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 (ISBE) Twining, Carole; University of Manchester, Imaging Science and Biomedical Engineering (ISBE) Petrovic, Vladimir; University of Manchester, Imaging Science and Biomedical Engineering (ISBE) Cootes, Timothy; University of Manchester, Imaging Science and Biomedical Engineering (ISBE) Crum, William; University College London, Centre for Medical Image Computing Taylor, Christopher; University of Manchester, Imaging Science and Biomedical Engineering (ISBE)
Keywords:	Non-rigid registration, Ground-truth validation, Registration assessment, 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 three different registration algorithms on a set of MR images of the brain, demonstrating that the method is able to discriminate between different methods of registration in a practical setting.**

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 [2]. The problem is highly under-constrained and many different algorithms have been proposed.

The aim of NRR is to find, automatically, a meaningful dense correspondence between a pair (*pairwise* registration), or across a group of images (*groupwise* registration). 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 with respect to the deformation fields. As different algorithms generally produce different results when applied to the same set of images [3], there is a clear need for methods to evaluate the results of NRR.

Various methods of evaluation have been proposed [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]. To build a model we exploit 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 three NRR

[Draft Placeholder] Manuscript submitted May 26th, 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].

algorithms applied to the registration of sets of 2D MR brain images, demonstrating the superiority of a fully groupwise registration algorithm 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 applied originally [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 [4], or overlap [4], [6] of anatomical structures annotated by an expert, or obtained as a result of (semi-)automated segmentation. This has the disadvantege that manual annotation is expensive to obtain and prone to subjective error, whilst 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 [1]. 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:

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

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 standard error in $\mathcal{O}$ can be estimated in the normal way from the standard deviation of the pairwise overlaps.

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 as we rely only on the generative property of such models 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

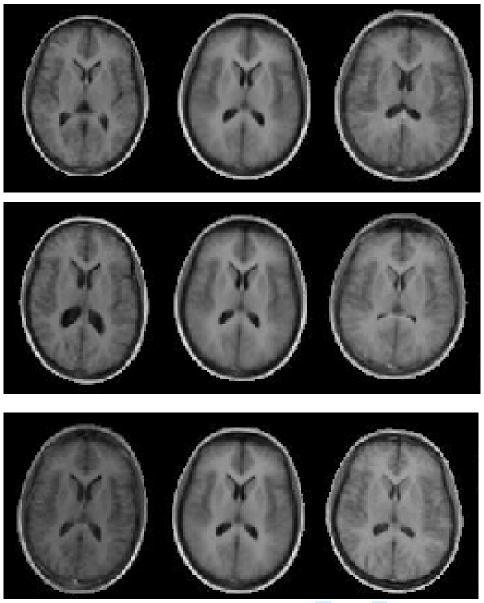


Fig. 1. The effect of varying the first (top row), second, and third parameter of a brain appearance model by ± 2.5 standard deviations

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) 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 image values (texture) sampled over a regular grid on the *registered* image. This means that the a given element in $\mathbf{g}$ is sampled from an equivalent point in each image, assuming the registration is correct.

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

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

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 obtain a combined statistical model of the more general form:

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

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. The number of modes (columns) in $\mathbf{Q}_s$ and $\mathbf{Q}_g$ is one less than the number of images. In practice, it is often possible to approximate images well, using fewer modes m.

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 $T_{\mathbf{t}}(\cdot)$:

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

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,

$$\mathbf{g}_{im} = T_{\vec{u}}(\mathbf{g}_{model}), \qquad (5)$$

where $T_{\vec{u}}$ 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 one of the registration strategies evaluated in this 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 image modelled. That is, a model built from brain images

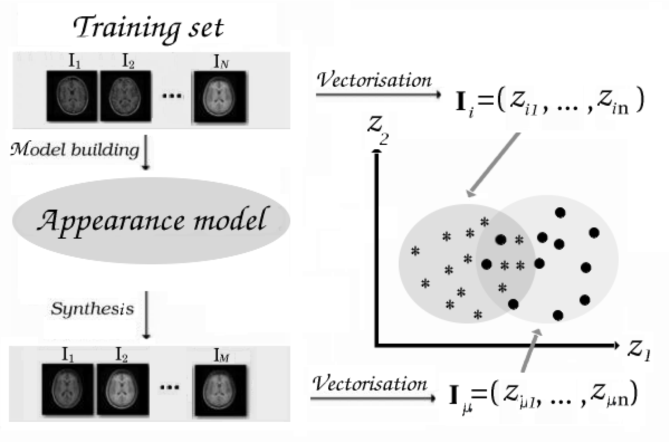


Fig. 2. The model evaluation framework: A model is constructed from the training set and used to generate synthetic images. The training set and the set generated by the model can be viewed as clouds of points in image space ($\mathbf{I}_i$ represented by stars, and $\mathbf{I}_\mu$ represented by dots).

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 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 (pdf) $p(\mathbf{z})$ defined on this space.

To be specific, let $\{\mathbf{I}_i : i = 1, \ldots \mathcal{N}\}$ denote the $\mathcal{N}$ 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 (for the remainder of this paper we will consider only the case λ = 1). 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 differing specificity.

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

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

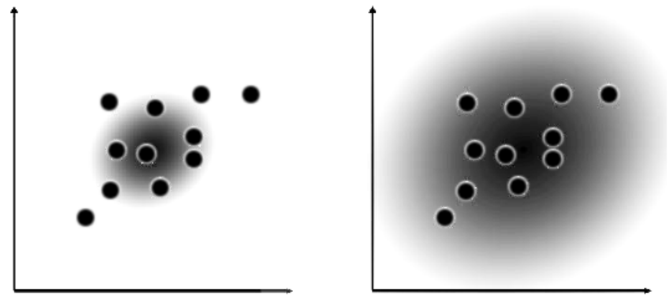


Fig. 3. Training set (points) and model pdf (shading) in image space. **Left:** A model which is specific, but not general. **Right:** A model which is general, but not specific.

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. Note that this definition of S does not require that we construct the space of images, we simply need to be able to define distances between images. This is discussed in Section III-B below.

We define a measure of generalisation similarly, simply reversing the direction of the nearest-neighbour distance measure:

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

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].

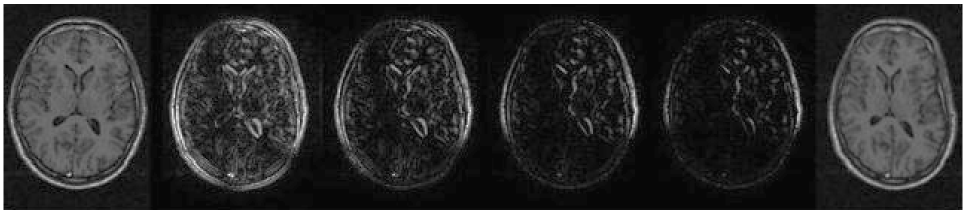


Fig. 4. A comparison between shuffle difference images evaluated using various size neighbourhoods (radius r). **Left:** original image, **right:** warped image, **centre, from the left:** shuffle distance with $r = 1$(Euclidean), $1.5, 2.9$ and 3.7 pixels.

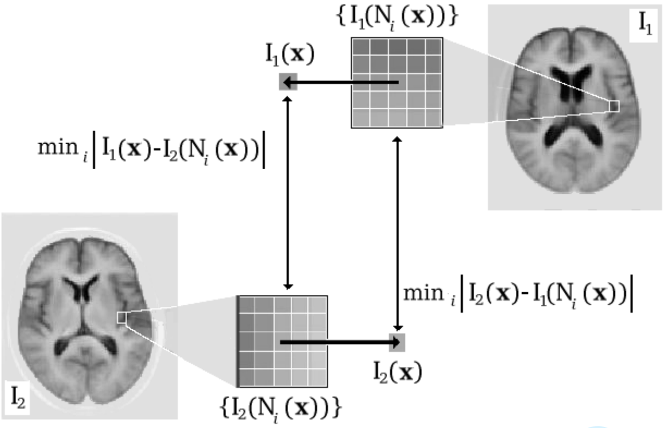

$\{I_1(N_i(\mathbf{x}))\}$
I_1
$I_1(\mathbf{x})$
$\min_i |I_1(\mathbf{x})\text{-}I_2(N_i(\mathbf{x}))|$
$\min_i |I_2(\mathbf{x})\text{-}I_1(N_i(\mathbf{x}))|$
$I_2(\mathbf{x})$
I_2
$\{I_2(N_i(\mathbf{x}))\}$

Fig. 5. The calculation of a shuffle difference image

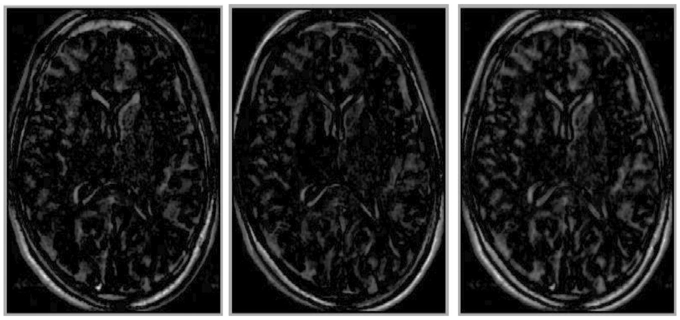


Fig. 6. Examples of the shuffle difference image: from first to second (left), from second to first (centre), and the symmetrical shuffle difference image (right)

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.

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 in both directions between a pair of images. We found, however, that the improvement obtained was not significant, and did not justify the increased computation time. In what follows, we use the asymmetric shuffle distance.

IV. Experimental Validation

We performed two sets of experiments, one designed to validate our model-based approach for evaluating NRR, the other to demonstrate its use in a practical application. In the first set of experiments our aim was to show that Specificity and Generalisation are valid measures of the degree of misregistration of a group of images. We took a set of registered images for which ground-truth labels were available, and applied a series of deformations which introduced progressively increasing misregistration. This allowed us to investigate how our measures of Specificity and Generalisation varied, as a function of the known misregistration. We also measured generalised overlap, using the ground-truth labels, to provide a comparison with an existing method. In the second set of experiments our aim was to demonstrate that we could usefully discriminate between different NRR algorithms, by comparing results for the same dataset.

A. Image Data

To conduct our experiments we used two different sets of MR images of the brain. The first, which we will refer to as the 'MGH Dataset' (see Acknowledgements), was a set of 2D transaxial mid-brain slices, extracted at an equivalent level from each of a set of affinely aligned T1-weighted 3D MR scans of $\mathcal{N} = 36$ normal subjects. As well as the images themselves, we had access to ground-truth data, in the form of dense (pixel by pixel) anatomical label maps for the gray and white matter, the caudate nucleus, and the lateral ventricles. These labels were further divided into left and right hemispheres. The anatomical labels were obtained by manual annotation under conditions of rigorous quality control. An example image and the corresponding label maps are shown in Figure 7.

The set of images was non-rigidly registered using a Minimum Description Length (MDL) NRR algorithm [16], and

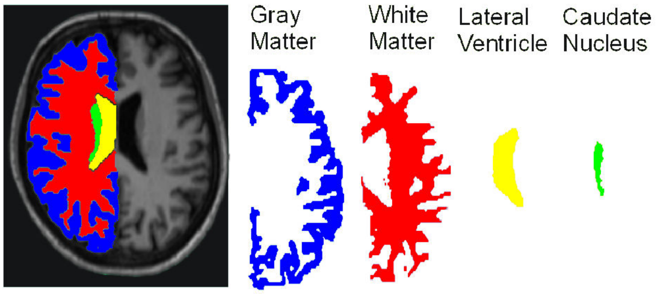


Fig. 7. An example affinely-aligned brain image and its accompanying anatomical labels, both overlaid and expanded, for gray matter, white matter, the lateral ventricles, and the caudate nucleus. The labels are also divided into left and right.

this registration was used as the starting point for a systematic evaluation of the effects of misregistration.

The second set of images, which we will refer to as the ’Dementia Dataset’, consisted of a set of 2D transaxial mid-brain slices, extracted at an equivalent level from each of a set of affinely aligned T1-weighted 3D MR scans of $\mathcal{N} = 104$ subjects entered into a clinical study of dementia.

B. Perturbing the Initial Registration

In order to perform a systematic evaluation of the effects of misregistration, we created multiple image sets, based on the MGH Dataset, but with controlled degrees of misregistration. To create a misregistered set, we took the original image set and applied a set of smooth pseudo-random spatial warps, based on biharmonic Clamped Plate Splines [20]. The warp for each image 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 degree of misregistration introduced. This provided a very general family of warps. We summarised the degree of misregistration by measuring d, the average magnitude of pixel displacement over the whole image. We generated a total of 70 misregistered image sets – 10 warp-set instantiations for each of 7 different values of d (0.0643, 0.249, 0.685, 1.36, 2.21, 2.76, and 4.15 pixels). Examples of warped images are shown in Figure 8.

C. Validation using Warped Images

Given the 70 image sets described above, each with known average misregistration, d, we investigated the relationship between d and Specificity, Generalisation, and Generalised Overlap, calculating the mean and standard error for each measure over the 10 warp instances for each value of d.

For each misregistered image set, we calculated Specificity and Generalisation, as described in Section III-A, using $m = 15$ modes of variation for the model and $\mathcal{M} = 1000$ synthetic images drawn from a Gaussian distribution, as described in Section II-C. This was repeated for values of shuffle radius, r, of 1 (Euclidean distance), 1.5, 2.1 and 3.7, as defined in Section III-B, corresponding to circular neighbourhoods contained within 1x1, 3x3, 5x5 and 7x7 pixel patches respectively. We also repeated these experiments with 2.5%, 5.0% and 10% Gaussian intensity noise added to the misregistered images, in order to investigate the sensitivity of the model-based measures to image noise.

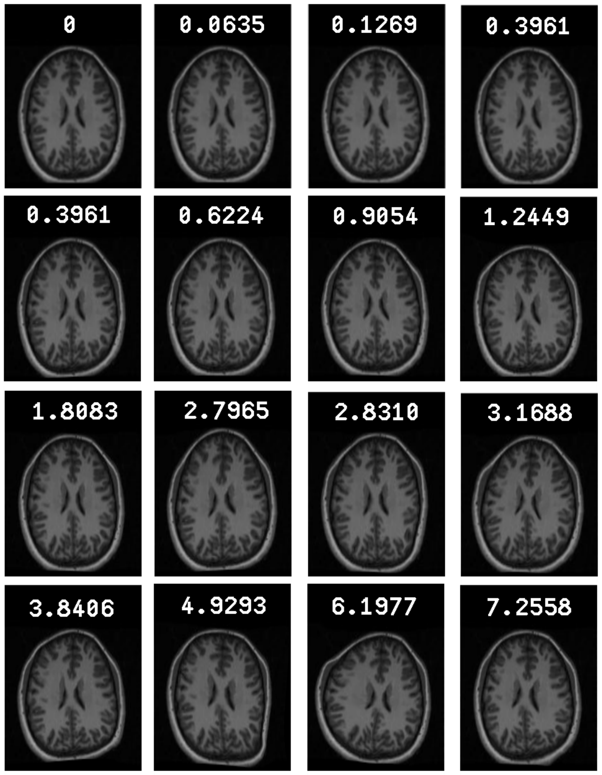


Fig. 8. An original image from the MGH Dataset (top left) and examples of warped versions of the same image obtained using different values of d, the mean pixel displacement (shown on each image).

Similarly, we calculated Generalised Overlap with volume, equal, inverse volume and complexity weightings, as defined in Section II-B.2 calculating the mean and standard error for each measure over the 10 warp instances for each value of d.

D. Sensitivity

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 standard error of those values. To quantify this, we define the sensitivity of a measure as follows.

$$D(m;d) = \frac{1}{\sigma_m}\left(\frac{m(d) - m(0)}{d}\right), \tag{12}$$

where $m(d)$ is the value of the measure for some degree of deformation d, σ_m is the standare error of the estimate of $m(d)$. $D(m;d) = 1$ 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. D is a function of d; to simplify comparison between different methods of evaluation, we also use the mean sensitivity over a range of values of d.

In order to compare the sensitivities of different methods of evaluation, we need also to estimate the expected error in D. Since the validation experiments provided repeated estimates of $m(d)$, we can obtain empirical estimates of the errors in

$m(d)$, $m(0)$, and σ_m. These can be combined, using error propagation, to estimate the uncertainty in the estimate of sensitivity.

E. Comparing Registration Algorithms

To illustrate the application of model-based evaluation in practice, we compared the NRR results obtained using three different methods for registering a group of images, as described in more detail below. We wished to establish whether it was possible, in a practical setting, to detect significant differences in performance between different NRR algorithms. All three registration methods used the same piecewise affine representation of image warps [23] and the same multi-resolution optimisation framework. The same number of iterations (function evaluations) were used in each case.

We applied the three registration algorithms to two datasets. The MGH Dataset was used because it allowed the evaluation results obtained using Specificity and Generalisation to be compared with an evaluation based on the Generalised Overlap measure (using ground truth). For these experiments $\mathcal{M}$ = 500 synthetic images were used to estimate Specificity and Generalisation. The Dementia Dataset was used because it was more representative of a typical clinical study, and we wished to demonstrate that our results were not dataset-specific. For these experiments we used $\mathcal{M}$ = 1000 synthetic images.

The three registration methods we used were as follows.

1) Pairwise Registration to a Reference: A commonly used approach to registering a group of images is to register each image in the group in turn to a reference image chosen from the group, using a pairwise objective function (e.g., [12]). We used this approach as a baseline, with a sum of absolute intensity differences objective function (which gave slightly better results than sum of squared differences or mutual information).

Pairwise approaches to registration can produce reasonable correspondences, but suffer from the problem that the results obtained depend on the choice of reference. Refinements of the basic approach are possible, where the reference is initialised and updated so as to be representative of the group of images as a whole. It is important to note, however, that even in this case the correspondence for a given image is determined solely by the information in the image and the reference. More recently, there has been considerable interest in *groupwise* methods which aim to make more systematic use of the information in the complete set of images when establishing correspondence. The remaining two methods we tested fall into this category.

2) Groupwise Congealing Algorithm: Learned-Miller et. al. [24] originally introduced their 'congealing' algorithm for registering a set of hand-written digits. The aim was to avoid the arbitrary selection of a co-ordinate frame, by repeatedly registering each image with an evolving "average" model. Given the current set of transformed images (initially the original images), for each pixel position, i, the probability density function of intensities, v, at that position across the set of images, $p_i(v)$ is estimated. The objective function is then the sum of entropies of these distributions across the whole image, $F = \sum_i \int p_i(v)\mathbf{log}p_i(v)dv$. A set of image deformations were optimised to minimise this. In later work on registering sets of 3D medical images [25], the objective function was approximated by$\sum_j \sum_i \mathbf{log}p_i(v_{ij})$, where v_{ij} is the value of pixel i in deformed image j. During optimisation, each image was warped so as to bring pixels with similar intensities into correspondence across the set. We implemented this later approach.

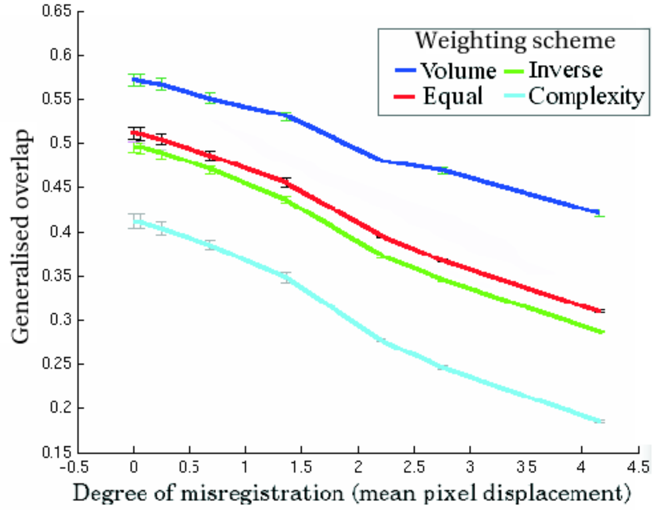


Fig. 9. Overlap measures (with corresponding ± one standard error errorbars) for the MGH dataset as a function of the degree of degradation of registration correspondence, d. The various graphs correspond to the various tissue weightings as defined in Section II-B.

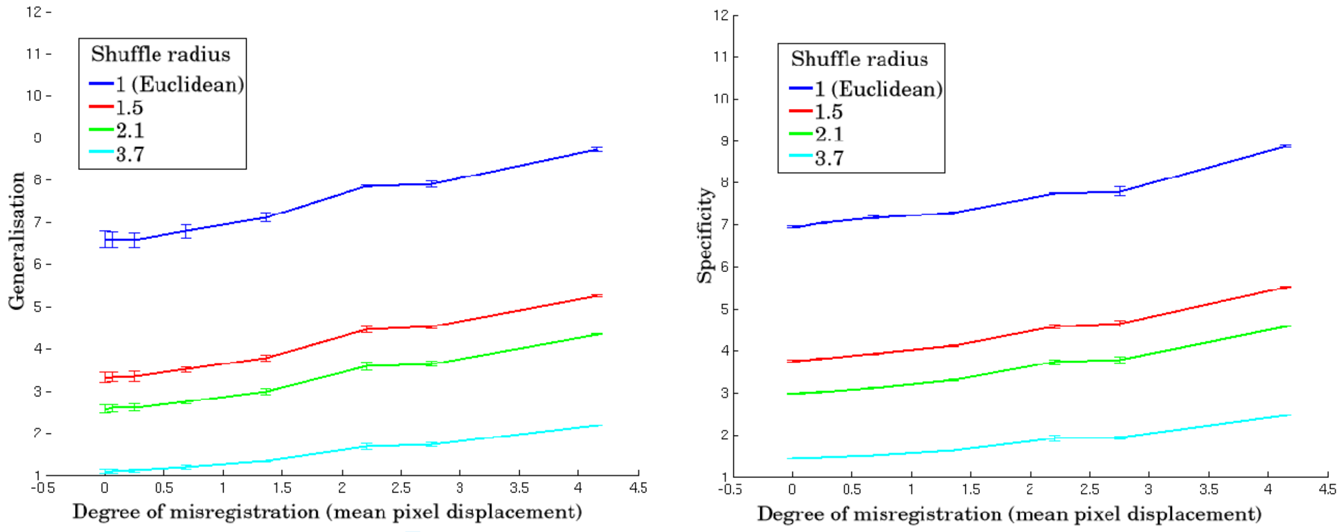
3) Groupwise MDL Algorithm: We have previously described a groupwise method which uses a Minimum Description Length (MDL)formulation [16]. The main idea is that the complete set of images can be encrypted as a coded message, and the description length [22] in bits used as an objective function. Rather than encoding the raw images, the encoding uses an appearance model, built using the estimated correspondences, to approximate the data; the encoding needs also to include details of the model itself and of the discrepancy between each image and its model approximation. As the registration proceeds, the correspondences, and hence the appearance model, are continually updated so as to minimise the description length.

V. RESULTS

A. Validation using Warped Images

Figure 9 plots each of the four variants of the generalised overlap measure, as a function of d, the degree of misregistration. As expected, the value decreases monotonically with increasing misregistration, in each case. This shows that our two gold-standard measures of misregistration (mean pixel displacement and ground-truth overlap) are in agreement, which validates the experimental framework.

Similarly, Figures 10(a) & 10(b) plot Generalisation and Specificity as functions of d, for different values of shuffle radius r. The results are qualitatively similar to those obtained for generalised overlap, except that both measures *increase*



(a) Generalisation

(b) Specificity

Fig. 10. Generalisation & Specificity for various definitions of image distance (varying shuffle radius) with corresponding ± one standard error errorbars as a function of the degree of degradation of the registration correspondence d for the MGH dataset

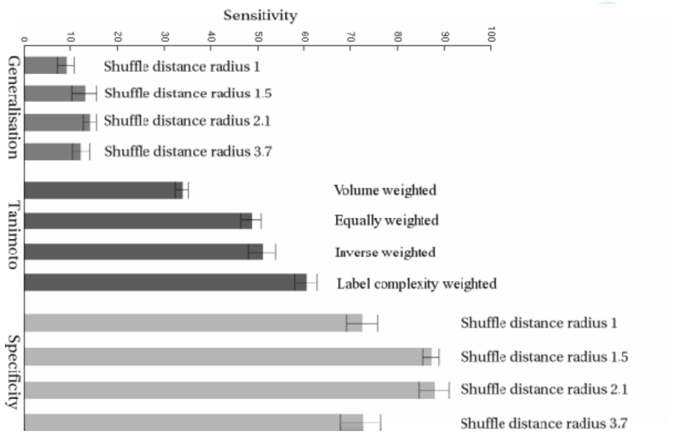


Fig. 11. Mean sensitivity of different NRR assessment methods over the full range of deformations d, shown with ±one standard error errorbars

monotonically with increasing misregistration, as expected (see Section III. These results show that, over the range of misregistrations investigated, the model-based measures are good surrogates for d, the mean pixel misregistration. Since the warps used to introduce controlled misregistration were of very general form, there is no reason to suppose that this result is dependent on the pattern of misregistration.

B. Sensitivity

Figure 11 shows the results of applying sensitivity analysis to the validation study. These demonstrate that Specificity is more sensitive (is able to detect smaller misregistrations) than the overlap-based approach, which is in turn more sensitive than Generalisation. Note from the error bars that these differences are statistically significant. Maximum sensitivity is achieved with a shuffle radius of 1.5 or 2.1. The most sensitive generalised overlap measure is obtained using label-complexity weighting.

C. Effect of Noise

We repeated the validation experiments and sensitivity analysis reported above with added image noise. Although the absolute values of the model-based measures were shifted upwards, as would be expected, there were no changes in the relative values, nor any systematic or statistically significant changes in sensitivity, even for 10% added noise.

D. Comparing Registration Algorithms

Figure 13 compares the performance of the three registration algorithms outlined in Section IV-E. All the measures tested in the previous section were computed, but we show results for only the most sensitive model-based method. Figures 13(a) and (c) show Specificity calculated using a shuffle radius of 2.1, for different values of m, the number of modes used to build the generative model. Figure 13(b) shows generalised overlap using different weightings. The results shown in Figure 13(a) suggest that the MDL groupwise approach gives the best registration result for the MGH Dataset, followed by Pairwise and Congealing in order of decreasing performance – irrespective of the value of m. Inspection of the error bars shows that these differences are statistically significant. The results for Generalised Overlap, shown in Figure 13(b), are more complicated, with the performance of the different NRR algorithms ordered differently for different weightings, though inspection of the error bars shows that many of the differences are not significant. Overall, the same general pattern emerges as for Specificity, with the Groupwise method generally best (statistically significantly in two cases), but with no significant difference between Pairwise and Congealing in most cases.

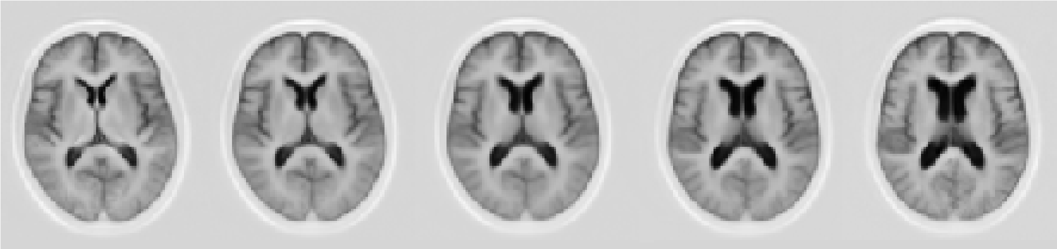


Fig. 12. The first mode (±2.5 standard deviations) of an appearance model built automatically by group-wise registration.

The results for inverse volume weighting generally lack significance, but are inconsistent with those obtained using the other weighting schemes. Volume weighting gives the best separation between the different variants, and places the three methods in the same order as Specificity. Overall, this supports the interpretation that Specificity give results that are generally equivalent to those obtained using Generalised Overlap, but with higher sensitivity. Finally, the Specificity results shown in Figure 13(c) for the Dementia Dataset, place the three methods in the same order.

VI. Discussion

The results of the validation experiment reported in Section V-A are the most important outcome of the work presented here. They demonstrate a causal relationship between our Specificity and Generalisation measures, and a known (up to an additive constant) mean pixel displacement, d. A strong correlation between these model-based measures and a Generalised Overlap measure, based on ground truth, adds further weight to this interpretation. The fact that the relationship with d held good over many different instantiations of a very general class of perturbing warps, makes it unlikely (though not impossible) that there is any significant pattern dependence.

The results obtained with added noise are also encouraging, since it is a reasonable concern that the use of an intensity-based distance measure might make the model-based measures sensitive to noise. In the event, the approach seems robust to quite significant levels of noise. The fact that the absolute values of specificity and generalisation change when noise is added, mean that they would not be useful for comparing registration results for different image sets. Their ability to compare the performance of different registration algorithms applied to the same set of images, the main intended use, is, however, unaffected.

Our results comparing the performance of different registration algorithms demonstrate that the model-based measures, and Specificity in particular, are sufficiently sensitive to misregistration to provide useful discrimination in a practical setting. There is, however, a potential concern that it is important to address. It might be argued that using a model-based approach to assessing registration favours methods which use a model-based objective function for registration (as in the experiments reported here). In practice, we do not believe that this is a problem.

First, as we have argued above, our validation results show that there is a causal relationship between the mean pixel displacement, d, and Specificity/Generalisation. It is thus irrelevant how a registration (or misregistration) has been obtained. Second, the MDL objective function we optimise in our model-based registration method measures a quite different property of the model to those we use in evaluation, so there is no element of ’self-fulfilling prophecy. In an ideal world it would, of course, be preferable to avoid even the possibility of bias, though it seems unlikely that one could devise a strategy for evaluation that had no relevance to achieving a good registration in the first place. We hope that, in due course, other ground-truth-free methods of evaluation will be developed, allowing a multi-perspective assessment of performance.

One obvious limitation of our approach to evaluation is that it can *only* be applied to groups of images. This could be considered an important restriction, since many practical applications involve registration of pairs or very short temporal sequences of images. We would argue that, in fact, this is a necessary restriction, because it is only possible to arrive at a meaningful assessment of registration in the context of a population of images.

The experiments we have reported were performed in 2D to limit the computational cost of running the 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. We have implemented the method in 3D and the time taken to evaluate the registration of 100 190x190x50 images using a shuffle radius of 2.1 and $\mathcal{M} = 1000$ is around 62.5 hours on a modern PC, which is short compared to most registration algorithms.

There are a number of issues that merit further investigation. We have studied a particular method of measuring image separation, but others, such as local correlation, would be worth exploring. Another interesting issue is whether it is possible within this framework to localise registration errors. We have performed some initial experiments, summing the shuffle difference maps between all pairs of images in the registered set. This gives some interesting results, highlighting areas of common misregistration, but it is not clear what quantitative interpretation could be placed on such maps. Finally, it is clear that our current measures of Specificity and Generalisation are not normalised – their values depend on the size of the set of registered images, the number of synthetic images generated and so on. We are currently exploring the possibility of measuring more fundamental properties of the relationship between the real and synthetic image distributions,

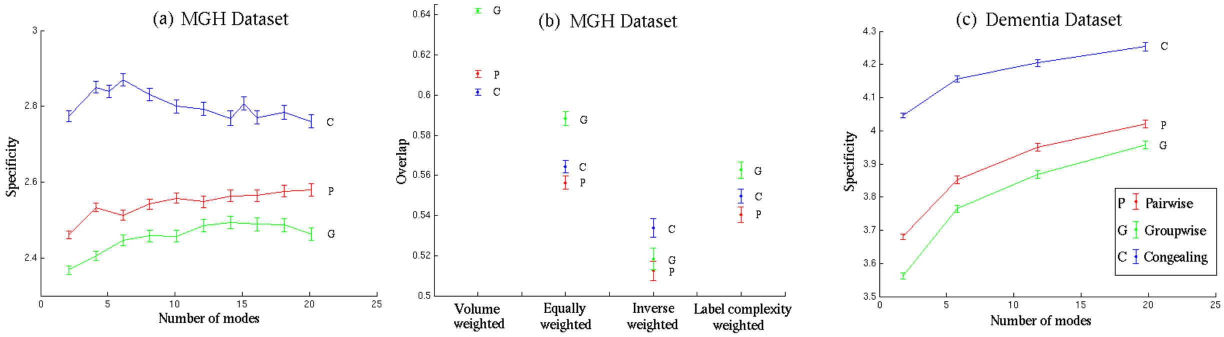


Fig. 13. **Left and right:** Generalisation and Specificity of the three registration methods as a function of the number of modes included in the appearance model.

with a view to achieving a ’natural’ normalisation.

VII. 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.

It is important to emphasise that our approach is not restricted to evaluating model-based NRR algorithms, though we presented results for one such method; the model-based measures of registration accuracy can be applied to any set of non-rigidly registered images, however they were obtained. We have discussed the possibility of a bias in favour of model-based methods of registration and conclude that there is no major problem, though it would be desirable to compare results obtained using a range of ground-truth-free methods of evaluation.

Acknowledgement

The authors would like to thank David Kennedy of the Center for Morphometric Analysis at MGH, for providing the fully-annotated brain images. Additional images from age-matched normals in a dementia study were generously provided by Prof. Alan Jackson, University of Manchester. We also acknowledge the interesting and constructive comments of the referees, which have helped us improve the paper.

References

[1] M. Beauchemin and K. P. B. Thomson, “The evaluation of segmentation results and the overlapping area matrix,” *International Journal of Remote Sensing*, 18(18):3895-3899, 1997.

[2] 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.

[3] B. Zitová and J. Flusser, “Image registration methods: A survey,” *Image and Vision Computing*, vol. 21, pp. 977 – 1000, 2003.

[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 images,” in IEEE 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.

[23] T. F. Cootes, C. J. Twining, V. Petrovic, R. Schestowitz, and C. J. Taylor, “Groupwise Construction of Appearance Models using Piecewise Affine Deformations.” in *Proceedings of the British Machine Vision Conference (BMVC’04)*, Kingston UK, 2004.

[24] E. G. Miller, N. E. Matsakis and P. A. Viola, “Learning from one example through shared densities on transforms,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Volume 1, 2000, pp. 464-471.

[25] L. Zollei, E. Learned-Miller, E. Grimson and W. Wells, “Efficient population registration of 3D data,” in *Workshop on Computer Vision for Biomedical Image Applications: International Conference of Computer Vision (ICCV05)*, 2005.

Response to Reviewers

Evaluating Non-Rigid Registration without
Ground Truth - TMI-2006-0096

We found the reviewer's comments helpful and thought-provoking. We hope that in responding we have significantly improved the paper. We felt able to address virtually all the points which were raised. In what follows the review is reproduced in italics, with our comments interspersed.

Associate Editor

The consensus of the reviewers is that the paper presents an interesting and valuable contribution. The primary concern raised, that must be addressed in a compelling way in preparing a revised version of the manuscript, is the issue of circularity in the assessment. This is the concern that the strategy for validation utilizing an AAM objective function finds an AAM registration method to be the best.

We accept that this is an important issue, and have addressed it in several ways, including both new experiments and revised text, as detailed below.

Reviewer 1

This paper presents an automatic method to evaluate the quality of elastic registrations without using manual segmentation or artificial data. The paper is dealing with a relevant issue since there is no real validation of non-linear registrations remains a controversy. The main idea of this work consists in transforming the evaluation of the registration into the evaluation of an appearance model of the aligned images. The appearance model is based on the variation of shape and texture, which are a priori considered as correlated. The quality of the model is evaluated regarding its ability to represent all the dataset called the generalisation, and also how close the model is from the dataset called the specificity.

A definition of the specificity and generalisation are given, based on the pseudo-distance between images called shuffle distance, supposed to be less sensitive to small (radius r) misalignments. A first experimental validation is achieved on 2D images which verifies the three measures (classical overlap, generalisation, specificity) monotonically increase when the misalignment increases. The sensitivity of the 3 measures are also compared. The specificity measure, which appears to be the most sensitive measure even before the overlap measure, and far before the generalisation measure which appears not very useful. Then the 2 measures of misalignments based of the appearance model (specificity and generalisation) are tested on a real dataset of 2D images study using one Minimal Description Length registration methods 3 different group registration strategies: pairwise, groupwise without and with constrained spatial deformations. The specificity of the model derived from the both groupwise registration strategies appears significantly better than from the pairwise approach. The generalisation shows no differences between strategies. The authors conclude by precising that any registration method can be tested with this model.

The paper is clearly written. Figures are useful and understandable. There are lots of different ideas and notions discussed. Because of that, the main message is sometimes hidden by mathematical equations, and conversely some explanations are not complete enough (texture representation, probability density function, measure of generalisation directly introduced in its discrete form).

The paper has been extensively rewritten from Section IV onwards. We hope that it is now clearer. The texture representation and form of distribution are now explained in Section IIC. The motivation for the definition of Generalisation is expanded Section IIIA.

The authors are talking in the long introduction about the 2 different ways of validating a non-rigid registration: the measure of overlap of manually segmented regions and the distance of the estimated transformation to an artificially applied one. The latter method is not further discussed although an artificial dataset is used. Actually, this measure is implicitly represented by the deformation factor d. What is validated is thus the consistency of the 3 measures (overlap, generalisation, specificity) with this d factor, which can be understood as the distance between the resulting registered image and the reference (d=0). This could be mentioned.

This was a helpful comment. We make the role of the artificially introduced deformation as a gold standard clearer throughout the paper, which strengthens the story we can tell about the results of the validation experiments.

The distance between images is basically a robust difference of intensities. The normalisation of intensities between images might be a critical point for the measure. A more general measure could thus be for instance the correlation coefficient between 2 blocks centered on the voxels that are compared in each image.

The model can in fact deal with simple (linear) intensity transformations, and this is explained in Section IIC. We agree that other measures such as local correlation could usefully be investigated and mention this in the Discussion (Section V1).

The use of "NRR" seems sometimes to mean the actual registration method used (MDL), and sometimes it means the registration strategy for a group of images (pairwise, groupwise). Also Non Rigid should theoretically be non affine (the images are initially registered using affine registration) but this terminology (as well as non-linear) is accepted.

We hope this confusion has been removed.

The conclusion of the article is that the specificity measure proposed is more sensitive than the classical overlap measure of manually segmented regions. That does not really mean that the proposed measure estimates better the anatomical matching, but rather that it takes into consideration the whole image. We can wonder if the method is not sensitive to outliers too, or intersubject variabilities outside of the anatomical regions of interest.

We agree with the sentiment in the second sentence. The question about noise is important. We have now run an additional set of experiments with added noise, described in Sections IVC and VC. The results show that the method is relatively insensitive to noise.

The practical scheme used for the registration study is not clear and could be precised: what pdf used, is there a MonteCarlo method used, how many random images generated?

The methods section (IV) has been completely rewritten and is hopefully clearer.

A critical point not really discussed in the paper is the number of images necessary to estimate a reliable model. Two relatively big datasets with 36 and 104 images are tested. It seems not to be applicable to the evaluation of the registration for only 2 images for instance.

We now discuss the restriction to groups of images in the Discussion (Section V1).

In conclusion, this paper is interesting and useful to evaluate the accuracy of registration of a large group of images, assuming that the computation time for 3D images is not prohibitive. My recommandation is acceptation with the revisions suggested.

We mention the time taken for a 3D evaluation in the Discussion (Section VI)

Specific corrections or suggestions: p2: Tanimoto's overlap coefficient: reference ? p3: "c" could be called "b" in reference to (2) p3: just before III: This is of one of the... p4: Fig 2: precise Ii represented by stars, and Imu represented by dots. p4: name the n-dimensional space and precise in the eq (6) p4: pdf stands for probability density function (not defined) p4: why M and N (not defined) are different? p5: N=36. Is N the same as p4? p7: Fig 11: which deformation coefficient d?

We believe we have dealt with all these detailed points**.**

Reference that could be mentioned (dual problem of evaluation of segmentations): Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004 Jul;23(7):903-21.

After some consideration we did not add this reference. Although the spirit of the STAPLE approach was an inspiration for this work, there is no direct similarity to our approach at a technical level.

Reviewer 2

This paper introduces a framework for evaluating nonrigid registrations with a ground truth, be it based on region overlap or known deformation fields. The technique instead analyzes the properties of a generative appearance model, which are demonstrated to be highly correlated with a region overlap measure, as well as sensitive to small misregistrations.

The work presented herein is highly original and extremely relevant for the difficult problem of validating nonrigid registration methods. Sections I-III are well written, but unfortunately, the same cannot be said for the experimental evaluation and discussion part. See details below. I have no doubt that, after suitable revisions, this will be an excellent and important paper. However, in my opinion it is just not quite there yet.

We accept this as entirely fair comment. We rushed to make the deadline for the special issue, and Sections IV onward did not receive the attention they should have. Those sections have now been completely rewritten and reorganized. We hope the reviewer will agree that we have now done a decent job.

The main flaw with the paper as it is right now is that it fails to address many of the constraints, assumptions, and limitations of the proposed method. Let us start with the results that suggest groupwise registration performed better than pairwise registration. This may be so, but the authors glance over the fact that their groupwise registration method itself uses an appearance model as part of its objective function. How is it surprising that an evaluation based on an appearance model finds this registration superior? In effect, you are validating a method (read: AAM-based registration criterion) against itself.

This hints to a fundamental limitation of validation methods, which is only shifted here but not resolved: any quality criterion other than the actual ground truth can also be used as a merit function to solve the problem that it was designed to evaluate. If this is done, the resulting method cannot be validated using this same criterion. In the present paper, this is not only a theoretical consideration, but the authors themselves, in my opinion, fell into this trap.

For the paper to be acceptable, I would suggest to either replace the AAM-based groupwise registration with a different algorithm, or demonstrate clearly, why the evaluation performed here is not a validation of a method against itself. As it is, I find it quite ironic that the authors state in Section IV, "the method is not restricted to evaluating model-based NRR algorithms", where on the contrary it appears that the method is fundamentally not capable of evaluating model-based algorithms.

We agree entirely that we glanced over this. We have now strengthened the paper in several ways.

1. In Section VA we take more care to emphasise what we believe can be deduced from the results of the validation experiments, in particular that Specificity (and Generalization) is a good surrogate for the degree of misregistration, *however it arises.*
2. As suggested we have implemented and evaluated another groupwise registration method (although plausible, unfortunately it turned out not to be very good!).
3. We have run the three registration methods on the MGH data as well as the Dementia data. This allows us to corroborate the result obtained using Specificity by using the ground truth to compute Generalised Overlap.
4. We now confront this issue directly in the Discussion (Section V1) and present what we hope is a persuasive but balanced argument.

We do not claim there is no issue, but believe that we make the case for the value of our approach.

Other issues I would like to see discussed are: can the proposed method localize registration errors, like comparison to a known deformation field can, and to some extent the overlap criterion can too? Can the method be applied to evaluate a single pairwise registration result? The authors rightfully point out that being able to perform in-line evaluations of NRR is an important property of their algorithm - yet, by far not every nonrigid registration situation is a groupwise problem. Also, how applicable is the proposed framework to situations that are not inter-subject registrations, e.g., temporal registrations?

These were helpful pointers to issues we should address. All these points are now dealt with in the Discussion (Section V1).

Another thing that had me wondering is the integration of transformations and textures in the appearance model. Only the transformation component is really influenced by the registration, whereas the texture is influenced, at least in part, by image noise. I am curious - how does the image noise level affect the performance of the proposed evaluation?

We have now run an additional set of experiments with added noise, described in Sections IVC and VC. The results show that the method is relatively insensitive to noise.

Detailed Comments

p. 1, right column, second paragraph - I don't think Ref. [3] is appropriate and relevant here, since it deals with point-based (and clearly not nonrigid) registration.

p 3, line 30 - "shape-free texture" - please explain briefly how this is obtained. Also, how is the "shape-free" property affected by the registration error?

p. 6, Eq (12) and after - how is $\bar\sigma$ computed? The explanation "mean error in the estimate of m" really doesn't tell me anything.

p. 6, l. 60 - "the uncertainties" - is this a statistical term? I have never heard of it. Is this standard deviations, ranges, certain percentiles? Please be more specific.

p. 7, Fig. 11 - please mention in figure caption what the errors bars represent

We believe that we have addressed all these points.