Data-Driven, Entropy-Based Measures for Assessing Non-Rigid Registration

Abstract

We present a method for assessing the performance of non-rigid registration algorithms, without the need for any form of ground truth. The 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 a measure which is is based on Shannon's entropy and show that it demonstrates the loss of registration as a set of correctly registered images is progressively perturbed. We also show the tight correlation between our newly-proposed method and an overlap-based measure, which is based on ground-truth anatomical labels of the brain.

1. Introduction

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 [5]. Many different approaches to NRR have been proposed, for registering both pairs and groups of images [3, 19]. 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.

Various methods have been proposed for assessing the results of NRR [8, 10, 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.

In this paper we present a method for evaluating the results of NRR that relies on the image data alone, and can thus be applied routinely without the need for ground truth. The method is based on the observation that, given a set of registered images, it is possible to construct a statistical model of appearance. If the registration is correct, this provides a concise description of the set of images. If it is incorrect, the performance of the model degrades. We base our assessment of the quality of registration on the quality of the resulting model, which can be evaluated using an entropy-based approach.

In the remainder of the paper we explore the background, explain the method in detail, and present validation results using data for which the correct registration is known.

2. Background

2.1. Assessing Non-Rigid Registration

One approach to assessing the results of NRR is to create a set of test images by taking original images and applying known spatial deformations. Evaluation involves comparing the deformation fields recovered by NRR to those known to have been applied [15, 16]. This approach can be used to test a given NRR method 'off-line', but cannot be used to evaluate the results when the method is applied to real data as part of a registration-based analysis.

An alternative approach involves measuring the coincidence of anatomical annotations following registration. Variants of this approach include measuring the mis-registration of anatomical landmarks [8, 10], and the overlap between anatomically equivalent regions obtained using manual or semi-automatic segmentation [10, 15]. These methods are of general application, but are labour-intensive and error prone.

This paper will use a generalised overlap-based approach to provide a 'gold standard' method of assessment. The method requires manual annotation of each image – providing an anatomical/tissue label for each voxel – and measures the overlap of corresponding labels following registration, using a generalisation of Tanimoto's overlap coefficient. Each label for a given image 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 [4], which provides a single figure of merit aggregated over all labels and all images in the set:

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

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 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 as the mean absolute voxel intensity gradient in the label).

2.2. Statistical Models of Appearance

Statistical models of shape and appearance (combined appearance models) were introduced by Cootes, Edwards, Lanitis and Taylor [1, 2, 7], and have since been applied extensively in medical image analysis [9, 13, 17]. The construction of an appearance model depends on establishing a dense correspondence across a training set of images using a set of landmark points marked consistently on each training image.

Using the notation of Cootes [2], the shape (configuration of landmark points) can be represented as a vector $\mathbf{x}$ and the texture (intensity values) represented as a vector $\mathbf{g}$.
The shape and texture are controlled by statistical models of the form

$$\begin{aligned}\mathbf{x} &= \overline{\mathbf{x}} + \mathbf{P}_s\mathbf{b}_s \\ \mathbf{g} &= \overline{\mathbf{g}} + \mathbf{P}_g\mathbf{b}_g\end{aligned} \tag{2}$$

where $\mathbf{b}_s$ are shape parameters, $\mathbf{b}_g$ are texture parameters, $\overline{\mathbf{x}}$ and $\overline{\mathbf{g}}$ are the mean shape and texture, and $\mathbf{P}_s$ and $\mathbf{P}_g$ are the principal modes of shape and texture variation respectively.
Since shape and texture are often correlated, we can take this into account in a combined statistical model of the form

$$\begin{aligned}\mathbf{x} &= \bar{\mathbf{x}} + \mathbf{Q}_s\mathbf{c} \\ \mathbf{g} &= \bar{\mathbf{g}} + \mathbf{Q}_g\mathbf{c}\end{aligned} \tag{3}$$

where the model parameters $\mathbf{c}$ control the shape and texture simultaneously and $\mathbf{Q}_s$, $\mathbf{Q}_g$ are matrices describing the modes of variation derived from the training set. The effect of varying one element of $\mathbf{c}$ for a model built from a set of 2D MR brain image is shown in Fig. 1.

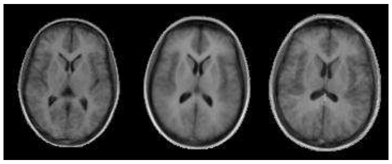

Fig. 1. The effect of varying the first model parameter of a brain appearance model by ±2.5 standard deviations.

3. Model-Based Evaluation

3.1. Model Entropy

Our approach to the assessment of NRR relies on the close relationship between registration and statistical model building, and extends the work of Davies *et al.* on evaluating shape models [6]. We note that NRR of a set of images establishes the dense correspondence which is required to build a combined appearance model. Given the correct correspondence, the model provides a concise description of the training set. As the correspondence is degraded, the model also degrades in terms of its ability to reconstruct images of the same class, not in the training set, and its ability to only synthesise new images similar to those in the training set. If we represent training images and those synthesised by the model as points in a high dimensional space, the clouds represented by training and synthetic images ideally overlap fully (see Fig. 2). The two clouds can be inter-connected to form a graph. Given a measure of the distance between images (see next section), graph entropy and its standard errors [11] can be defined as follows:

$$EQ1 \tag{4}$$

$$Error = EQ2 \tag{5}$$

where $\{I_j : j = 1..m\}$ is a large set of images sampled from the model, $|\cdot|$ is the distance between two images and SD is standard deviation.

Entropy is used as a measure of model compactness. A good model will generate images which are similar to its training set and thus, inter-image distances will be small (entropy values are low for a good model).

Entropy estimates the distance between images generated by the model and their closest neighbours in the training set, but it can also estimates the mean distance between images in the training set and their closest neighbours in the synthesised set. The approach is illustrated diagrammatically in Fig. 3.

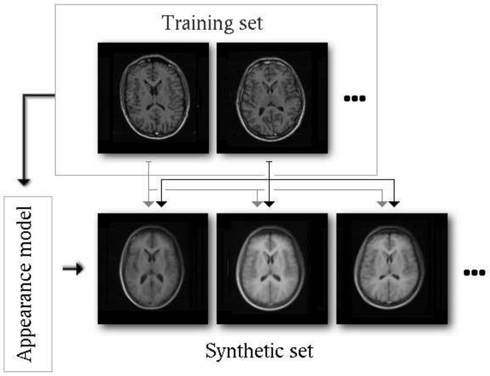

Fig. 3. The model evaluation framework. Each image in the training set is compared against every image generated by the model

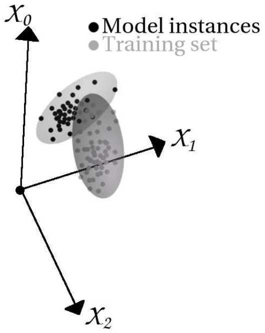

Fig. 2. Training set and model synthesis in hyperspace

3.2. Measuring Distances in Between Images

The most straightforward way to measure the distance between images is to treat each image as a vector formed by concatenating the pixel/voxel intensity values, then take the Euclidean distance. Although this has the merit of simplicity, it does not provide a well-behaved distance measure since it increases rapidly for quite small image misalignments. This observation led us to consider an alternative distance measure, based on the 'shuffle difference', inspired by the 'shuffle transform' [12]. The idea is illustrated in Fig. 4. Instead of taking the sum of squared differences between corresponding pixels, the minimum absolute difference between each pixel in one image and the values in a shuffle neighbourhood around the corresponding pixel is used. This is less sensitive to small misalignments, and provides a more well-behaved distance measure.

4. Validation of the Approach

4.1. Perturbing Ground-Truth

We conducted a series of experiments to test the hypothesis that reduced registration accuracy can be detected using model entropy. An equivalent 2D mid-brain T1-weighted slice was obtained from each of 37 subjects using a 3D acquisition. Fixed binary labels were positioned manually on the right- and left-hand-side grey matter, white matter, lateral ventricles, and caudate nucleus. These labels were used to establish ground truth for our overlap-based assessment approach.

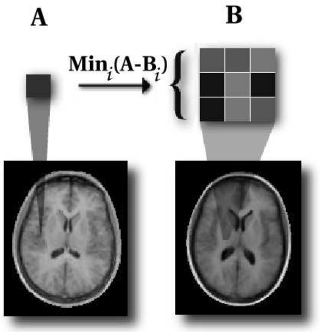

Fig. 4. The calculation of a shuffle difference image

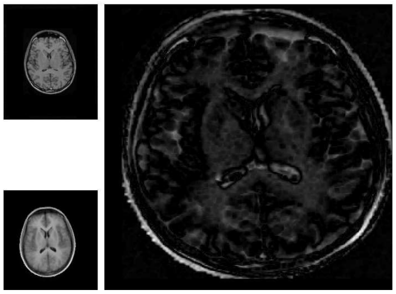

Fig. 5. An example of the shuffle difference image (right) when applied to two MR slices (left)

The images and their accompanying labels were then non-rigidly registered using a groupwise, minimum description length-based algorithm. A statistical appearance model was constructed using the methods described in Section 2. It used the set of landmark coordinates, which had been extracted from the registration, to form the shape vector **x** for each image. We then applied a series of warps, based on biharmonic clamped-plate splines, to the training images and labels, resulting in successively decreasing registration. Each warp resulted in increased displacement, which corresponds to degraded NRR performance. Entropy results were obtained for a range, using $m = 1000$. Corresponding results were obtained for the overlap-based method, which used the resampled anatomical labels to assess label overlap.

The intent of these validation experiments was to show that the model-based approach lies in tight agreement with the overlap-based approach, which uses ground truth.

4.2. Effects of the Shuffle Transform

The experiment described in the previous section was repeated for shuffle neighbourhoods with radii of 1 (Euclidean distance), 1.5, 2.1, and 3.7, to test the hypothesis that this would extend the range over which different degrees of mis-registration could be discriminated.

IMAGE entropy | IMAGE overlap

Fig. 6. [CHANGE TO ENTROPY - ISBI experiments] increasing mis-registration of different shuffle neighbourhood sizes.

4.3. Comparing Entropy-Based Assessment and Ground Truth

IMAGE entropy | IMAGE overlap

Fig. 6. [CHANGE TO ENTROPY - ISBI experiments] increasing mis-registration of different shuffle neighbourhood sizes.

By super-imposing the plots obtained by the different assessment methods, one should able to see the agreement and disagreement. Also of interest is the ability to discern between the sensitivity of the two measures, namely entropy and label overlap fraction.

5. Results

The results of the experiment which test the effect of increasing mis-registration are shown in Fig. X. These demonstrate that, for all sizes of shuffle neighbourhood, entropy values increase (get worse) with increasing mis-registration, with few exceptions at the start (minor misalignments). The results for different sizes of shuffle neighbourhood demonstrate that the range of mis-registration over which distinct values of entropy are obtained increases as the neighbourhood size increases.

6. Discussion and 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. We have also shown that this approach is a good surrogate to the generalised overlap measure, which is an assessment method representative of ground-truth-dependent assessors.

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.

References

[1] 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.

[2] T.F. Cootes, G.J. Edwards and C.J.Taylor. Active appearance models. In *European Conference on Computer Vision*, 2:484-498, 1998.

[3] *Anonymised*

[4] 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 Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 3749:99-106, 2005.

[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] H. Neemuchwala, A. O. Hero, and P. Carson. Image registration using entropy measures and entropic graphs. In *European Journal of Signal Processing*, 2003.

[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 nonrigid 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] *Anonymised*

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