideration. $A_{kli}$ 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 a fuzzy set. This generalised overlap measures the consistency with which each set of labels partitions the image volume. The parameter $\alpha_l$ 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 $\alpha_l$ weights for the inverse label volume (which makes the relative weighting of different labels equal), where $\alpha_l$ weights for the inverse label volume squared (which gives labels of smaller volume higher weighting) and where $\alpha_l$ 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 as the quality of a generative, statistical appearance model, constructed from registered images. The idea is that a correct registration produces a true dense correspondence between the images, resulting in a better statistical appearance model of the images.

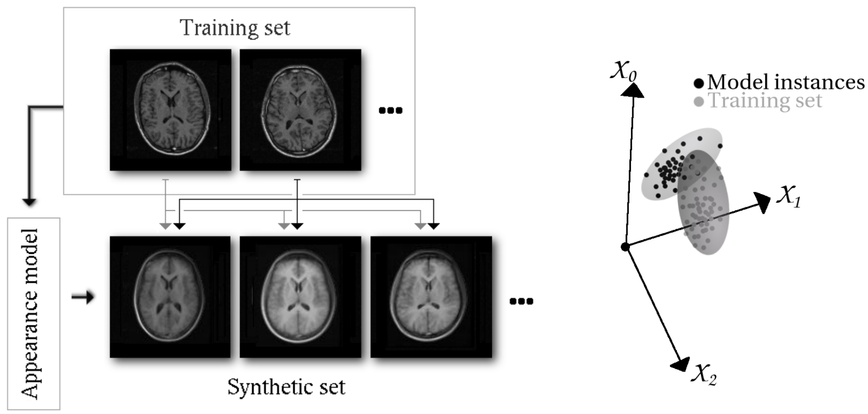


**Fig. 1**. Left: The model evaluation framework. Each image in the training set is compared against every image generated by the model; Right: Training set and model in hyperspace.

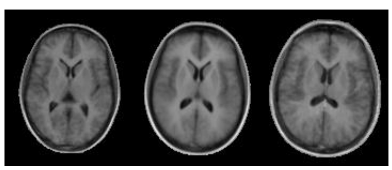

**Fig. 2**. The effect of varying the first model parameter of a brain appearance model by $\pm 2.5$ standard deviations.

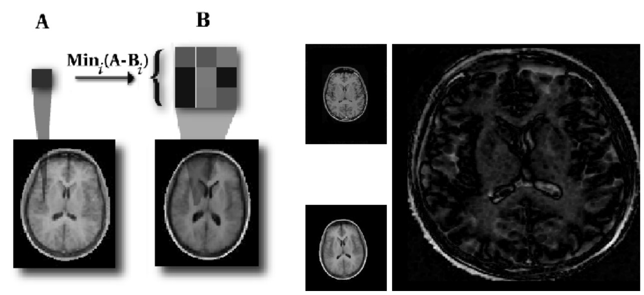


**Fig. 3**. Left: The calculation of a shuffle difference image; Right: An example of the shuffle difference (right) when applied to two MR brain slices

Registration is then evaluated through specificity and generalisation ability [17] of the model, or the ability of the model to i) generate realistic examples of the modelled entity and ii) represent well both seen and unseen examples of the modelled class. In practice, these are evaluated by using generative properties of the model to produce a large number of synthetic examples (in this case brain images) that are then compared to real examples in the original set using some predefined image distance measure. Minimum distances of synthetic examples to examples in the original set and vice versa, give model specificity and generalisation respectively. Image distance is measured as a mean shuffle distance, or minimum Euclidean distance between a pixel in one image and a corresponding neighbourhood of pixels in the other.

To test the validity of the proposed methods, the brain images were annotated with 6 tissue classes including gray, white matter and CSF that provided the ground truth for image correspondence. Initially, the images were brought into alignment using an NRR algorithm based on the MDL optimisation. A test set of different registrations was then created by applying random perturbations to each image in the registered set using diffeomorphic clamped-plate splines. By choosing a different perturbation seed for each image and gradually increasing the magnitude of the perturbations, a series of image sets of progressively worse spatial correspondence and thus registration quality were obtained. By measuring the quality of the registration at each step, the proposed registration assessment measures can be validated.

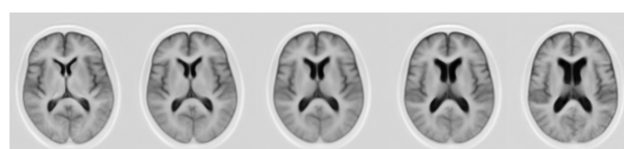

**Fig. 5**. Appearance model which was built automatically by group-wise registration. First mode is shown, $\pm 2.5$ standard deviations.

## 3. RESULTS

Overall, the above approach was applied 10 times using 10 different perturbation seeds to ensure that both methods are consistent and results unbiased. Results of the proposed measures for increasing registration perturbation are shown in Fig-

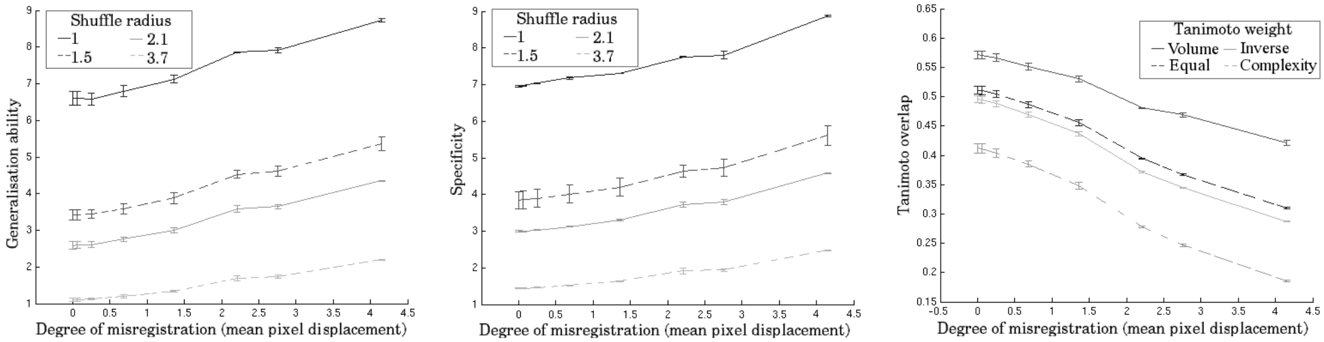


**Fig. 4**. Behaviour of proposed metrics with increasing registration perturbation: a) Generalisation, b) Specificity and c) Tantimoto overlap

ure 4. Note that Generalisation and Specificity plotted for different shuffle neighbourhood radius are in error form, i.e. they increase with decreasing performance. All metrics are generally well-behaved and show a monotonic decrease in registration performance. Such results directly validate the model-based metrics, which are shown be in agreement with the ground truth embodied in the region overlap based measure.

These results also demonstrate that, for all sizes of shuffle neighbourhood, the specificity and generalisation values increase (get worse) with increasing mis-registration. The results for different sizes of shuffle neighbourhood demonstrate that the range of mis-registration over which distinct values of specificity and generalisation are obtained increases as the neighbourhood size increases. We observe similar behaviour as the value of $\alpha_l$ is altered.

Finally, in order to obtain a quantitative comparison of the proposed algorithms we explore sensitivity of the proposed metrics, where the slighter the difference which can be detected reliably, the more sensitive the method. Sensitivity is in this case defined as the rate of change in the measure for a given perturbation range, normalised by the average uncertainty in the measurement over that range. More formally, sensitivity can be defined thus:

$$\frac{m - m_0}{d} / \overline{\sigma} \tag{2}$$

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 over the error bars. Sensitivity is evaluated for all three of the proposed metrics and shown in Figure 3 with errors bars based on both an inter-instantiation error and a measure-specific error. The Specificity measure is the most sensitive for any radius of the shuffle distance followed by the overlap metric and Generalisation, with shuffle radii of 1.5 and 2.1 (equivalent to 3x3 and 5x5 neighbourhoods) giving optimal sensitivity.

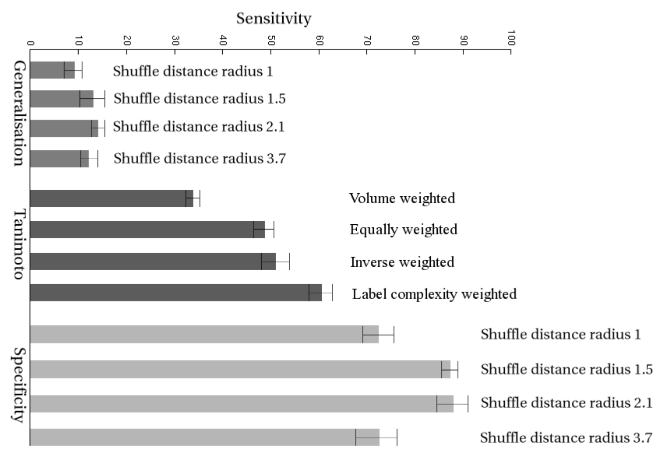


**Figure 2.** The sensitivity of the different registration assessment methods.

## 4. 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 distance improves the range of mis-registration over which we can detect significant changes in registration accuracy.

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 performance, 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.

## 5. 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] 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] 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.

[8] 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.

[9] 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 Imagin*g, 21:1151-66, 2002.

[10] 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.

[11] K. N. Kutulakos. Approximate N-view stereo. In *European Conference on Computer Vision*, 1:67-83, 2000.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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. To be presented in *Information Processing in Medical Imaging*, 2005.

[18] B. Zitova and J. Flusser. Image registration methods: a survey. *Image Vision Computing*, 21:977-1000, 2003.