Data-Driven Evaluation of Non-Rigid Registration and Appearance Models

Roy S. Schestowitz*, Vladimir S. Petrovic, Carole J. Twining, Timothy F. Cootes, William R. Crum, and Christopher J. Taylor

***Abstract*— The paper presents a generic approach, which can be used assess the quality of appearance models of the brain. Moreover, this approach is fully capable of assessing and comparing non-rigid registration (NRR) algorithms without exploiting any form of ground truth. We base this approach on the observation that a statistical appearance model can be constructed from a set of non-rigidly registered images. A model can be evaluated by comparing images generated by it with the image set from which it was constructed. The quality of the model depends on the quality of its seminal registration. A registration can also be evaluated by constructing and evaluating models that it produces. Indices are derived which reflect on model specificity and generalisation. We show that these indices are surrogates of Shannon's entropy, which can in itself be used to assess NRR. All of these measures are negatively affected as a set of correctly registered images is progressively perturbed. We compare our results against those which are obtained using overlap-based NRR assessment, which is based on ground-truth anatomical labels. We demonstrate that not only is our approach capable of assessing NRR without ground truth, but it is also more sensitive than the ground-truth-dependent approach. Finally, to demonstrate the practicality of our method, different NRR algorithms – both pairwise and groupwise– are compared in terms of their performance on MR brain data.**

***Index Terms*— Non-rigid registration, ground-truth validation, registration assessment, correspondence problem, appearance models, minimum description length (MDL), Shannon's entropy.**

I. Introduction

NON-RIGID registration (NRR) is ubiquitously used as a basis for medical image analysis. Its applications include atlas matching, analysis of change over time or subjects [7], and structural analysis. A wide variety of approaches exist, which solve the NRR problem. They differ in terms of the objective function that defines mis-registration, the representation of spatial deformation fields, and the approach used to minimize mis-registration by selecting good deformations. Ideally, a composition of aggregated deformations brings a set of images into full alignment, which means that corresponding structures across those images overlap.

Most commonly, pairs of images are being registered [27] at any one time, though groups can be considered too [5]. In the former case, which is referred to as pairwise registration, NRR is applied to just two images in isolation. In the latter case, all the different images are handled simultaneously. This approach has become more popular in recent years and is referred to as groupwise registration in the literature. It has real merits since, given a couple of very dissimilar images, the set as a whole can compensate for that dissimilarity, making its contribution in the form of additional information. It has become a recurrent contention that groupwise registration is the more valid approach.

NRR is an under-constrained problem that suffers from subjectivity in its solution. The solution comprises the set of spatial deformations, which deform one image to match another. For any set of images to be registered, different approaches are likely to produce different results. The different objective functions have different minima, which is the direct effect of the way they define similarity between images.

One obvious way to assessing a given solution is by making use of the ground-truth solution. This idea is based on the principle that any solution can be – in one way or another – numerically evaluated in terms of its divergence from the *correct* solution. Several methods have been demonstrated, which work along these lines [10], [12], [21], [19]. These methods, however, require access to some form of ground truth, which is usually difficult to obtain.

One approach involves the construction of artificial test data, which limits application to 'off-line' evaluation. Furthermore, that approach relies on conditions which are unrealistic, so should be taken with a grain of salt. 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 3D, matters become even more complex. As the correct solution – that which is often based on anatomy – is indeed hard to obtain, NRR assessment without ground truth appears highly valuable.

We consider appearance model, which have been extensively used as the basis for interpretation by synthesis. Such models are derived from sets of training images and, in effect, the models capture statistics about variability within these images. Any model acquires knowledge from its training set and is able to use that knowledge in a variety of ways. Any set of images, which is used to construct an appearance model, is

[DRAFT Placeholder] Manuscript received January 20, 2006 for the TMI special issue on validation; revised March 1, 2006. The work of R. S. Schestowitz was supported by the EPSRC. The work of W. R. Crum was also supported by the EPSRC and fell under the IBIM project 'umbrella'. *Asterisk indicates corresponding author*

*R. S. Schestowitz is with the Division of Imaging Science and Biomedical Engineering, Stopford Building, Oxford Road, University of Manchester, M13 9PT Manchester, United Kingdom.

W. R. Crum is with the Centre for Medical Image Computing, Department of Computer Science, Gower Street, University College London, London WC1E 6BT, United Kingdom. All other authors are with the Division of Imaging Science and Biomedical Engineering, University of Manchester, M13 9PT Manchester, United Kingdom.

Publisher Item Identifier [placeholder].

directly related to that model's quality. When the images are not correspondent, the model is fuzzy and often not valuable. When the images are properly correspondent, the model is improved.

As NRR aims to bring sets of images to a state of full pixel-to-pixel correspondence, the output of a good NRR algorithm builds a good model. We embrace this key observation and exploit the relationship between models and NRR. We use existing methods from both ends of the problem and unify the two as to benefit from both.

The paper presents a framework for building appearance models automatically and then evaluating them. In turn, this method facilitates the assessment of NRR, which requires only the image data, and can therefore be applied routinely, while oblivious to any form of ground truth. The method relies on the fact that, for a given set of registered images, a statistical model of appearance can be constructed. When the registration is correct, the model provides the most concise description of the set of images. As the solution to NRR degrades, so does the performance of model synthesis. Thus, the quality of registration affects the quality of the resulting model and the model itself reflects on the quality of NRR, which makes evaluation of the two somewhat mutual.

The remainder of this paper is structured as follows: it begins by covering background on registration (assessment in particular) and statistical models. It outlines some existing NRR assessment methods, explains about the proposed methods, and presents results which support ideas and theory behind our new method. Validation experiments are then performed where brain models are advertently degraded, by misregistering their training set. Our validation results confirm our method to in tight correlation with ground truth. We show this to be the case by using a generalised measure of label overlap, which uses hand-annotated brain anatomy. Lastly, several registration algorithms are compared to demonstrate one main application of our approach, as applied to brain data. We also show that groupwise registration algorithms produce better results than these of pairwise equivalents. All the same, an algorithm which is based on the minimum description length (MDL) principle, produces results that are comparable, if not better, than the standard groupwise NRR algorithm.

II. Background

A. Non-Rigid Registration

Medical image interpretation is a difficult problem due to cross-invdividual anatomical variation. Additionally, there are factors such as image acquisition errors and soft tissue deformation. In order to perform an analysis of medical images, there needs to be a degree of commonality across these images. Above all, the images must have spatial relationships between them identified. Only by identifying these relationship, can a one-to-one pixel correspondence be obtained. This establishment of inter-image correspondences is possible owing to non-rigid registration.

NRR is a process where images get warped by the means of spatial transformations and their similarity then measured. Warps are chosen which increase this similarity. A good registration algorithm is one which is able to select and apply the 'correct' composition of warps to an image and is able to faithfully estimate similarity between images. In the medical domain, however, there is rarely a solution which is objectively correct. There is no single approach to solving the NRR problem either. The different algorithms in existence use different objective functions, which comprise the way spatial deformation fields are represented, the similarity measure, and the method for selecting deformations to maximise similarity.

Certain algorithms choose to warp one image at a time, fitting it to another arbitrary image in the set, which is known as the reference image or the template. Other algorithms rid the registration framework from bias by comparing any image with the remainder of the set (either in full or partially), directly or implicitly. The results are then not subjected to an arbitrary choice of a reference image. As many ways exist for registering images, solutions remain subjective. Each NRR algorithm will, in principle, lead to a different result, so the need to compare the algorithms becomes more apparent.

B. Assessment of Non-Rigid Registration

1) Deformation Fields Recovery: A common approach to assessment of the results of NRR involves the generation of test images. Such images are created by taking the original images and then applying known deformations to them. The process of evaluation is based on comparison between the deformation fields recovered by NRR and those which have originally been applied [19], [21]. This type of approach can be used to test NRR methods 'off-line'. It cannot, however, be used to evaluate the results when the method is applied to real data as part of a registration-based analysis. Moreover, such artificial deformations fail to resemble real-world situations where there is an innate anatomical variation, which deformation are unable to capture. For instance, there may not be a one-to-one pixel/voxel relationship if images were acquired from different subjects. This property cannot be emulated by any fundamental deformation field.

2) Overlap-based Assessment: The overlap-based approach involves measuring the overlap between of anatomical annotations before and after registration. A good NRR algorithm will be capable of aligning similar image intensities – in particular these which indicate the location of anatomical structures. Alignment of image intensities leads to better overlap between anatomical structures, so the two are closely-correlated.

Similar approaches involve measurement of the misregistration of anatomical regions of significance [10], [12], and the overlap between anatomically equivalent regions obtained using segmentation. This process is either manual or semi-automatic [12], [19]. Although these methods cover a general range of applications, they are labour-intensive andare often prone to errors. They also rely one's ability to faithfully extract anatomical structures from the image intensities alone.

This paper explores one such method, which assesses registration using the spatial overlap. The overlap is 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, in the opposite case – low overlap with non-corresponding structures. A generalised overlap measure [6] is used to compute a single figure of merit for the overall overlap of all labels over all subjects.

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

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 α_l affects the relative weighting of different labels. With $\alpha_l = 1$, label contributions are implicitly volume weighted with respect to one another. This means that large labels contribute more to the overall measure. We have also consider 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. We define label complexity rather arbitrarily as the mean absolute voxel intensity gradient in the label.

More formulations of overlap, other than Tanimoto's, have also been investigated. Their results were shown to be less accurate and they are omitted in the interest of brevity. While our main focus remains assessment that requires no ground truth, the approach above provides a good reference to compare against for validity with respect to ground-truth annotation.

C. Statistical Models of Appearance

Statistical models of shape and appearance (combined appearance models) were introduced by Cootes, Edwards, Lanitis and Taylor [2], [3], [9]. They have been applied extensively in medical image analysis [11], [17], [24], among other related domains. Brain morphometry has been one main point of focus while cardiac imaging incorporated a third and fourth dimension, which is time series [23].

The construction of an appearance model depends on establishing a dense correspondence across a training set of images. That correspondence is defined using a set of landmark points marked consistently on each training image. Landmark points are often prominent anatomical positions, which can easily be identified as they lies on stronger edges. Moreover, they have meaningful properties such as being markers of a boundary of an organ, or as in our case – a brain compartment or the skull.

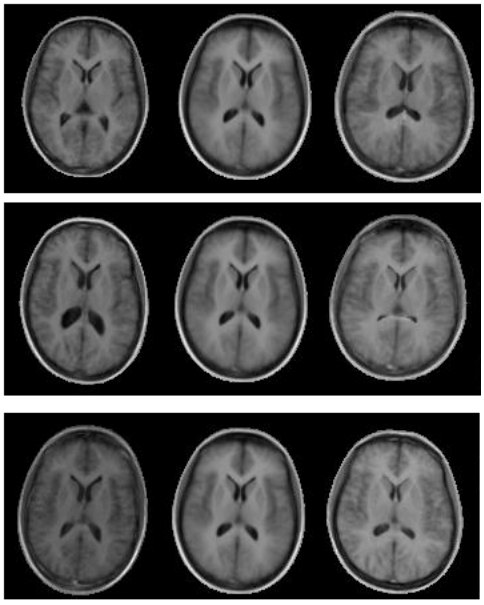

Fig. 1. The effect of varying the first (top row), second, and third model parameters of a brain appearance model by ± 2.5 standard deviations

Using the notation of Cootes [3], 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 two vectors are formed by simple concatenation of values, either intensity (usually grayscale) values or geometric position of landmark points in the image. Using Principal Component Analysis (PCA) [13], the variation in terms of shape and texture can be learned and decomposed. The shape and texture are controlled by a linear statistical models of the form

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

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. By varying $\mathbf{b}_s$ and $\mathbf{b}_g$, the image produced by the model can be altered.

Since shape and texture are often correlated, we can take this into account by applying yet another stage that involves PCA. We then obtain a combined statistical model (encapsulating both shape and intensity) of the form

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

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.

To generate the positions of points in an image we use

$$\mathbf{X} = T_{\mathbf{t}} \mathbf{x} \quad (4)$$

where $\mathbf{x}$ are the points in the model frame, $\mathbf{X}$ are the points in the image, and $T_{\mathbf{t}}\mathbf{x}$ applies a global transformation

with parameters $\mathbf{t}$. For instance, in 2D, $T_{\mathbf{t}}\mathbf{x}$ is commonly a similarity transform with four parameters describing the translation, rotation and scale.

The texture in the image frame is generated by applying a scaling and offset to the intensities, $\mathbf{g}_{im} = T_{gtrans}\mathbf{g}$ where $\mathbf{u}$ is the vector of transformation parameters.

D. The Correspondence Problem

A very key step in construction of combined appearance models is that of identifying dense correspondence across a given set of training images. This is often achieved by marking up the training set by hand, simply identifying significant points in the images and interpolating between these points. In recent years, automation of this process was a problem of great interest. Denser correspondence, which is also accurate, builds a better model. However, that dense correspondence is arduous to obtain. In 3-D, identification of correspondences is hard to obtain objectively. More points of correspondence must be identified as well.

One approach to solving this problem automatically is to use NRR and bring the images to alignment by optimising a similarity measure [11], [17]. A different approach refines initial estimates of the correspondence so as to code the set of images in themost efficient way [1]. We have recently outlined an approach which is based on optimising the total description length of the training set, using its model [25]. A model will be most concise when its training set is fully correspondent.

In Section IV our approach is validated by deliberately perturbing the correspondence in models, i.e. decreasing the registration. Such models were built using manual annotation that establishes a reliable correspondence. In Section V our approach is used to compare common registrations methods [11], [17], as well as our minimum description length approach.

III. Evaluation Method

This section presents the evaluation method that can assess NRR in a model-based fashion. More broadly, it explains the use of the approach for evaluation of appearance models, which in turn make ground-truth-free NRR assessment possible. The section also explores a variety of different analytical methods for estimating the quality of a given model. Ultimately, the aim is to use the most robust and most principled way rather than embrace *ad-hoc* solutions.

A. Specificity and Generalisation

Our approach to model evaluation is based on directly measuring key properties of a given model. An effective model is one which is able to generate a broad range of example of the class of modelled images. This property is referred to as *Generalisation ability*. This property is not sufficient since the model must also generate examples that are *consistent* with the class of modelled images. This property is referred to as *Specificity*. As will be shown later, the two properties are related to (and their estimates can even be substituted by) the notion of Shannon's entropy [22].

The 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 [8]. 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 (Generalisation), and its ability to only synthesise new images similar to those in the training set (Specificity). 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). Given a measure of the distance between images (as described in the next subsection), Specificity, S, Generalisation, G, and their standard errors $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{G}}$ can be defined as follows:

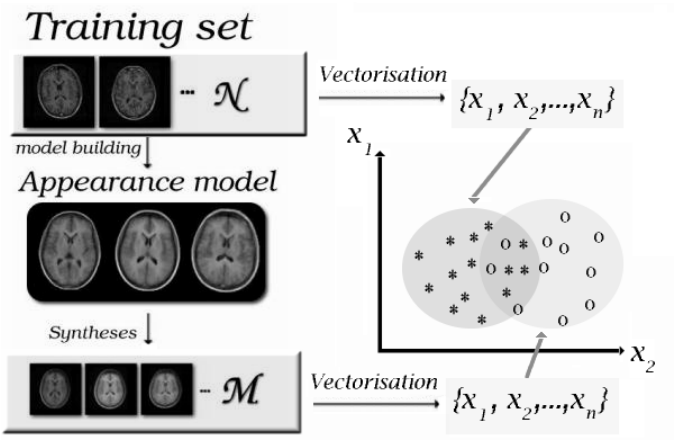

Fig. 2. The model evaluation framework: A model is constructed from the training and images are generatedfromthe model. Each image is vectorised and embedded in hyperscape. Many such points can be visualised as though they form a cloud.

Let $\{I_a(X_0) : a = 1, ...m\}$ be a large image set which has been sampled from the model and has the same distribution as the model. The distance between two images is described by $|\cdot|$ which enables us to define:

$$G = \frac{1}{n}\sum_{i=1}^{n} min_j\,|I_i - I_j|, \tag{5}$$

$$S = \frac{1}{m}\sum_{j=1}^{m} min_i\,|I_i - I_j|. \tag{6}$$

$$\sigma_{\mathcal{G}} = \frac{SD(min_j\,|I_i - I_j|)}{\sqrt{n-1}}, \tag{7}$$

$$\sigma_{\mathcal{S}} = \frac{SD(min_j\,|I_i - I_j|)}{\sqrt{m-1}}. \tag{8}$$

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.

Both values are low for a good model as short distances imply image proximity. Specificity measures the mean distance

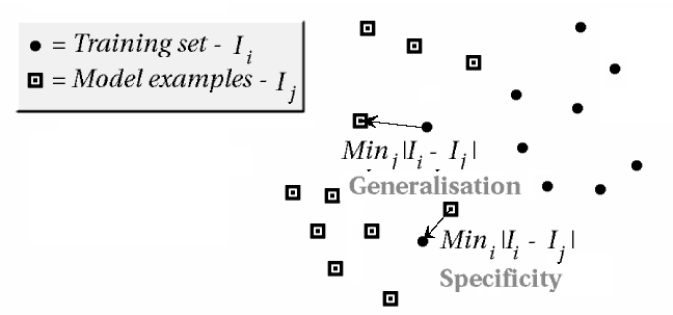

Fig. 3. Simplified hyperspace representation of the model indices calculation. Specificity and Generalisation are derived from the distances between images.

between images generated by the model and their closest neighbours in the training set, whilst Generalisation measures 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.

It can be observed that Specificity and Generalisation fail to account for many of image distances. This might lead to poor and incomplete results. While these measures provide good approximations, we strive to make use of a less simplistic method and exploit work that is related to the MDL principle. The principle was shown to be valuable when dealing with shapes alone.

B. Entropic Graphs

According to the aforementioned definition of Specificity and Generalisation, only nearest image distances get accounted for. This prevents us from attaining a robust measure that is dependent upon the set of images as a whole. Image distances can be perceived as a graph with a network of distances between nodes. We come to consider K nearest neighbours (kNN), wherein *several* nearest neighbours contribute to the measure. Making use of literature in the area, it is possible to treat the problem using entropic graphs analysis, as proposed by Neemuchwala *et al.* [16]. Rather than dealing with two isolated yet reciprocal measures like Specificity and Generalisation, overlap between data clouds can be estimated using an approximation of Shannon's entropy. We adopt the Jensen's dissimilarity measure, which is defined thus

xxxx formula xxxx use similified one? xxxx

where...

To make the calculation even simpler we compute the entropy in the following way

xxxx formula xxxx

In our experiments we consider minimal spanning tree (MST) with just one nearest node. As Fig. /refentropy suggests, we are able to get a measure that is closely-related to both Generalisation and Specificity, only with a higher (?hopefully?) level of certainty. The results also indicate that entropy is by all means a good surrogate of Specificity and Generalisation. It is considered to be a more principled way of measuring clouds overlap and it incorporates normalisation. This means that set sizes do not play any significant role, so a variety of models can be compared regardless of these free parameters.

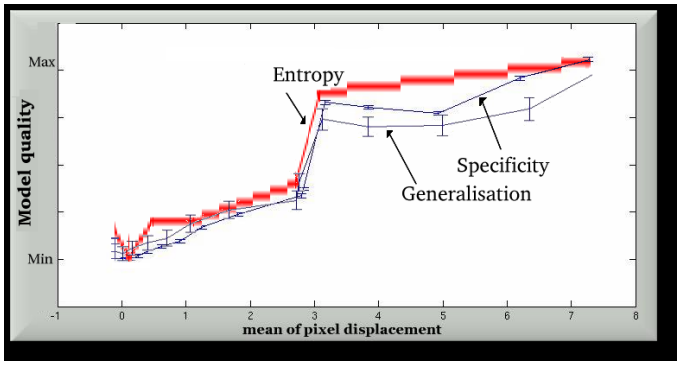

Fig. 4. [PLACEHOLDER-FIG]Specificity, Generalisation and graph entropy (with corresponding error bars) for degraded registration

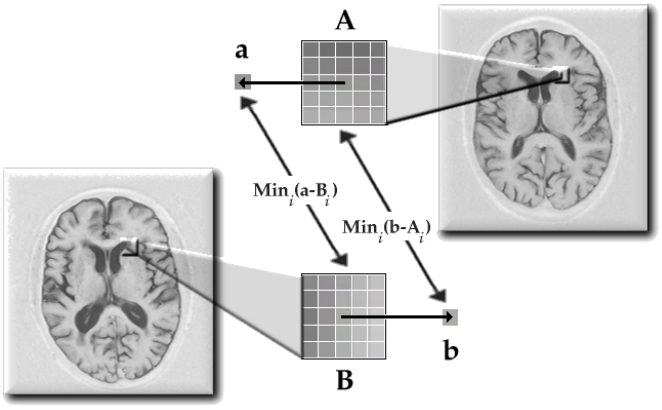

Fig. 5. The calculation of a shuffle difference image

C. 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. It means that each point in one image is compared against its spatially corresponding point in another image. Although this has the merit of simplicity, it does notprovide a very well-behaved distance measure since it increases rapidly for quite small image misalignments [26]. This observation led us to consider an alternative distance measure, based on the 'shuffle difference', inspired by the 'shuffle transform' [14]. The idea is illustrated in Fig. 5. 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 better-behaved distance measure. The tolerance for misalignment is dependent on the size of the neighbourhood which is considered, so we investigate the effect of expanding that neighbourhood (see Fig. 7).

On several occasions we considered the symmetrical shuffle distance, as illustrated in Fig. 6. It applies the shuffle transform in both direction and averages over the sum of the two, thus accounting the the contribution of both. We noted that it entailed no significant improvements. Therefore, experiments

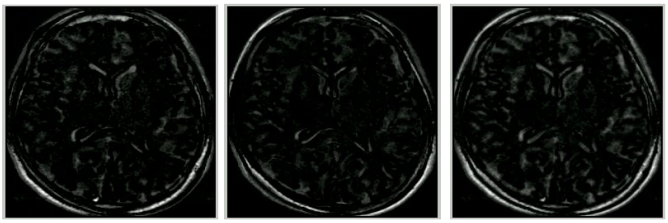

Fig. 6. Examples of the shuffle difference image: from one image to a second image (left), from the second image to the first (centre), and the symmetrical shuffle distance image (right)

in the remainder of this paper choose one image and compute the shuffle distance in just one direction, which is more more efficient and provides equally-valid results.

IV. Validation of the Approach

This section outlines experiments which demonstrate the validity of our registration assessment method. The principal idea is that by creating sets of data where a monotonic degradation exists, algorithms can be put to the test by detecting that degradation. The degree to which an algorithm is able to detect change is another matter, which is further explored in Section VI.

A. Annotated Brain Data

The overlap-based and model-based approaches were validated and compared, using a dataset consisting of 36 transaxial mid-brain slices. These were extracted at equivalent levels from a set of T1-weighted 3D MR scans of different subjects. Brain images were annotated with eight tissue classes including gray matter, white matter, the caudate nucleus and CSF (both left and right) that provided the ground truth for image correspondence. An example image and its corresponding labels are shown in Fig. 8. Initially, the images were brought into alignment using an NRR algorithm, which is based on the MDL optimisation. This alignment is assumed to be the 'correct' solution although, as a rule of thumb and as Section I explains, no registration algorithm is known to give a correct solution.

B. Perturbing Ground Truth

A test set of different registrations was created by applying smooth pseudo-random spatial warps (based on biharmonic Clamped Plate Splines) to each image in the registered set. 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. Registration quality was measured, for each level of registration degradation (perturbation), using several variants of each of the proposed assessment methods.

Overall, the above approach was applied 10 times using 10 different random seeds to ensure that both methods are consistent and the results unbiased. The 10 different warp instantiations were generated for each image for each of seven progressively increasing values of average pixel displacement.

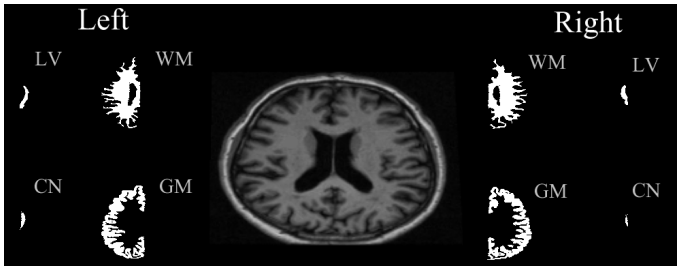

Fig. 8. An example brain image and its accompanying anatomical labels, which include the whitematter, graymatter, caudate nucleus, and lateral ventricle

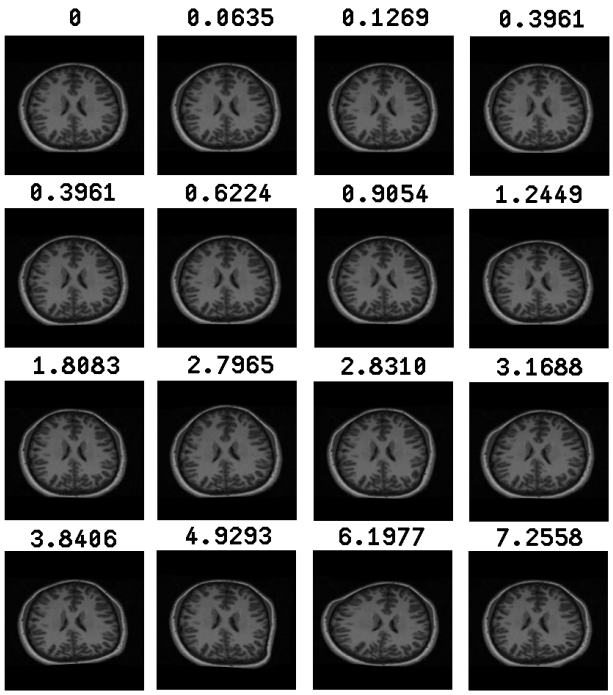

Fig. 9. Examples of registration degradation for increasing scales of smooth CPS warps. Meanpixel displacement for each image is shown at its top.

Figure 9 provided examples from the data as perturbation extent is increased.

C. Validation Results

Results of the proposed measures for increasing registration perturbation are shown in Fig. 11 and Fig. 12. Note that Generalisationand and Specificity plotted for different shuffle neighbourhood radius are in error form, i.e. they increase with decreasing performance. Also shown in Fig. 10 are results from the Tanimoto overlap-based measure, which computes the measure that is based on ground truth. 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.

D. Fine-Tuning Free Parameters

1) The Effects of the Shuffle Transform: The experiments described in the previous section were repeated for shuffle

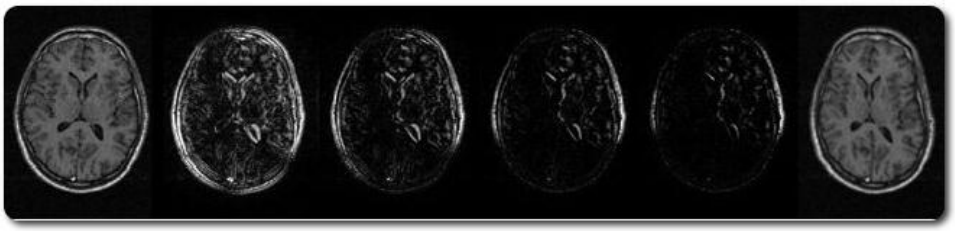

Fig. 7. A comparison between shuffle distance evaluation types. On the left: original image; on the right: warped image; in the centre (from left): shuffle distance with $r = 0$(absolute difference), $1.5, 2.9$ and 3.7.

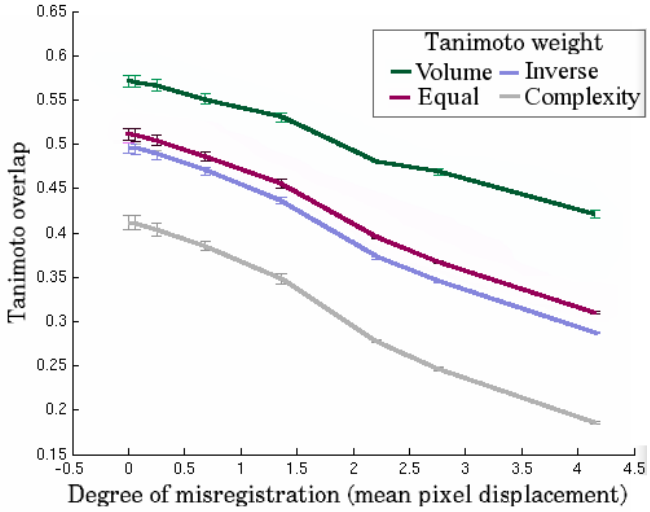

Fig. 10. Overlap (with corresponding errorbars) of brains as their registration degrades

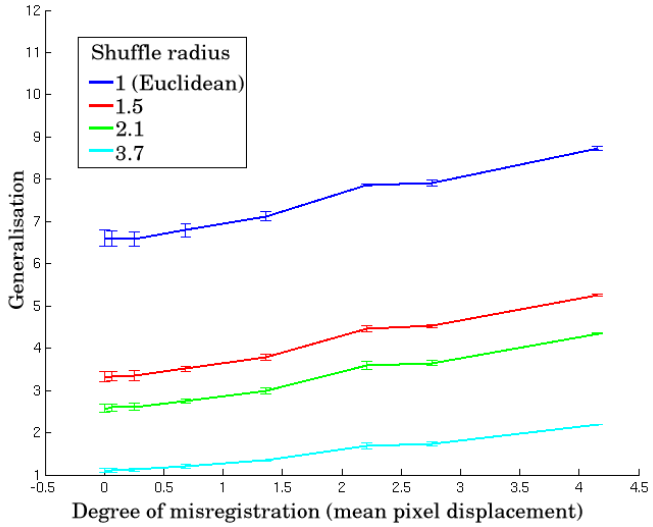

Fig. 11. Generalisation (with corresponding error bars) of brains as their registration degrades

neighbourhoods with radius 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. It can observed, however, that there is no obvious gain when going past a certain neighbourhood size. This is proven numerically in Subsection IV-E.

2) Overlap-based Assessment Weighting Variants: As well as putting Tanimoto overlap to the test, a different overlap formation, known as Dice overlap was considered. All the same, it was shown to be less valuable and was thus not included among the figures. It is worth pointing out that each and every weighting scheme results in a similar behaviour, wherein the decrease in overlap is monotonic and almost linear.

E. Sensitivity Measures

To identify which weighting scheme provides the best behaviour, a measure of sensitivity was desirable. Performance of a measure can be described it term of its ability to discern a good registration from a worse one. Put differently, the problem in question is revealing howsmall a degradation can be and still get detected. This enables us to compare the merits of our model-based assessement method and comparing it with an overlap-based method. It also makes it clearer to see which weighting scheme works best in terms of performance.

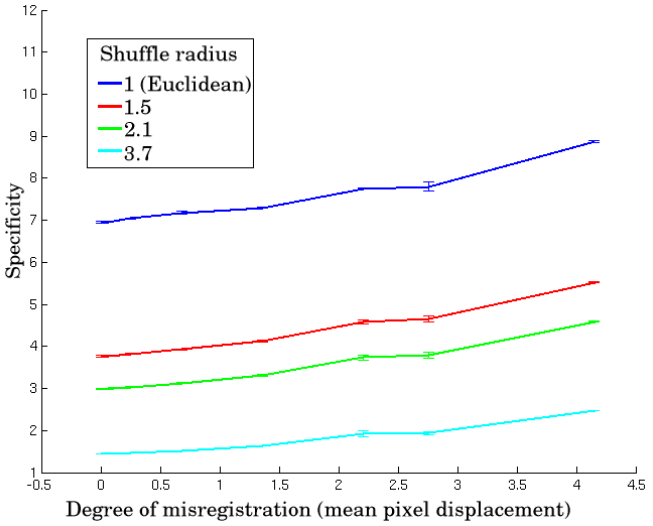

Fig. 12. Specificity (with corresponding error bars) of brains as their registration degrades

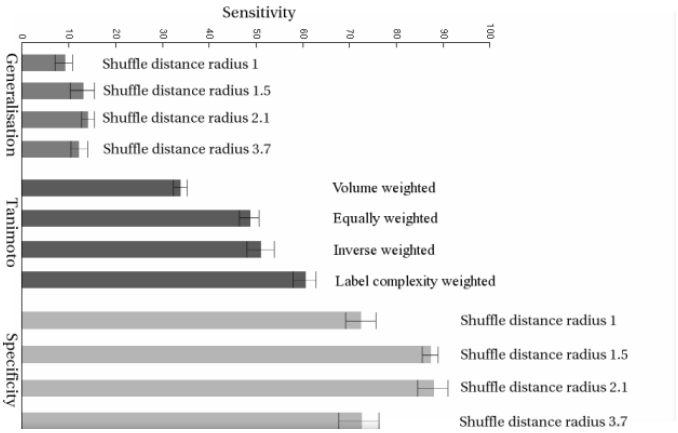

Fig. 13. Sensitivity of different NRR assessment methods

It is worth paying attention to the fact that 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 figures 10, 11, and 12 is plotted in Figure 13 for all the methods of assessment. This shows that the Specificitymeasure with shuffle radius 1.5 or 2.1 is the most sensitive of the measures studied, and that this difference is statistically significant.

V. Applications of the Approach

A. Comparing Appearance

Of lower relevance to this paper are experiments that examine the quality of combined models of appearance. The method we have described enabled us to analyse and thus learn which methods construct appearance models, often by fine-tuning a variety of parameters or establishing the dense correspondence in a most valuable way.

B. Assessing and Comparing Different Methods of NRR

As explained in Section I and II, a common task in medical image analysis is the estimation of correspondences across a group of images, to allow mapping of effects into a common co-ordinate frame when performing population studies. A widely used approach is to use a non-rigid registration algorithm to map a chosen reference image onto each example, defining the correspondence across the group [17]. However, it has been argued [5] that this *pairwise* approach does not take advantage of the full information in the group, and thus may lead to sub-optimal registration. We have been investigating *groupwise* methods of registration which aim to make the best use of the group as a whole when estimating the correspondence. We work within a minimum description length (MDL) framework. The aim is to construct a statistical appearance model which can exactly synthesize each example in the training set as efficiently as possible

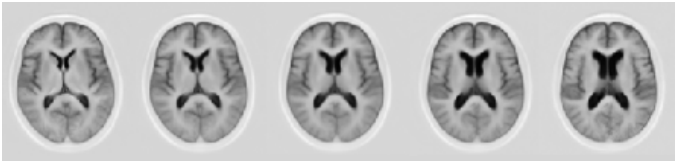

Fig. 14. Appearance model which was built automatically by group-wise registration. First mode is shown, ± 2.5 standard deviations.

[25]. It has been observed that the more the compact the representation, the better the correspondences. The general approach is to define a deformation field between reference frame and each training image. For a given choice of sets of fields, one can compute the cost of encoding the images (a combination of the coding cost of the model, the cost of the parameters and the cost of residuals between the synthesized images and the training images). The effect on this total description length of modifying the deformation fields can be evaluated - the correspondence problem becomes a (very high dimensional) optimisation problem. Within this general framework we compare three different approaches (for details see [25]):

1) Pairwise registration, using the first image as a reference
2) Groupwise registration in which the reference model is just the current mean of the shape and intensities across the training set, and no constraints are placed on the deformations
3) Groupwise registration to the mean including a term encouraging a compact representation of the set of deformations.

Though the algorithms will work in 3D, for the evaluation experiments we concentrate on a 2D implementation (allowing more large-scale experiments to be performed). We have a dataset of 104 3D MR images of normal brains[1] , which have been affine aligned and a single slice at equivalent location extracted from each. Fig. 5 (left) shows examples of extracted slices. In order to evaluate the different registration algorithms outlined above, we register the 104 2D slices using the different techniques, construct statistical models from them and calculate the specificity and generalisation measures.

VI. Results

The results of assessing the generalisation and specificity for each of the three models is shown in Fig. 15. This shows that the full groupwise method is better than the partial method (without shape constraints), which in turn is better than a simple pairwise approach. The evaluation technique allows us to compare different algorithms and make quantitative judgements on the effect of different approaches.

The results of the experiment to test the effect of increasing mis-registration were shown in Fig. 11 and Fig. 12. These demonstrates that, for all sizes of shuffle neighbourhood, the specificity and generalisation values increase (get worse) with increasing mis-registration.

[1]The age matched normals in a dementia study generously provided by X (*anonymised*).

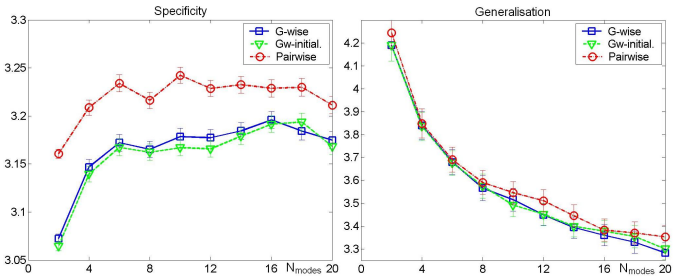

Fig. 15. Specificity and generalisation of the three registration methods

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.

The results of the comparison between three different methods of NRR are shown in Fig. 15 These show that, particularly in terms of specificity, we can distinguish between the three approaches, with the fully groupwise method performing best, as anticipated. A model built using this approach is shown in Fig. 14.

VII. 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. More importantly, we should that out method lied in tight correlation with a successful assessment that uses ground-truth annotation. We have shown that all approaches are capable of detecting statistically significant differences in registration accuracy between three different (plausible) approaches to NRR.

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. It is also worth using the method to investigate a wider variety of registration algorithms, possibly extending our method to 3D.

Acknowledgement

The authors would like to thank David Kennedy of the Center for Morphometric Analysis at MGH. He should be attributed for the fully-annotated brain images, which comprise detailed anatomical labels. An EPSRC grant (GR/S48844/01) for Oscar Camara helped support studies that were based on ground truth. Another EPSRC grant (GR/S82503/01) of the IBIM project helped and encouraged cross-site collaboration.

References

[1] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26:1380-1384, 2004.

[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, G. J. Edwards, and C. J. Taylor. Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23:681-685, 2001.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. M. Fitzpatrick and J. B. West. The distribution of target registrationerror in rigid-body point-based registration. *IEEE Transaction Medical Imaging*, 20:917-27, 2001.

[11] 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.

[12] P. Hellier, C. Barillot, I. Corouge, B. Giraud, G. Le Goualher, L. Collins, A. Evans, G. Malandain, and N. Ayache. Retrospective evaluation of intersubject brain registration. In *Medical Image Computing and Computer-Assisted Intervention*, 2208:258-265, 2001.

[13] I.T. Joliffe. Principal component analysis. In *Springer Series in Statistics*,Springer, New York, 1986.

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

[15] Y. Li, S. Gong, and H. Liddel. Constructing facial identity surfaces in a nonlinear discriminating space. In Proceedings of Computer Vision and Pattern Recognition, pages 258-263, 2001.

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

[17] 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.

[18] 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.

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

[20] S. Romdhani, S. Gong, and A. Psarrou. A multi-view nonlinear active shape model using kernel PCA. In Proceedings of the British Machine Vision Conference, pages 483-492, 1999.

[21] 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.

[22] C. E. Shannon. A mathematical theory of communication. In /empthBell Syst. Tech. J., 27:379-423;623-656, 1948.

[23] M. B. Stegmann. Analysis of 4D cardiac magnetic resonance images. In *Journal of The Danish Optical Society*, 4:38-39, 2001

[24] 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.

[25] 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. Presented in *Information Processing in Medical Imaging*, 2005.
[26] L. Wang, Y. Zhang, and J. Feng. On the Euclidean distance of images. In *Transactions on Pattern Analysis and Machine Intelligence*, 27:1334-1339, 2005.
[27] B. Zitova and J. Flusser. Image registration methods: a survey. *ImageVision Computing*, 21:977-1000, 2003.

PLACE PHOTO HERE

Christopher J. Taylor Biography text here.

Roy Schestowitz Biography text here. Example which excludes photo.

PLACE PHOTO HERE

Vladimir S. Petrovic Biography text here.

PLACE PHOTO HERE

Carole J. Twining Biography text here.

PLACE PHOTO HERE

Timothy F. Cootes Biography text here.

PLACE PHOTO HERE

William R. Crum Biography text here.