IEEE Transactions on Medical Imaging

Evaluating Non-Rigid Registration without Ground Truth

Journal:	*Transactions on Medical Imaging*
Manuscript ID:	draft
Manuscript Type:	Full Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Schestowitz, Roy; University of Manchester, Imaging Science and Biomedical Engineering Twining, Carole; University of Manchester, Imaging Science and Biomedical Engineering Petrovic, Vladimir; University of Manchester, Imaging Science and Biomedical Engineering Cootes, Timothy; Manchester University, Imaging Science and Biomedical Engineering Crum, William; University College London, Centre for Medical Image Computing Taylor, Christopher; University of Manchester, Imaging Science and Biomedical Engineering
Keywords:	Non-rigid registration, Ground-truth validation, Registration assessment, Correspondence problem correspondence problem, Minimum description length

Evaluating Non-Rigid Registration without Ground Truth

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

***Abstract*— We present a generic method for assessing the quality of non-rigid registration (NRR), that *does not* require ground truth, but rather depends solely on the registered images. We consider the case where NRR is applied to a *set* of images, providing a dense correspondence between images. Given this correspondence, it is possible to build a generative statistical model of appearance variation for the set. We observe that the quality of the resulting model will depend on the quality of the correspondence. We define measures of model *specificity* and *generalisation* that can be used to assess the quality of the model and, hence, the quality of the correspondence from which it is derived. The approach does not depend on the specifics of the registration algorithm or the form of the model. We validate the approach by measuring the change in model quality, as the correspondence of an initially registered set of MR images of the brain is progressively perturbed, and compare the results with those obtained using a method based on the overlap of ground-truth anatomical labels. We demonstrate that, not only is the proposed approach capable of assessing NRR reliably without ground truth, but that it also provides a more sensitive measure of misregistration than the overlap-based approach. Finally we apply the new method to compare the performance of repeated pairwise and fully groupwise registration of MR images of the brain.**

I. Introduction

NON-RIGID registration (NRR) of both pairs and groups of images is used widely as a basis for medical image analysis. Applications include structural analysis, atlas matching and change analysis [1]. The problem is highly under-constrained and many different algorithms have been proposed.

The aim of non-rigid registration is to find automatically a meaningful, dense correspondence between a pair (*pairwise* registration), or across a group (*groupwise* registration) of images. A typical algorithm consists of a representation of the deformation fields that encode the spatial variation between images, an objective function that quantifies the degree of misregistration, and a method of optimising the objective function. As different algorithms tend to produce different results when applied to the same set of images [2], there is a need for methods to evaluate the results of NRR.

Various methods of evaluation have been proposed [3], [4], [6], [7]. One approach is to construct artificial test data, applying known deformations to real or synthetic images. This allows algorithms to be evaluated by attempting to recover the applied deformations, but does not allow the results of NRR to be assessed 'in-line' in real applications. An alternative approach is to provide anatomical ground truth for the images to be registered, then measure the degree of anatomical correspondence following NRR. We have used one such method in this paper as a 'gold standard', but the need for expert annotation of the images renders the approach too time-consuming and subjective for routine application. These problems motivate the search for a method of evaluation that can be used routinely in real applications, without the need for ground truth.

The approach we have adopted is based on the observation that, given a set of non-rigidly registered images – however obtained – it is possible to construct a statistical model of appearance that takes account of both the shape and texture variation across the set. Models of this type have been used extensively as a basis for image interpretation by synthesis [9], [10]. We build models by exploiting the dense correspondence across the set of images established by the NRR. The key idea that underpins our approach is that, if the correspondence is poor, the resulting appearance model will be unsatisfactory. This observation allows us to transform the problem of evaluating non-rigid registration into one of evaluating the model generated from the result of registration.

The structure of the paper is as follows. We first provide a brief description of the background to both the assessment of registration, and the construction of appearance models, explaining in more detail the link between the two. We then define two quantitative measures of model (and thus registration) quality, and discuss their implementation. The behavior of these measures is investigated by measuring the effect of deliberately perturbing the registration of an initially registered set of images The results are compared to those obtained using a 'gold standard' method of assessment, based on measuring the overlap of manually annotated ground truth. The results demonstrate that our new measures are closely correlated with those based on ground-truth, and that the proposed approach is actually *more* sensitive to misregistration. Finally, we use the measures we have developed to compare various NRR algorithms applied to the registration of sets of 2D MR

[Draft Placeholder] Manuscript submitted February 7th, 2006 for the TMI special issue on validation

This research was supported by the MIAS IRC project, EPSRC grant No. GR/N14248/01, UK Medical Research Council Grant No. D2025/31 (*"From Medical Images and Signals to Clinical Information"*), and also by the IBIM project, EPSRC grant No. GR/S82503/01 (*"Integrated Brain Image Modelling"*)

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].

brain images, demonstrating the superiority of fully groupwise registration over a repeated pairwise approach.

II. Background

A. Non-Rigid Registration

The aim of non-rigid registration is to find an anatomically meaningful, dense (i.e., pixel-to-pixel or voxel-to-voxel) correspondence across a set of images. This correspondence is typically encoded as a set of spatial deformation fields, one for each image, such that when the deformations are applied to the images, corresponding structures are brought into alignment.

A typical registration algorithm proceeds by optimising some objective function that depends on the similarity of the images after alignment, with respect to the set of deformations. As well as the objective function, it is necessary to define the representation used for the deformation fields and the method for finding the optimum of the objective function. Different choices lead to different registration results, and thus competing methods of NRR – hence the need for an objective and easily applied method of assessment.

B. Evaluation of NRR

Two main approaches to assessing the accuracy of NRR algorithms have been described previously – one based on the recovery of known deformation fields, the other based on measuring the overlap of ground-truth annotations after registration. Both approaches are valid, but neither is easy to apply routinely, and both are better suited to off-line evaluation of algorithms, rather than *in-line* evaluation of the results of NRR in practical applications.

1) Recovery of Deformation Fields: One obvious way to test the performance of a registration algorithm is to apply it to some *artificial* data where the correct correspondence is known. Such test data is typically constructed by applying sets of known deformations (either spatial or textural) to real images. This artificially-deformed data is then registered, and evaluation is based on comparing the deformation fields recovered by the registration algorithm with those that were originally applied [6], [7]. This approach can be used to compare the performance of different NRR algorithms, but since it relies on the creation of artificial test data, cannot be applied in-line. Also, the validity of the approach depends on the ability to construct artificial deformations which mimic the variability found in real images of a given type, which is difficult to guarantee.

2) Overlap-Based Methods: An alternative approach is based on measuring the alignment [3], [4], or overlap [4], [6] of anatomical structures annotated by an expert, or obtained as a result of (semi-)automated segmentation. Manual annotation is expensive to obtain and prone to subjective error. Reliable automated or semi-automated segmentation is extremely difficult to achieve – indeed if it was available it would often obviate the need for NRR.

We have used an 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 [8] which provides a single figure of merit aggregated over all labels and all images in the set:

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

where i indexes voxels in the registered images, l indexes the labels and k indexes image pairs (all permutations are considered). A_{kli} and B_{kli} represent voxel label values for 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 fuzzy sets. 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 structures contribute more to the overall measure. We have also considered the cases where α_l weights labels by the inverse of their volume (which makes the relative weighting of different labels equal), where α_l weights labels by the inverse of their volume squared (which gives regions of smaller volume higher weighting), and where α_l weights labels by their complexity, which we define as the mean absolute voxel intensity gradient over the labelled region.

An overlap score based on a generalisation of the popular Dice Similarity Coefficient (DSC) would also be possible but, since DSC is related monotonically to the Tanimoto Coefficient (TC) by DSC = 2TC/(TC+1) [5] we have not considered this further.

C. Statistical Models of Appearance

Our approach to ground-truth-free evaluation of NRR depends on the ability, given a set of registered images, to construct a generative statistical model of appearance. We have adopted the approach of Cootes et al [9], [10], who introduced models that capture variation in both shape and texture (in the graphics sense). These have been used extensively in medical image analysis in, for example, brain morphometry and cardiac time-series analysis [11]–[13]. Other approaches to appearance modelling could also be considered – we rely only on the generative property in this application

The key requirement in building an appearance model from a set of images, is the existence of a dense correspondence across the set. This is often defined by interpolating between the correspondences of a limited number of user-defined landmarks. Shape variation is then represented in terms of the motions of these sets of landmark points. Using the notation of Cootes et al [9], the shape (configuration of landmark points)

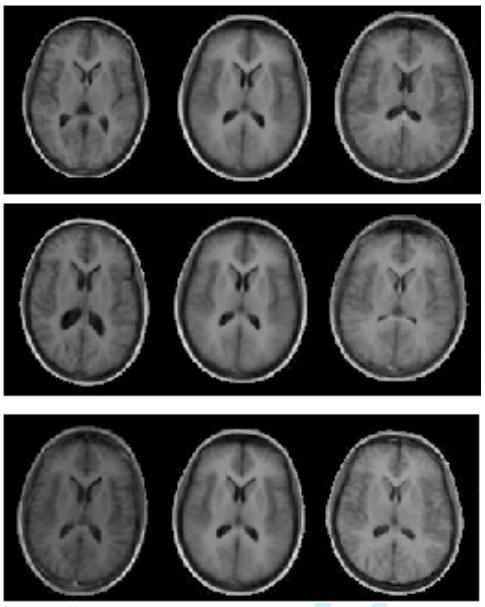


Fig. 1. The effect of varying the first (top row), second, and third model parameters of a brain appearance model by ± 2.5 standard deviations

of a single example can be represented as a vector $\mathbf{x}$ formed by concatenating the coordinates of the positions of all the landmark points for that example. The texture is represented by a vector $\mathbf{g}$, formed by concatenating the image values for the shape-free texture sampled from the image.

In the simplest case, we model the variation of shape and texture in terms of multivariate gaussian distributions, using Principal Component Analysis (PCA) [15], obtaining linear statistical models of the form:

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

where $\mathbf{b}_s$ are shape parameters, $\mathbf{b}_g$ are texture parameters, $\bar{\mathbf{x}}$ and $\bar{\mathbf{g}}$ are the mean shape and texture, and $\mathbf{P}_s$ and $\mathbf{P}_g$ are the principal modes of shape and texture variation respectively.

In generative mode, the input shape ($\mathbf{b}_s$) and texture ($\mathbf{b}_g$) parameters can be varied continuously, allowing the generation of sets of images whose statistical distribution matches that of the training set.

In many cases, the variations of shape and texture are correlated. If this correlation is taken into account, we then obtain a combined statistical model of the more general 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 both shape and texture, and $\mathbf{Q}_s$, $\mathbf{Q}_g$ are matrices describing the general modes of variation derived from the training set. The effect of varying different elements of $\mathbf{c}$ for a model built from a set of 2D MR brain images is shown in Figure 1.

Generally, we wish to distinguish between the meaningful shape variation of the objects under consideration, and the apparent variation in shape that is due to the positioning of the object within the image (the pose of the imaged object). In this case, the appearance model is generated from an (affinely) aligned set of images. Point positions $\mathbf{x}_{im}$ in the original image frame are then obtained by applying the relevant pose transformation $T_{\mathbf{t}}(\cdot)$:

$$\mathbf{x}_{im} = T_{\mathbf{t}}(\mathbf{x}_{model}) \tag{4}$$

where $\mathbf{x}_{model}$ are the points in the model frame, and $\mathbf{t}$ are the pose parameters. For example, in 2D, $T_{\mathbf{t}}$ could be a similarity transform with four parameters describing the translation, rotation and scale of the object.

In an analogous manner, we can also normalise the image set with respect to the mean image intensities and image variance,

$$\mathbf{g}_{im} = T_{gtrans}(\mathbf{g}_{model}), \tag{5}$$

where T_{gtrans} consists of a shift and scaling of the image intensities. For further implementation details see [9], [10].

As noted above, a meaningful dense groupwise correspondence is required before an appearance model can be built. NRR provides a natural method of obtaining such a correspondence, as noted by Frangi and Rueckert [11], [12]. It is this link that forms the basis of our new approach to NRR evaluation.

The link between registration and modelling is further exploited in the Minimum Description Length (MDL) [16] approach to groupwise NRR, where modelling becomes an integral part of the registration process. This is of one of the registration strategies evaluated later in the paper.

III. Model-Based Evaluation of NRR

In the previous section, we described how the results of NRR can be used to build a generative statistical model of image appearance. In this section, we present our method for quantitatively assessing the quality of the model built from the registered data and, hence, the quality of the NRR from which the model was derived. We introduce several variants of the approach, with the aim of finding one which is both robust and sensitive to small misregistrations.

A. Specificity and Generalisation

A good model of a set of training data should possess several properties. Firstly, the model should be able to extrapolate and interpolate effectively from the training data, to produce a range of images from the same general class as those seen in the training set. We will call this *generalisation ability*. Conversely, the model should not produce images which cannot be considered as valid examples of the class of object imaged. That is, a model built from brain images should only generate images which could be considered as valid images of possible brains. We will call this the *specificity* of the model. In previous work, quantitative measures of *specificity* and *generalisation* were used to evaluate shape models [17]. We present here the extension of these ideas to images (as opposed to shapes). Figure 2 provides an overview of the approach.

Consider first the training data for the model, that is, the set of images which were the input to NRR. Without loss of

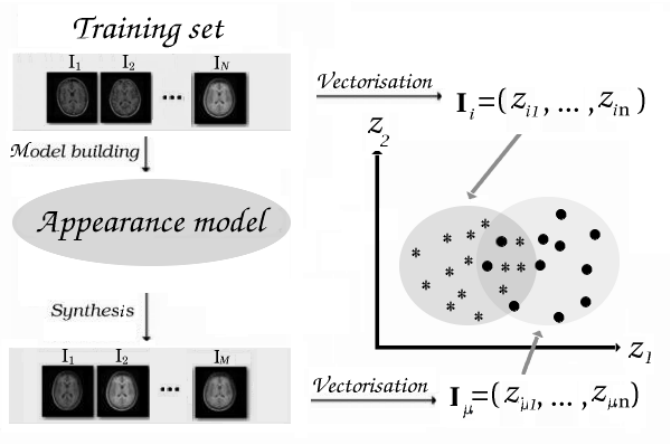


Fig. 2. The model evaluation framework: A model is constructed from the training set and then used to generate synthetic images. The training set and the set generated by the model can be viewed as clouds of points in image space.

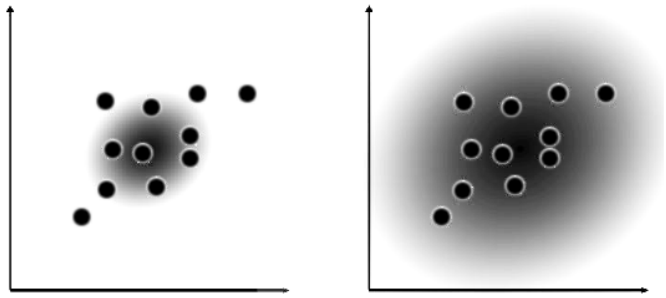


Fig. 3. Training set (points) and model pdf (shading) in image space. **Left:** A model which is specific, but not general. **Right:** A model which is general, but not specific.

generality, each training image can be considered as a single point in an n-dimensional image space. A statistical model is then a probability density function $p(\mathbf{z})$ defined on this space.

To be specific, let $\{\mathbf{I}_i : i = 1, \dots N\}$ denote the N images of the training set when considered as points in image space. Let $p(\mathbf{z})$ be the probability density function of the model. We define a quantitative measure of the *specificity* S of the model with respect to the training set $\mathcal{I} = \{\mathbf{I}_i\}$ as follows:

$$S_\lambda(\mathcal{I}; p) \doteq \int p(\mathbf{z}) \mathbf{min}_i \left(|\mathbf{z} - \mathbf{I}_i|\right)^\lambda \, d\mathbf{z}, \tag{6}$$

where $|\cdot|$ is a distance on image space, raised to some positive power λ. That is, for each point $\mathbf{z}$ on image space, we find the nearest-neighbour to this point in the training set, and sum the powers of the nearest-neighbour distances, weighted by the pdf $p(\mathbf{z})$. Greater specificity is indicated by *smaller* values of S, and vice versa. In Figure 3, we give diagrammatic examples of models with varying specificity.

The integral in equation 6 is approximated using a Monte-Carlo method. A large random set of images $\{\mathbf{I}_\mu : \mu = 1, \dots \mathcal{M}\}$ is generated, having the same distribution as the model pdf $p(\mathbf{z})$. The estimate of the specificity (6) is:

$$S_\lambda(\mathcal{I}; p) \approx \frac{1}{\mathcal{M}} \sum_{\mu=1}^{\mathcal{M}} \mathbf{min}_i \left(|\mathbf{I}_i - \mathbf{I}_\mu|\right)^\lambda, \tag{7}$$

with standard error:

$$\sigma_S = \frac{SD_\mu \left\{ \mathbf{min}_i \{ |\mathbf{I}_i - \mathbf{I}_\mu|^\lambda \} \right\}}{\sqrt{\mathcal{M} - 1}}, \tag{8}$$

where SD_μ is the standard deviation of the set of μ measurements.

A measure of generalisation is defined similarly:

$$G_\lambda(\mathcal{I}; p) \doteq \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \mathbf{min}_\mu \left(|\mathbf{I}_i - \mathbf{I}_\mu|\right)^\lambda, \tag{9}$$

with standard error:

$$\sigma_G = \frac{SD_i \left\{ \mathbf{min}_\mu \{ |\mathbf{I}_i - \mathbf{I}_\mu|^\lambda \} \right\}}{\sqrt{\mathcal{N} - 1}}. \tag{10}$$

That is, for each member of the training set $\mathbf{I}_i$, we compute the distance to the nearest-neighbour in the sample set $\{\mathbf{I}_\mu\}$. Large values of G correspond to model distributions which do not cover the training set and have poor generalisation ability, whereas small values of G indicate models with better generalisation ability.

We note here that both measures can be further extended, by considering the sum of distances to k-nearest-neighbours, rather than just to the single nearest-neighbour. However, the choice of k would require careful consideration and in what follows, we restrict ourselves to the single nearest-neighbour case.

B. Measuring Image Separation

The definitions we have provided for specificity and generalisation require a measure of separation in image space. 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. This means that each pixel/voxel in one image is compared against its spatially corresponding pixel/voxel in another image. Although this has the merit of simplicity, it does not provide a very well-behaved distance measure since it increases rapidly for quite small image misalignments [18]. This observation led us to consider an alternative distance measure, based on the 'shuffle difference', inspired by the 'shuffle transform' [19]. If we have two images $\mathbf{I}_1(\mathbf{x})$ and $\mathbf{I}_2(\mathbf{x})$, then the shuffle distance between them is defined as

$$D_s(\mathbf{I}_1, \mathbf{I}_2) = \frac{1}{n} \sum_{\mathbf{x}} \mathbf{min}_i \|\mathbf{I}_1(\mathbf{x}) - \mathbf{I}_2(\mathbf{N}_i(\mathbf{x}))\| \tag{11}$$

where $\|\cdot\|$ is the absolute difference, there are n pixels (or voxels) indexed by $\mathbf{x}$, and $\{\mathbf{N}_i(\mathbf{x})\}$ is the set of pixels in a neighbourhood of radius r around $\mathbf{x}$.

The idea is illustrated in Figure 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 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 (r), as is illustrated in Figure 4.

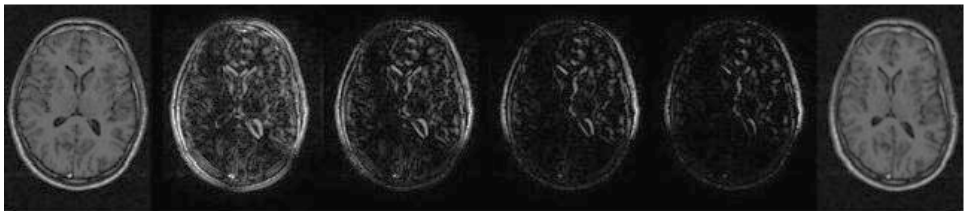


Fig. 4. A comparison between shuffle distance using varying size neighbourhoods (radius r). **Left:** original image, **right:** warped image, **centre, from the left:** shuffle distance with $r = 1$(Euclidean), $1.5, 2.9$ and 3.7 pixels.

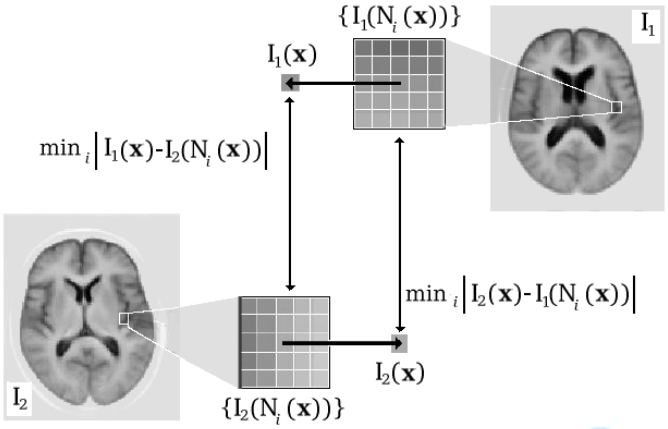


Fig. 5. The calculation of a shuffle difference image

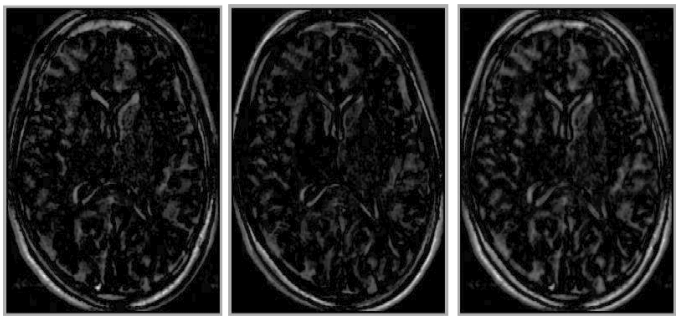


Fig. 6. Examples of the shuffle difference image: from one image to a second image (left), from the second image to the first (centre), and the symmetrical shuffle distance image (right)

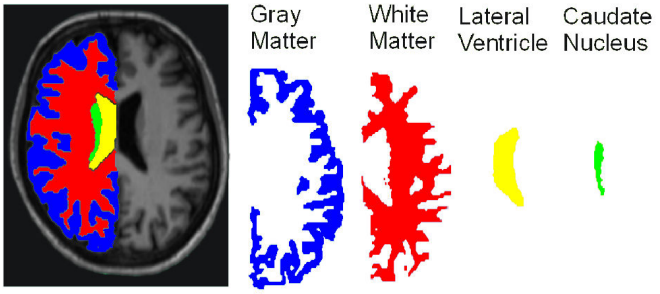


Fig. 7. An example affinely-aligned brain image and its accompanying anatomical labels, both overlaid and expanded, for gray matter, white matter, the lateral ventricles, and the caudate nucleus. Labels are also divided into left and right.

It should be noted that the shuffle distance as defined above depends on the direction in which it is measured (see Figure 6), hence is not a true distance. It is trivial to construct a symmetric shuffle distance, by averaging the distance calculated both ways between a pair of images. However, it was found that the improvement obtained using this was not significant, and did not justify the increased computation time. In what follows, we use the asymmetric shuffle distance.

IV. Experimental Validation

In this section, we discuss the design of experiments to investigate the behaviour of different methods of evaluating NRR. The main idea is that progressive misregistration of initially registered datasets should result in monotonically increasing values of specificity and generalisation (decreasing performance). We also derive a measure of sensitivity to misregistration that can be used to compare methods of NRR evaluation.

A. Brain Dataset with Ground Truth

Our initial dataset consisted of $\mathcal{N} = 36$ transaxial mid-brain 2D slices, extracted at equivalent levels from a set of T1-weighted 3D MR scans of normal subjects. The ground-truth data for this set consists of dense (pixel by pixel) binary tissue labels for the gray and white matter, the caudate nucleus tissue classes and CSF within the lateral ventricles. These labels were further divided into left and right. An example image and its labelling is shown in Figure 7.

The training set was non-rigidly registered using a Minimum Description Length (MDL) NRR algorithm [16]. This registration was used as the starting point for the evaluation.

B. Perturbing the Registration

A test set of different registrations was created by applying smooth pseudo-random spatial warps (based on biharmonic Clamped Plate Splines [20]) 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. Example images from the test set are shown in Figure 8. Ten different warp instantiations were generated for each image and for each of seven progressively increasing values of average pixel displacement.

The correspondence from the initial registration was applied to the *deformed* images resulting in a controlled degree of misregistration. The correspondence becomes progressively worse as the degree of image deformation increases.

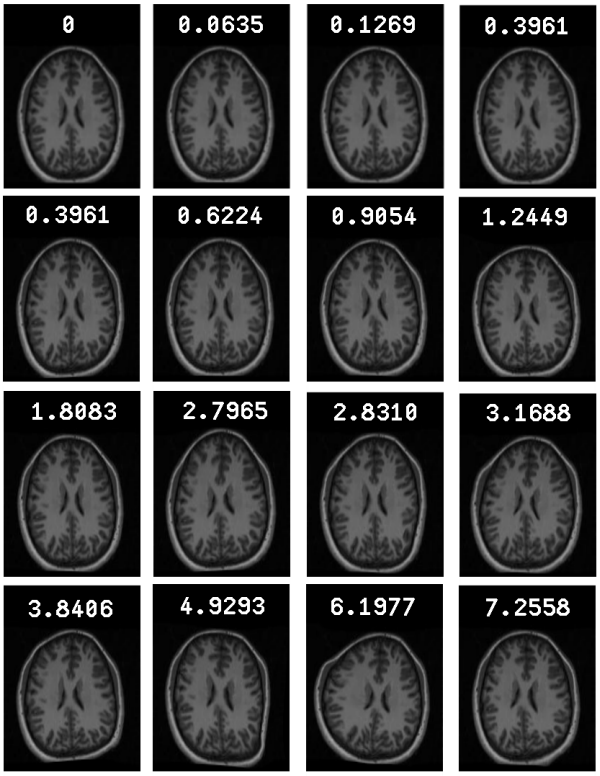


Fig. 8. Examples registration degradation through image deformation for increasing scales of smooth CPS warps. Mean pixel displacement for each image is shown.

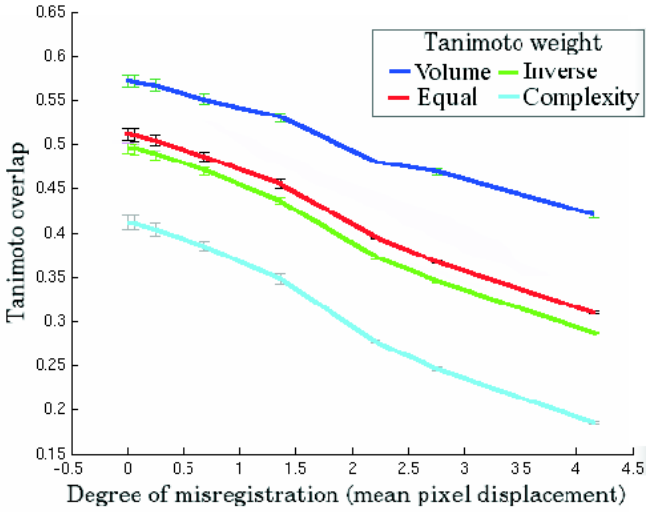


Fig. 9. Overlap measures (with corresponding errorbars) for the brain dataset as a function of the degree of degradation of registration correspondence. The various graphs correspond to the various tissue weightings as defined in section II-B.

C. Measuring Sensitivity

As well as being consistent with ground truth, a good measure of registration quality should also exhibit good sensitivity (measurement accuracy). That is, it should enable us to detect small misregistrations. By evaluating sensitivity we can also assess the effect of varying the parameters of the two approaches that we investigated: the shuffle neighbourhood radius r for the model-based measures and the alternative label weighting options for the generalised Tanimoto overlap.

The size of perturbation that can be detected in the validation experiments will depend both on the change in the values of the measures as a function of misregistration and the mean error on those values. To quantify this, we define the sensitivity of a measure as follows.

$$D(m;d) = \frac{1}{\overline{\sigma}}\left(\frac{m(d)-m(0)}{d}\right), \qquad (12)$$

where $m(d)$ is the value of the measure for some degree of deformation d, $\overline{\sigma}$ is the mean error in the estimate of m over the range. $D(m;d)=1$ is the change in d required for $m(d)$ to change by one noise standard error, which indicates the lower limit of change in misregistration d which can be detected by the measure.

We computed the sensitivity for the data shown in Figures 9, 10(a), & 10(b). The averaged sensitivity over the range of deformations is plotted in Figure 11 for the various measures. The uncertainties on the measurements of sensitivity can also be derived and are shown as error bars on Figure 11. In particular, there are two separate sources of uncertainty: i) errors associated with the finite number of deformation instantiations and ii) errors associated with the finite number of synthetic images used in the evaluation of the figure of merit for NRR. Considering (12), we can evaluate the standard errors in measured quantity m (for a given d) and σ_{mi}, SE_m and SE_{σ_m}, analogously to (8) and (10). Using error propagation the uncertainty on the numerator (T) in (12) is the sum of standard errors on the two measurements, $\sigma_T^2 = (SE_{m(d)})^2 + (SE_{m(0)})^2$, while the uncertainty on the denominator (B) is simply $\sigma_B^2 = SE_{\sigma_m}^2$. Using error propagation for a ratio of variables the uncertainty on the sensitivity becomes:

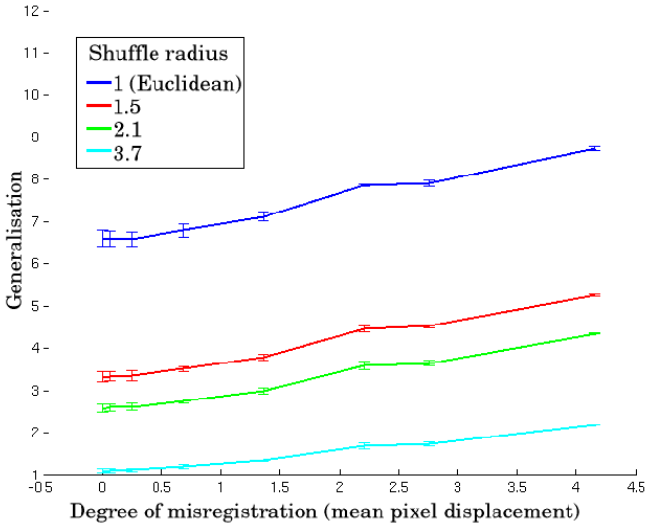
$$\sigma_{D(m;d)} = D(m;d)\sqrt{\left(\frac{\sigma_T}{T}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2 - 2\left(\frac{\sigma_{TB}}{T}\right)\left(\frac{\sigma_{TB}}{B}\right)} \qquad (13)$$

Finally, when sensitivity is aggregated across the deformation range, total uncertainty on the sensitivity, using the addition error propagation rule again becomes:

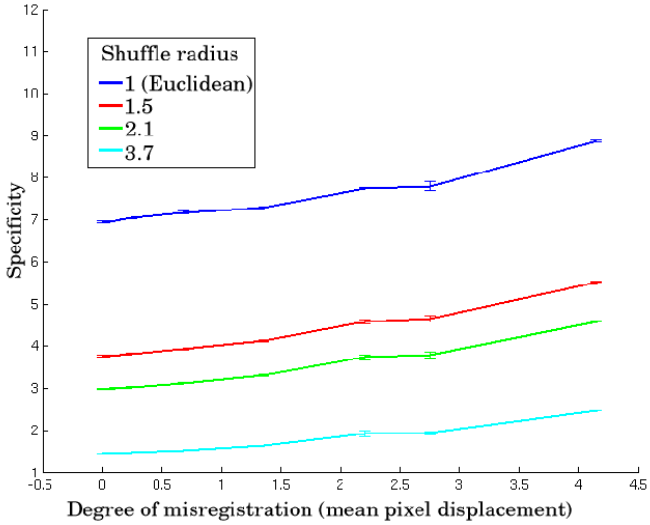
$$\sigma_{Aggr}^2 = \sum_j \sigma_{D(m;d_j)}^2 + \sigma_{D(m;d_{j+1})}^2 - 2\sigma_{D(m;d_j)}\sigma_{D(m;d_{j+1})}. \qquad (14)$$

D. Comparing Registration Algorithms

NRR algorithms can be divided into two general classes: *pairwise* and *groupwise*. Pairwise algorithms register a pair of images at a time. Registration across a group is then achieved by repeated applications of the pairwise algorithm. For example, all images in the training set can be pairwise-registered to some chosen reference example (e.g., [12]). However, this suffers from the general problem that the result obtained depends on the choice of reference. Refinements of this basic approach are possible, where the reference is initialised and



(a) Generalisation



(b) Specificity

Fig. 10. Generalisation & Specificity (with corresponding error bars) for the brain dataset as a function of the degree of degradation of the registration correspondence, and for varying definitions of image distance (varying radius of the shuffle neighbourhood).

updated so as to be representative of the group of images as a whole. The important point to note, however, is that the correspondence for a given training image is defined w.r.t. this reference (which enables consistency of correspondence to be maintained across the group), and the information used in determining the correct correspondence is limited to that contained in that training image and the reference image.

It can be seen that this approach does not take advantage of the full information in the group of images when defining correspondence [21]. Making better use of all the available information is the aim of *groupwise* registration algorithms, where correspondence is determined across the whole set in a principled manner. One such groupwise method is the Minimum Description Length (MDL) formulation as developed by the authors [16]. The main idea is that the appearance model generated from the current correspondence is made an integral part of the process of further registration, the model being continually updated as the process of registration proceeds. The objective function for this groupwise registration is a description length [22], which considers encrypting the entire training set as a coded message, the length of the message in bits being the objective function. Rather than encoding the raw images, the encoding proceeds by describing each training set image as a series of shape and texture deformations applied to some reference. That is, the encoding explicitly uses the model representation of each image from the appearance model built using the found correspondence. Thus the full encoding must also contain the details of the model itself, and the discrepancy between the actual image and the appearance model representation of that image.

We expect the groupwise approach to give significantly better registration results than the repeated pairwise approach. We compare the performance of two variants of the MDL groupwise approach and a pairwise method. These three

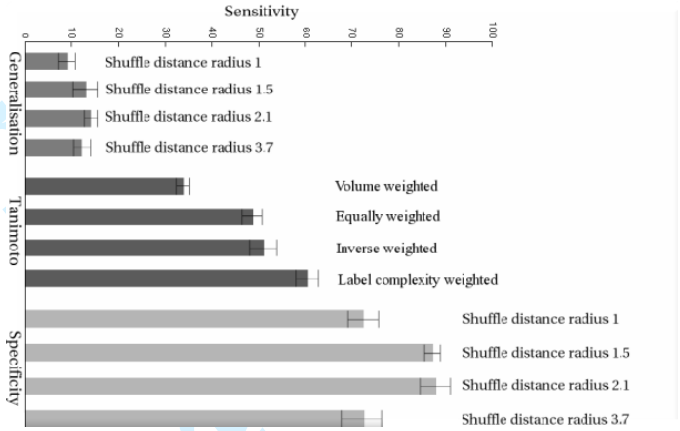


Fig. 11. Sensitivity of different NRR assessment methods

algorithms present a suitable test of the discrimination ability of our proposed evaluation framework.

The different NRR algorithms were compared using 2D images, which allowed large-scale experiments to be performed. 104 T1-weighted 3D MR brain images from a dementia study were affinely aligned, and a mid-brain slice extracted from each, at equivalent locations. The set of images was registered using each of the three registration algorithms. An example of one of the resulting models is shown in Figure 12. In each case the specificity and generalisation were computed.

V. Results

A. Overlap, Specificity and Generalisation

Registration quality was measured, for each level of registration degradation (perturbation), using several variants of each of the proposed assessment methods:

- **Generalised Tanimoto overlap** of the ground-truth data labels (1) for varying values of the label weighting α_l.
- **Specificity & Generalisation** ((7) & (9), $\lambda = 1$), for varying values of the shuffle neighbourhood radius.

Figure 9 shows the results for the Tanimoto overlap measure (1). All overlap variants decay monotonically as a function of misregistration, showing that our perturbed dataset does indeed have the systematic behaviour we require.

Results for the proposed specificity S (7) and generalisation G (9) measures as a function of the displacement magnitude are shown in Figures 10(a) & 10(b). Results are given for varying values of the shuffle neighbourhood radius r, including Euclidean distance, $r = 1$. Note that Generalisation and Specificity are in error form, and increase monotonically with increasing misregistration, for all values of shuffle radius. The strong qualitative agreement with the results for the overlap measure demonstrates the validity of the model-based measures.

B. Sensitivity

Figure 11 compares the sensitivities of the different methods of evaluating NRR. This shows that specificity is more sensitive (is able to detect smaller misregistrations) than either generalisation or the overlap-based approach. Note from the error bars that these differences are statistically significant. Maximum sensitivity is achieved with shuffle radii of 1.5 and 2.1. Generalisation is shown to be a valid but not particularly sensitive measure of misregistration.

C. Comparing Registration Algorithms

We compared the results of three registration algorithms as outlined in Section IV-D:

- Pairwise registration of each training set image to a fixed reference image, using an image from the training set as a reference.
- Groupwise registration based on the MDL algorithm described above, with no constraints on the spatial deformations during the registration process (Groupwise 1).
- Groupwise registration based on the MDL algorithm described above, using a statistical shape model to constrain the allowed spatial deformations between the images during registration [16] (Groupwise 2).

The results are shown in Figure 13. The specificity obtained for the two groupwise methods is significantly better than that obtained using the pairwise approach, implying better registration, but it is not possible to distinguish between the two groupwise methods. As might be expected from the sensitivity results presented above, it is not possible to distinguish between any of the methods using generalisation.

VI. Discussion and Conclusions

We have described a model-based approach to evaluating the results of NRR of a group of images. The most important advantage of the new method is that it does not require any ground truth, but depends solely on the registered images themselves.

We have validated the approach by studying the effect of perturbing, progressively, the registration of an initially registered set of images, comparing the results with those obtained using a 'gold standard' measure based on the overlap of ground-truth anatomical labels. We have shown that our new method provides measures of registration accuracy that are monotonic functions of the known misregistration, and that one, *specificity*, provides a more sensitive measure of misregistration than the approach based on ground truth. The model-based approach requires a distance measure in image space, and we have also demonstrated that the use of shuffle distance, rather than Euclidean distance, improves the sensitivity of the approach.

We have further validated the approach and illustrated its application by performing a comparative evaluation of the results obtained using three different NRR algorithms, demonstrating the superiority of a fully-groupwise algorithm over a repeated pairwise approach.

The experiments were performed in 2D to limit the computational cost of running a large-scale evaluation for a range of parameter values and with repeated measurements. The extension to 3D is, however, trivial, though the calculation of shuffle distance for 3D images increases the computational cost significantly.

In the experiments we have reported we used linear appearance models in the evaluation, but any generative model-building approach could, in principle, be used. It is important to emphasise that the method is not restricted to evaluating model-based NRR algorithms, though we presented results for one such approach; our model-based measures of registration accuracy can be applied to any set of non-rigidly registered images, however they were obtained.

At first sight, the result that one of the model-based measures is more sensitive than the method based on the overlap of ground-truth labels seems counter-intuitive. On further reflection this is not, however, so surprising – since the model-based approach uses the full intensity image, which provides a far richer description of local alignment than that provided by the relatively featureless label images.

Overall, we believe that our approach provides a powerful approach to evaluating NRR methods, allowing subtle differences to be detected without the need for any additional information. This should prove valuable both in helping to guide the development of new NRR methodology and in providing quality control in routine applications of NRR.

Acknowledgement

The authors would like to thank David Kennedy of the Center for Morphometric Analysis at MGH, for providing the fully-annotated brain images. Images from age-matched normals in a dementia study were generously provided by Prof. Alan Jackson, University of Manchester.

References

[1] T. H. W. R. Crum, T Hartkens and D. L. G. Hill, "Non-rigid image registration: theory and practice," *British Journal of Radiology*, vol. 77, pp. 140–153, 2004.

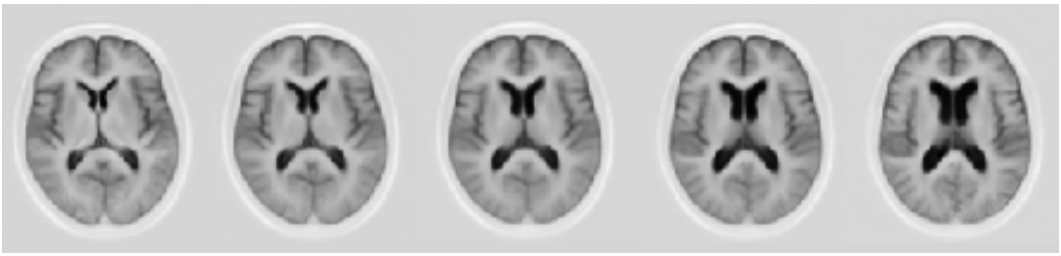


Fig. 12. Appearance model which was built automatically by group-wise registration. First mode is shown, ± 2.5 standard deviations.

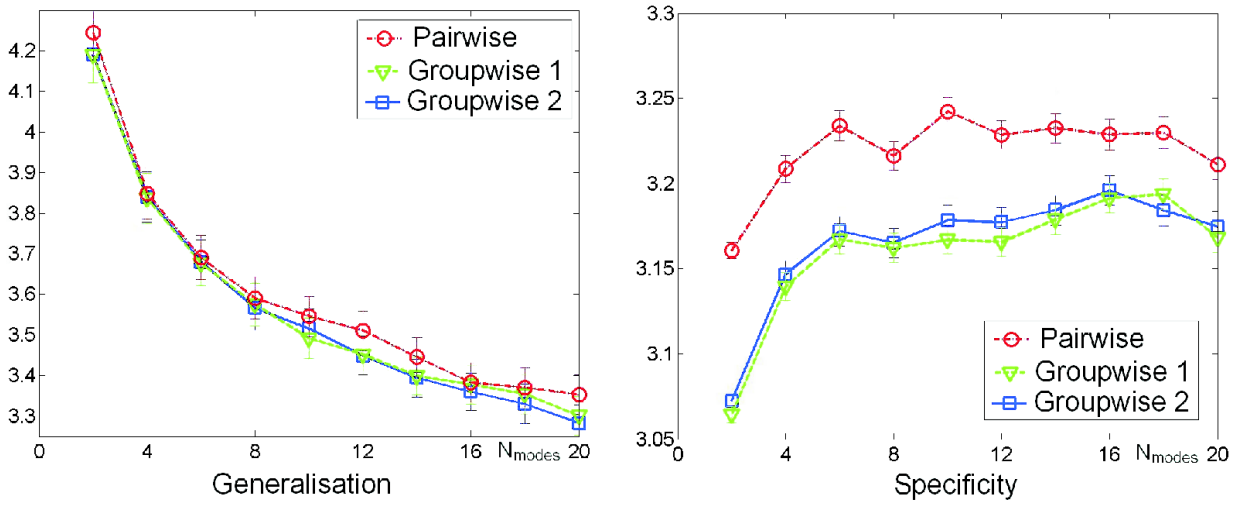


Fig. 13. Generalisation and Specificity of the three registration methods as a function of the number of modes included in the appearance model.

[2] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image and Vision Computing*, vol. 21, pp. 977 – 1000, 2003.
[3] J. M. Fitzpatrick and J. B. West, "The distribution of target registration error in rigid-body point-based registration," *IEEE Trans. Med. Imag.*, vol. 20, pp. 917–927, 2001.
[4] P. Hellier, C. Barillot, I. Corouge, B. Giraud, G. L. Goualher, L. Collins, A. Evans, G. Malandain, and N. Ayache, "Retrospective evaluation of inter-subject brain registration," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), Lecture Notes in Computer Science*, vol. 2208. Springer, 2001, pp. 258–265.
[5] D.W. Shattuck, S.R. Sandor-Leahy, K.A. Schaper, D.A. Rottenberg and R.M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, pp. 856–876, 2001.
[6] P. Rogelj, S. Kovacic, and J. C. Gee, "Validation of a nonrigid registration algorithm for multimodal data," in *Proceedings of Medical Imaging 2002, Image Processing, SPIE Proceedings*, vol. 4684, 2002, pp. 299–307.
[7] J. A. Schnabel, C. Tanner, A. Castellano-Smith, A. Degenhard, M. O. Leach, D.R. Hose, D. L. G. Hill, and D. J. Hawkes, "Validation of non-rigid registration using finite element methods: Application to breast MR imagesIEEE Transactions on Medical Imaging, vol. 22(2):238-247, 2003
[8] 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), Lecture Notes in Computer Science*, vol. 3749. Springer, 2005, pp. 99–106.
[9] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *Proceedings of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science*, vol. 1407. Springer, 1998, pp. 484–498.
[10] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, vol. 2. Springer, 1998, pp. 581–595.
[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 Trans. Med. Imag.*, vol. 21, pp. 1151–1166, 2002.
[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 Trans. Med. Imag.*, vol. 22, no. 8, pp. 1014–1025, 2003.
[13] M. B. Stegmann, B. K. Ersboll, and R. Larsen, "FAME - a flexible appearance modeling environment," *IEEE Trans. Med. Imag.*, vol. 22, no. 10, pp. 1319–1331, 2003.
[14] M. B. Stegmann, "Analysis of 4d cardiac magnetic resonance images," *Journal of The Danish Optical Society*, vol. 4, pp. 38–39, 2001.
[15] I. Joliffe, *Principal component analysis*. New York: Springer, 1986.
[16] C. J. Twining, T. F. Cootes, S. Marsland, V. Petrovic, R. Schestowitz, and C. J. Taylor, "A unified information-theoretic approach to groupwise non-rigid registration and model building." in *Proceedings of Information Processing in Medical Imaging (IPMI), Lecture Notes in Computer Science*, vol. 3565. Springer, 2005, pp. 1–14.
[17] 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 Trans. Med. Imag.*, vol. 21, no. 5, pp. 525–537, 2002.
[18] L. Wang, Y. Zhang, and J. Feng, "On the euclidean distance of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, pp. 1334–1339, 2005.
[19] K. N. Kutulakos, "Approximate n-view stereo," in *Proceedings of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science*, vol. 1842. Springer, 2000, pp. 67–83.
[20] C. J. Twining, S. Marsland, and C. J. Taylor, "Measuring geodesic distances on the space of bounded diffeomorphims," in *Proceedings of the British Machine Vision Conference (BMVC'02)*, 2002.
[21] T. F. Cootes, S. Marsland, C. J. Twining, K. Smith, and C. J. Taylor, "Groupwise diffeomorphic non-rigid registration for automatic model building," in *Proceedings of European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science*, vol. 2034. Springer, 2004, pp. 316–327.
[22] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. World Scientific Press, 1989.