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 in recent years increasingly been used as a basis for medical image analysis. The problem is highly under-constrained and a host of algorithms that have become available will, given a set of images to be registered, in general produce different results. We present two methods for assessing the performance of non-rigid registration algorithms. One of the methods requires ground-truth solution 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 method are in fact closely correlated if not interchangeable.

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 labels of distinct image regions (in this case brain tissue classes), produced by manual mark-up of the original images (ground-truth labels). A correctly registered image set will exhibit high relative overlap between corresponding brain structures in different images and the other way around. A generalised overlap measure is used to compute a single figure of merit for the overall overlap of all labels over all subjects.

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

In Equation 1, m indexes voxels in the registered images, l indexes the label and l indexes the two images under consideration. Akli and Bkli represent label voxel 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 affects the relative weighting of different labels. With $alpha1 = 1$, label contributions are implicitly volume weighted with respect to one another. We have also considered the cases where alpha weights for the inverse label volume (which makes the relative weighting of different labels equal), where alpha weights for the inverse label volume squared (which gives labels of smaller volume higher weighting) and where alpha 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. Registration is then evaluated through specificity and generalisation ability 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 pre-defined 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.

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 regis-

tration perturbation are shown in Figure 1. 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.

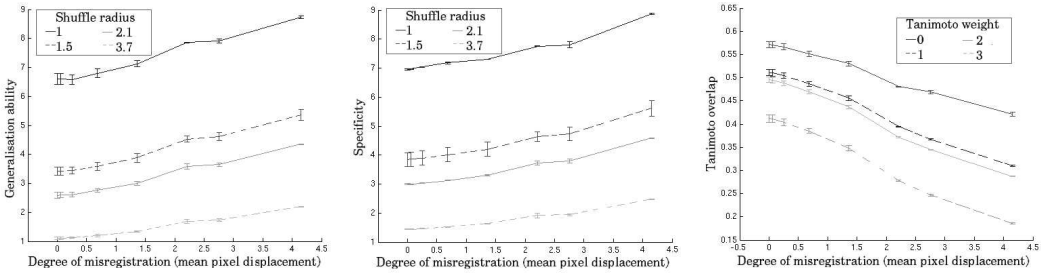

Figure 1. Behaviour of proposed metrics with increasing registration perturbation:
a) Generalisation, b) Specificity and c) Tantimoto overlap

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 by the following formulation:

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

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 2 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.

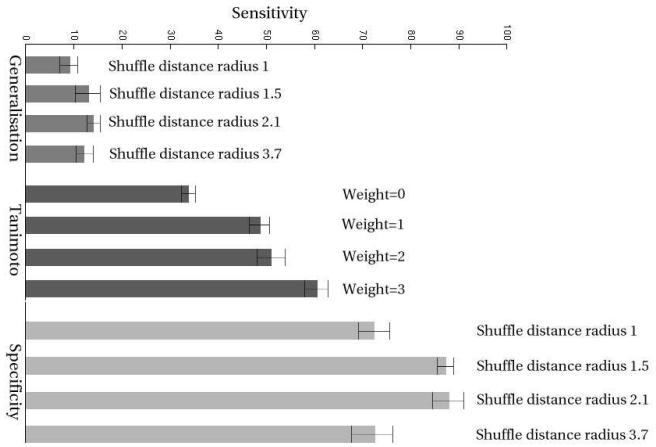

Figure 2. The sensitivity of registration assessment methods.

The results shown in this abstract indicate that 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.