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

We present two methods for assessing the performance of non-rigid registration algorithms. One of them requires ground-truth solutions, whereas the other does not need any form of ground truth. The former method is based on label overlap, which can be computed using Tanimoto's formulation. The method which requires no ground truth 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, and show that they demonstrate the loss of registration as a set of correctly registered images is progressively perturbed. We finally compare the two methods of assessment and show that the latter method, which requires no ground truth, is in fact more sensitive than the one that does.

Over the past few years, non-rigid registration (NRR) has been used increasingly as a basis for medical image analysis. Applications include structural analysis, atlas matching and change analysis. Many different approaches to NRR have been proposed, for registering both pairs and groups of images. These differ in terms of the objective function used to assess the degree of mis-registration, the representation of spatial deformation fields, and the approach to minimizing the mis-registration with respect to the deformations. The problem is highly under-constrained and, given a set of images to be registered, each approach will, in general, give a different result. This leads to a requirement for methods of assessing the quality of registration.

Hereby we outline two methods for assessment. one of which requires ground-truth solutions to be provided *a priori*, yet the other does not. We shall present results which confirm that both methods are valid and proceed to calculating their sensitivities. We find that the method which requires ground-truth solutions is not as sensitive as the method which need not have anything but the raw images and the corresponding deformation fields.

The first of the methods to be described relies on the existence of ground-truth data such as boundaries of image structures, produced by manual markup of distinguishable points. Having registered an image set, the method can measure overlap between elements that have been annotated, thereby implying how good a registration was.

Our latter method is able to assess registration without ground truth of any form. The approach involves automatic construction of appearance models from the registered data, subsequently evaluating, using model syntheses, the quality of that model. Quality of the registration is tightly-related to the quality of its resulting model and the two tasks, namely model construction and image registration, are innately the same. Both involve the identification of corresponding points, also known as landmarks in the context of model-building. Expressed differently, a registration produces a dense set of corresponding points and models of appearance require nothing but the images and the correspondences in order to be built.

To put the validity of both methods to the test, we assembled a set of 2-D 38 MR images of the brain. Each of these images was carefully annotated to identify different compartments within the brain. These anatomical compartments can be perceived as simplified labels that faithfully define brain structure. Our first method of assessment uses the Tanimoto overlap measure to calculate the degree to which labels across the image set overlap. In that respect, it exploits ground truth, which has been identified by an expert, to reason about registration quality.

The second method takes an entirely different approach. It feeds on the results of a registration algorithm, where correspondences have been highlighted, and builds an appearance model given the images and their correspondences. From that model, many synthetic brain images are derived. Vectorisation of these images allows us to embed (or mentally visualise) them in a high-dimensional space. We can then compare the spatial cloud that these synthetic images form with the cloud that is composed from the original image set – the set from which the model has been build. Computing the overlap between these clouds gives insight into the quality of the registration. Simply put, it is a model fit evaluation paradigm. The better the registration, the greater the overlap between those clouds will be.

To compute overlap between two clouds of data, we have devised measures that we refer to as Specificity and Generalisablity. The former tells how well the model fits its seminal data, whereas the latter tells how well the data fits its derived model. It is a reciprocal relationship that 'locks' a data to its model and vice versa. We calculate Specificity and Generalisablity by measuring distances in space. As we seek a measure that is tolerant to slight differences, we use the shuffle distance, not neglecting to compare it against Euclidean distance.

Our assessment framework, by which we test both methods, uses *non-rigid* registration, whereby many degrees of freedom are involved in image transformations. To systematically generate data over which our hypotheses can be tested, we perturb the brain data using clamped-plate splines. In this brain data, correspondences among images are said to be perfect so they can only ever be degraded. We then wish to show that as the degree of perturbation increases, so do the measures of our registration assessment methods.

In our extensive batch of experiments we perturbed the datasets at progressively increasing levels, which led to well-understood misregistration of the data. We repeated these experiments 10 times to demonstrate that both approaches to assessment are consistent are all results unbiased. Having investigated and plotted the measures of *overlap* for each perturbation extent, we see a rather linear decrease in the amount of overlap (Figure X). This means that, as ground-truth-based registration is eroded, the overlap-based measure is able to detect that and the response is very well-behaved, thus meaningful and reliable.

./Graphics/1.eps not found! ./Graphics/2.eps not found!

Figures X&Y. The measured quality of registration as perceived by the overlap-based evaluation (left) and the model-based evaluation (right).

We then undertake another assessment task, this time exploiting the method which *does not* use ground truth. We notice a very similar behaviour (Figure Y), which is evidence that the latter is a powerful and reliable method of assessing the degree of misregistration, or conversely the quality of registration.

As a last step, we embark on the task of comparing the two algorithm, identifying sensitivity as the factor which is most important. Sensitivity reflects on our ability to *confidently* tell apart a good registration from a worse one. The slighter the difference which can be correctly detected, the more sensitive the method. To calculate sensitivity, we compute the amount of change in terms of mean pixel deformation – deformation from the correct solution, that is. We then look at differences in our assessor's value, be it overlap, or Specificity, or Generalisation. We also stress the need to take account of the errors bars as there is both an inter-experiment error and measure-specific error; the two must be composed carefully. The derivation of sensitivity can be expressed as follows:

$$placeholder \tag{1}$$

where X is something... (TODO)

./Graphics/3.eps not found!

Figure Z. The sensitivity of registration assessment methods.
note to self: exclude Gen.?
-.

Figure Z suggests that, for roughly any selection of shuffle distance neighbourhood, the method which does not require ground truth is more sensitive than the method which depends on it. If the trends of these curves are looked at closely, it can be observed that they approximately overlap, which implies that the two methods are very closely correlated.

In summary, we have shown two valid methods for assessing non-rigid registration. The methods are correlated in practice, but the principles they build upon are quite different. Their pre-requisites – if any – likewise. Registration can be evaluated with or without ground-truth annotation and the behaviour of the measures are consistent across distinct data, are well-behaved, and are sensitive. Both methods have been successfully applied to assessment of non-rigid registration algorithms and both methods led to the expected conclusions. That aspect of the work, nonetheless, is beyond the scope of this paper.