Data-Driven Evaluation of Non-Rigid Registration via Appearance Modelling

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

***Abstract*— This paper presents a generic method for assessing the quality of non-rigid registration (NRR) algorithms, that *does not* depend on the existence of any ground truth, but depends solely on the data itself. The data is taken to be a set of images. The output of any non-rigid registration of such a set of images is a dense correspondence across the whole set. Given such a dense correspondence, it is possible to build a generative statistical model of appearance variation across the set. Evaluating the quality of the registration algorithm is hence mapped to the problem of evaluating the quality of the resultant statistical model; that is, when the model is compared to the image data from which it was generated. It should be noted that this approach does not depend on the specifics of the registration algorithm used or on the specifics of the modelling approach used.**

We derive indices of model specificity and generalisation that can be used to assess the quality of such models. This approach is validated by comparing our assessment of registration quality with that derived from ground-truth anatomical labelling. We demonstrate that not only is our approach capable of reliably assessing NRR without ground truth, but it is also more sensitive than the ground-truth-dependent approach. Finally, to demonstrate the practicality of our method, different NRR algorithms – both pairwise and groupwise– are compared in terms of their performance on MR brain data.

I. Introduction

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

The aim of non-rigid registration is to automatically find a meaningful, dense correspondence across a pair (hence *pairwise* registration), or group (hence *groupwise*) 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. And different algorithms tend to produce slightly different results when applied to the same set of images [2] - there is a need for methods to assess the results of such registrations.

Various methods have been proposed for assessing the results of NRR [3]–[6]. One obvious approach is to compare the results of the registration to anatomical ground truth. However, this suffers from the problem that such ground truth is often difficult to obtain. For instance, expert annotation is time consuming, subjective, and very difficult in 3D. Other evaluation approaches involve the construction of artificial test data, which limits application to 'off-line' evaluation. Furthermore, such artificially generated and manipulated correspondence does not necessarily capture the type of deformation seen in real data. These problems motivate the search for a method of evaluation that does not depend on the existence of ground-truth data, or on making possibly unrealistic assumptions about the nature of the actual correspondence.

The method we will present here is based on the idea of constructing statistical models of sets of images, models which consider both the shape and texture variation of the objects imaged (appearance models). Such models have been extensively used as the basis for image interpretation by synthesis. The link between registration and modelling is given by the fact that the output of registration is a dense correspondence across the set of images. Such a set of correspondences is required to construct the shape and texture models [8], [9]. Varying the correspondence across a set varies the appearance model built upon this correspondence. The obvious corollary is that a better correspondence ought to produce a better appearance model. This allows use to map the problem of evaluation of registration to that of evaluating the model generated from the output of the registration.

The structure of this paper is as follows. We first give a brief description of the background to both the assessment of registration, and of the construction of appearance models, and explain in more detail the link between the two. We present quantitative measures which can be used to assess the quality of such models, hence of the registration upon which we will build such models. The behavior of these measures is investigated, and in particular, their behavior when compared to an assessment based on ground-truth data. Our validation results confirm our method to be in tight correlation with ground truth. Finally, we use the measures we have developed to compare various registration algorithms when applied to the registration of sets of 2D MR images of human brains. In particular, we are able to show the quantitative superiority of groupwise registration over a pairwise method.

[DRAFT Placeholder] Manuscript received February .., 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].

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 spatial deformation field between each pair of images, so that when one image is deformed onto another, corresponding structures are brought into alignment. Such non-rigid registration of medical images is a difficult problem, due to the size and complexity of cross-individual anatomical variation.

A typical registration algorithm proceeds by optimising some objective function. The objective function depends on, for example, the degree of deformation present in the spatial deformation fields defining the correspondence, and the image similarity that remains after the deformation has been applied. Also to be defined are the representation used for the deformation fields, and the method used for finding the optimum of the objective function. Varying any of these factors produces a different registration algorithm, which in general, tends to produce a slightly different resulting correspondence.

B. Assessment of Non-Rigid Registration

We here describe several commonly-used approaches to the problem of assessing the results of registration.

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 actual correspondence is known. Such test data is typically constructed by applying sets of known deformations (either spatial or textural) to actual images. This artificially-deformed data is then registered. The process of evaluation is based on comparison between the deformation fields recovered by the registration and those which have originally been applied [5], [6]. This type of approach can be used to test NRR methods 'off-line'. However, the validity of this method presumes that we have the ability to construct artificial deformations which are sufficiently close to the types of deformation seen in real-world situations. Furthermore, there are situations where such artificial data sets are a poor representation of the actual variation between images. For example, images taken from different subjects may display a much more complicated and extensive variation that that which can be simulated by such simple deformations.

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

Similar approaches involve measurement of the misregistration of anatomical regions of significance [3], [4], and the overlap between anatomically equivalent regions obtained using segmentation. This process is either manual or semi-automatic [4], [5]. Although these methods cover a general range of applications, they are labour-intensive and are often prone to errors. They also rely on one's ability to faithfully extract anatomical structures from the image intensities alone.

This paper explores one such method, which assesses registration using the spatial overlap. The overlap is defined using Tanimoto's formulation of corresponding regions in the registered images. The correspondence is defined by labels of distinct image regions (in this case brain tissue classes), produced by manual mark-up of the original images (ground-truth labels). A correctly registered image set will exhibit high relative overlap between corresponding brain structures in different images and, in the opposite case – low overlap with non-corresponding structures. A generalised overlap measure [7] is used to compute a single figure of merit for the overall overlap of all labels over all subjects:

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

where i indexes voxels in the registered images, l indexes the label and k indexes the two images under consideration. A_{kli} and B_{kli} represent voxel label values in a pair of registered images and are in the range $[0, 1]$. The $MIN()$ and $MAX()$ operators are standard results for the intersection and union of a fuzzy set. This generalised overlap measures the consistency with which each set of labels partitions the image volume.

The parameter α_l affects the relative weighting of different labels. With $\alpha_l = 1$, label contributions are implicitly volume weighted with respect to one another. This means that large labels contribute more to the overall measure. We have also considered the cases where α_l weights for the inverse labelled region volume (which makes the relative weighting of different labels equal), where α_l weights for the inverse label volume squared (which gives regions of smaller volume higher weighting) and where α_l weights for a measure of label complexity. We define label complexity rather arbitrarily as the mean absolute voxel intensity gradient in the labelled region.

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

C. Statistical Models of Appearance

There are many approaches to building statistical models of the appearance variation of objects which encompass the variation of both shape and texture that underlies such appearance variation. In particular, we use the generative appearance models as introduced by Cootes et al. [8], [9]. They have been applied extensively in medical image analysis [10]–[12], among other related domains, and successfully applied to brain morphometry, and also to the time-series analysis of cardiac data (e.g., [13]).

The construction of such an appearance model from a set of images depends on the existence of a dense spatial

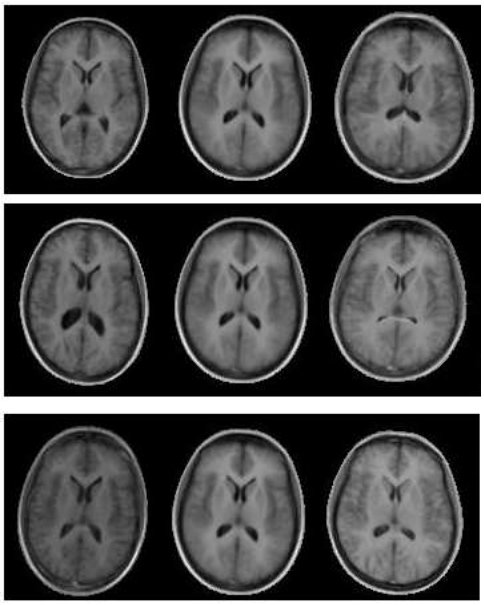


Fig. 1. The effect of varying the first (top row), second, and third model parameters of a brain appearance model by ± 2.5 standard deviations

correspondence across the set. In many manual or semi-automatic methods of model building, this dense correspondence is extrapolated and interpolated from the correspondence of some set of anatomically or user-relevant landmark points. In the automatic method that will be used here, the dense correspondence is given directly as the output of the NRR algorithm. Hence the relevant landmark positions in this case are in effect as dense as the pixels/voxels in the images registered.

In either case, the shape variation is represented in terms of the motions of these sets of landmark points. Using the notation of Cootes [8], 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 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) [14]. We hence obtain linear statistical models of the form:

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

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

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

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 one element of $\mathbf{c}$ for a model built from a set of 2D MR brain image is shown in Fig. 1.

In many cases, we wish to distinguish between the meaningful shape variation of the objects under consideration, and that apparent variation in shape that is due to the positioning of the object within the image (the pose of the imaged object). In that 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 details as regards the exact implementation of appearance models, see [8], [9].

As noted above, a meaningful dense groupwise correspondence is required before an appearance model can be built. One way to obtain such a correspondence is by extrapolating from expert annotation. However, this annotation process is extremely time-consuming and subjective, particularly for 3D data.

The output of groupwise NRR is such a correspondence, hence it was a natural next step to build automatic statistical models using the results of NRR algorithms [10], [11].

This link between registration and modelling is further exploited in the Minimum Description Length (MDL) [15] algorithm for non-rigid registration, where modelling becomes an integral part of the registration process. This latter work will be one of the registration strategies used later in this paper.

III. Evaluation Method

In the previous section, we described how the results of a non-rigid registration algorithm 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, hence for evaluating the quality of the non-rigid registration algorithm from which this model was derived. We also investigate several of the possible choices for model evaluation, the aim being to find one which is both robust, and gives the greatest sensitivity.

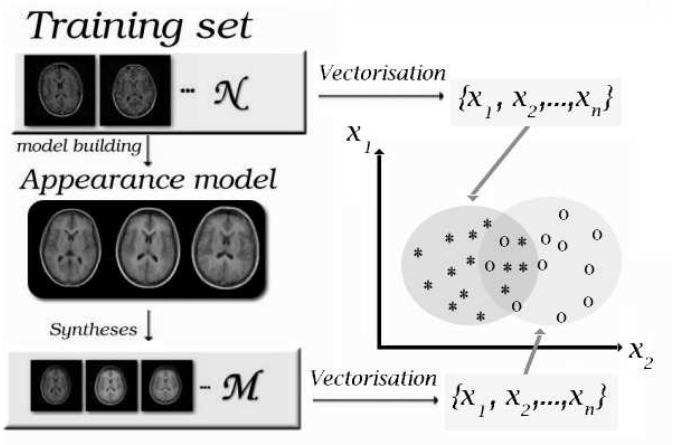


Fig. 2. The model evaluation framework: A model is constructed from the training set and then images are generated from the model. The training set of images and the set generated by the model can be viewed as clouds of points in image space.

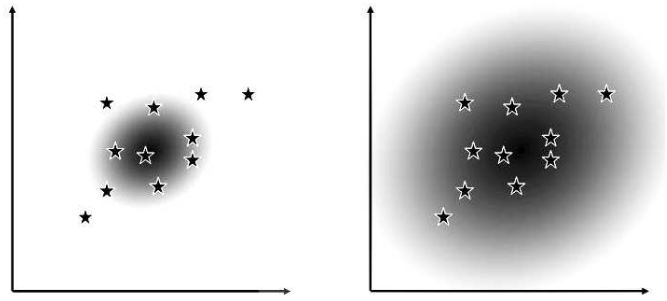


Fig. 3. Training set (points) and model pdf (shading) in image space. **Left:** A model which is specific, but not general. **Right:** A model which is general, but not specific.

A. Specificity and Generalisation

A good model of some training set of data should possess several properties. Firstly, the model should be able to effectively extrapolate and interpolate 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 [16]. We here present an extension of these quantitative measures.

Consider first the training data for our model, that is, the set of images which were the input to our NRR algorithm. Without loss of generality, each training image can be considered as a single point in image space (see Figure 2). A statistical model is then a probability density function $p(z)$ defined on this space. To be specific, let $\{I_i : i = 1, \dots N\}$ denote the N images of the training set when considered as points in image space. Let $p(z)$ be the probability density function of the model.

We then define our basic quantitative measure of the *specificity* S of the model with respect to the training set $\mathcal{I} = \{I_i\}$ as follows:

$$S_\lambda(\mathcal{I}; p) \doteq \int p(z) \min_{\text{w.r.t.}i} \left(|z - I_i|\right)^\lambda \, dz, \tag{6}$$

where $|\cdot|$ is a distance on image space, raised to some positive power λ. That is, for each point 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 $p(z)$. Greater specificity is indicated by *smaller* values of S, and lesser by *larger*. In Figure 3, we give diagrammatic examples of cases with varying specificity.

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

$$S_\lambda(\mathcal{I}; p) \approx \frac{1}{\mathcal{M}} \sum_{\mu=1}^{\mathcal{M}} \min_i \left(|I_i - I_\mu|\right)^\lambda, \tag{7}$$

with standard error:

$$\sigma_{\mathbf{S}} = \frac{\underset{\mu}{SD}\left\{\min_i\{|I_i - I_\mu|^\lambda\}\right\}}{\sqrt{\mathcal{M} - 1}}, \tag{8}$$

where $\underset{\mu}{SD}$ is the standard deviation of the set of measurements for the set of values of μ.

The measure of generalisation is then defined in an analogous manner:

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

with standard error:

$$\sigma_{\mathbf{G}} = \frac{\underset{i}{SD}\left\{\min_\mu\{|I_i - I_\mu|^\lambda\}\right\}}{\sqrt{\mathcal{N} - 1}}. \tag{10}$$

That is, for each member of the training set I_i, we compute the distance to the nearest-neighbour in the sample set $\{I_\mu\}$. 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, in what follows, we restrict ourselves to just the single nearest-neighbour case.

B. Distances 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. It 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

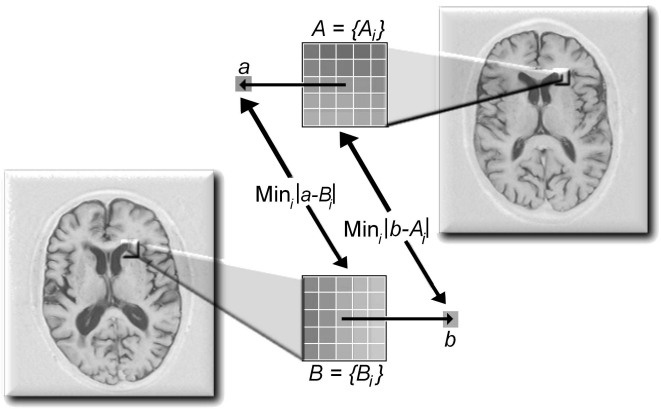


Fig. 4. The calculation of a shuffle difference image

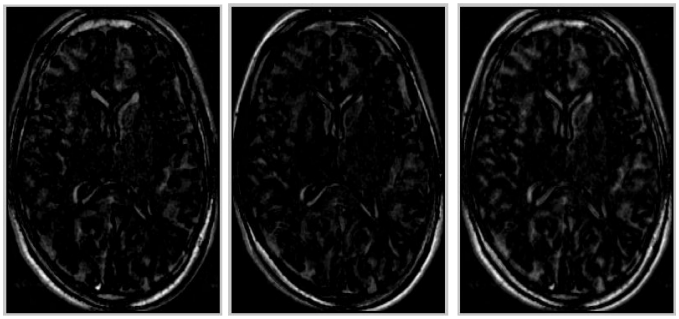


Fig. 5. Examples of the shuffle difference image: from one image to a second image (left), from the second image to the first (centre), and the symmetrical shuffle distance image (right)

does not provide a very well-behaved distance measure since it increases rapidly for quite small image misalignments [17]. This observation led us to consider an alternative distance measure, based on the 'shuffle difference', inspired by the 'shuffle transform' [18]. If we have two images $I_1(x)$ and $I_2(x)$, then the shuffle distance between them is defined as

$$D_s(I_1, I_2) = \frac{1}{n_p} \sum_x \min_{y \in N_r(x)} |I_1(x) - I_2(y)| \qquad (11)$$

where there are n_p pixels (or voxels) indexed by x, and $N_r(x)$ is the set of pixels in a neighbourhood of radius r around x.

The idea is illustrated in Figure 4. Instead of taking the sum-of-squared-differences between corresponding pixels, the minimum absolute difference between each pixel in one image and the values in a 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 6. It should be noted that the shuffle distance as defined above depends on the direction in which it is measured (see Figure 5), 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.

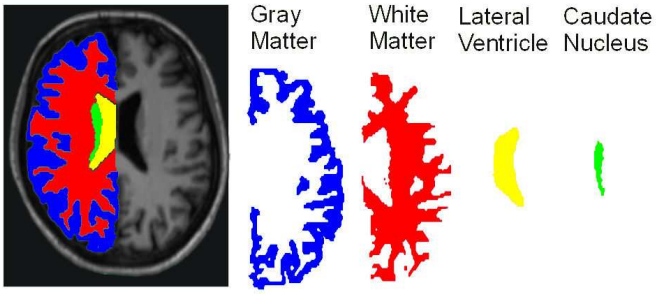


Fig. 7. An example affinely-aligned brain image and its accompanying anatomical labels, both overlaid and expanded, for gray matter, white matter, the lateral ventricles, and the caudate nucleus. Labels are also divided into left and right.

IV. Validation of the Approach

In this section, we present experiments which investigate the behaviour of our evaluation method. The principal idea is that of producing perturbed datasets, with progressive degrees of degradation. Ideally, our specificity and generalisation measures should vary monotonically with the degree of degradation.

The sensitivity of our method, that is, the degree of change that it can reliably detect, is an important issue, which is further explored in Section V-A.

A. Brain Dataset with Ground Truth

Our initial dataset consisted of $\mathcal{N} = 36$ transaxial midbrain 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 consisted of dense (pixel by pixel) binary tissue labels, the tissue classes being gray and white matter, the caudate nucleus, and CSF within the lateral ventricles. These labels were further divided into left and right. An example image and it's labelling is shown in Figure 7. The training set was non-rigidly registered using the Minimum Description Length (MDL) algorithm [15]. This registration was used as the starting point for the evaluation.

B. Perturbing Ground Truth

The evaluation now proceeded by considering perturbations about this found registration.

A test set of different registrations was created by applying smooth pseudo-random spatial warps (based on biharmonic Clamped Plate Splines [19]) 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.

Overall, the above approach was applied 10 times using 10 different random seeds. The 10 different warp instantiations were generated for each image and for each of seven progressively increasing values of average pixel displacement.

The perturbed correspondence across the set is then that given by applying the originally-found correspondence from

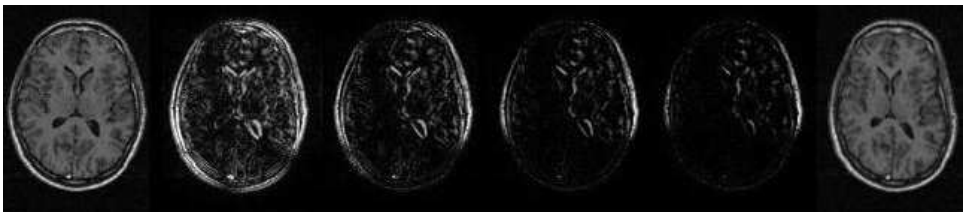


Fig. 6. A comparison between shuffle distance using varying size neighbourhoods (radius r). **Left:** original image, **right:** warped image, **centre, from the left:** shuffle distance with $r = 1$(Euclidean), $1.5, 2.9$ and 3.7 pixels.

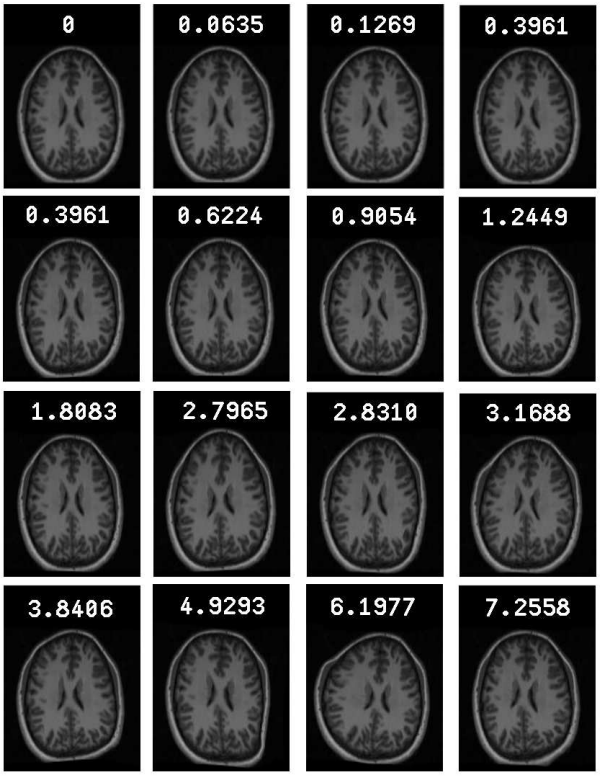


Fig. 8. Examples of registration degradation for increasing scales of smooth CPS warps. Mean pixel displacement for each image is shown.

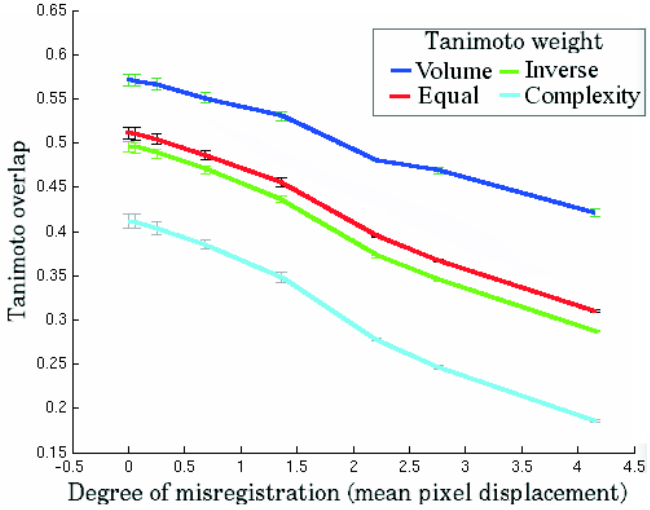


Fig. 10. Overlap measures (with corresponding errorbars) for the brain dataset as a function of the degree of degradation of the registration correspondence. The various graphs correspond to the various overlap measures as defined in section II-B.

the initial registration, but now applied to the *deformed* image sets. Hence the correspondence becomes progressively worse as the degree of image deformation increases.

C. Validation Results

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

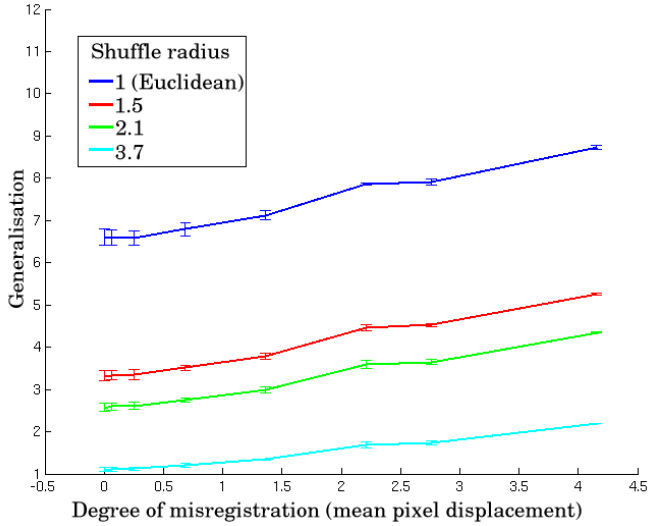
- **Tanimoto overlap** for the ground-truth data labels (1) for varying values of the label weighting α_l.
- **Specificity & Generalisation** ((7) & (9), $\lambda = 1$), for varying definitions of image distance (Euclidean and shuffle distances), and for varying values of the shuffle neighbourhood radius.

In Figure 10 are the results from the Tanimoto overlap-based measure (1), which computes a measure that is based on ground truth, that is, the overlap of the annotated labels. As can be seen from the Figure, all overlap measures decay monotonically as a function of misregistration, showing that our perturbed dataset does indeed have the systematic behaviour we require.

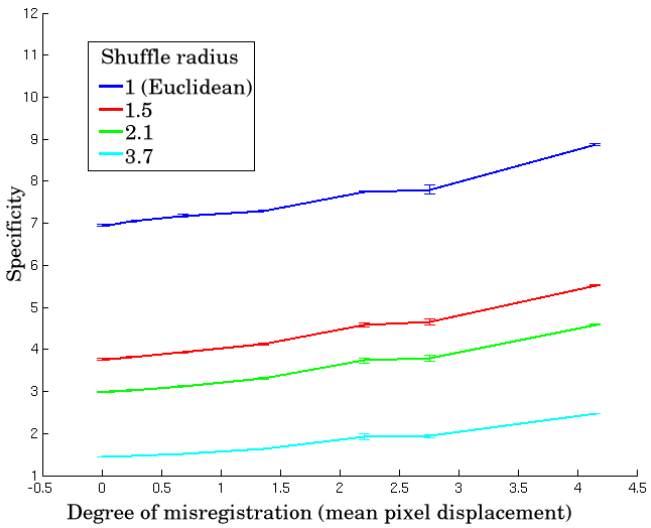
Results for the measures of specificity S (7) and generalisation G (9) as a function of the magnitude of the displacement are shown in Figures 9(a) & 9(b). Note that the values for Generalisation and Specificity are in error form, i.e. they increase with decreasing performance. The various graphs are for differing choices of the distance on image space, encompassing Euclidean distance, as well as shuffle distance for varying values of the shuffle neighbourhood radius r.

It should be noted that both measures show a monotonic decrease in performance with respect to the size of the registration degradation, for all choices of image distance. Since the overlap measure also shows such a monotonic decrease, this validates the model-based metrics inasmuch as they then also vary monotonically with respect to the ground truth measure.

What remains to be investigated are the effects of varying the various parameters in the definitions of the model-based measures. For the shuffle distance, the parameter is the neighbourhood radius r, the effect of which is studied in the next section. We also investigate the various forms of the Tanimoto overlap.



(a) Generalisation



(b) Specificity

Fig. 9. 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, that is, varying radius of the shuffle neighbourhood.

We note here that various other overlap measures as possible. For instance, we also considered the Dice overlap, but it was found to be inferior to the Tanimoto, and so is not considered further.

D. Measuring Sensitivity

As well as showing monotonicity, a good measure of registration quality should also sensitive. That is, it should enable us to measure small deviations from the optimum. If we can evaluate the sensitivity of a measure we will be able to fully compare the merits of various options.

The size of perturbation than can be detected in the validation experiments will depend both on the slope of the graphs of measure against degree of deformation, and also on the error on the measure. To quantify this, we define the sensitivity of a measure as follows.

Suppose $m(d)$ is the value of the measure for some degree of deformation d. We then define the measure sensitivity as:

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

where $\bar{\sigma}$ is the mean error in the estimate of m over the range. $D(m;d)$ is the change in d required for $m(d)$ to change by one noise standard deviation, which indicates the limit of changes in misregistration d which can be detected by the measure.

We computed the sensitivity for the data shown in Figures 10, 9(a), & 9(b). The averaged sensitivity over the range of deformations is plotted in Figure 11 for the various measures.

The first point to note is that there are statistically-significant differences between the various measures. Specificity is shown to be superior both to generalisation and most importantly, superior to the ground-truth based measure of Tanimoto overlap. Furthermore, we can see that shuffle radii of 1.5 and 2.1 for specificity give the most sensitive measure of all those studied.

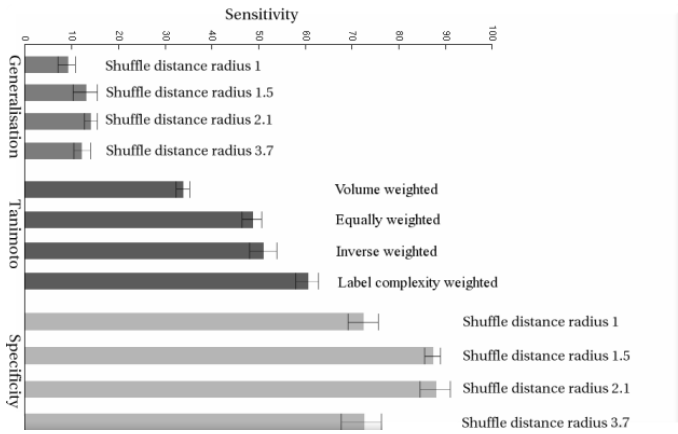


Fig. 11. Sensitivity of different NRR assessment methods

V. Assessing and Comparing Registration Algorithms

Having shown the validity of the model-based measures of registration quality, we now proceed to the reason for defining these measures, that is, to enable comparison of the performance of various non-rigid registration algorithms in cases where ground truth data is not available.

NRR algorithms can be divided into two general classes: *pairwise* and *groupwise*. Pairwise algorithms can be defined as those which register a pair of images. Registration across a group is then defined by successive applications of the basic pairwise algorithm. For example, all images in the training set can each be pairwise-registered to some chosen reference example (e.g., [11]). However, this suffers from the problem that, in general, the result obtained depends on the choice of reference. Refinements of this basic approach are possible, where the reference is artificially generated and updated so as to be representative of the group of images as a whole.

But the important point to note is that the correspondence for a single training image is defined w.r.t. this reference (which enables consistency of correspondence to be maintained across the group), and that the information used in determining the correct correspondence is limited to that contained in the single training set image and the single reference image.

It can be seen that this approach explicitly does not take advantage of the full information in the group of images when defining correspondence [20]. 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 [15]. 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 minimum description length [21] one, which envisages encrypting the entire training set as a coded message, the length of the message in bits being the objective function. But 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. The full encoding hence also has to contain the details of the model itself, and the discrepancy between the actual image and the appearance model representation of that image.

For the purposes of comparing NRR algorithms, we consider the following:

- 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, but admitting two slight variants of the algorithm.

In effect, the differences between the two groupwise variants considered are the way they constrain or do not constrain the allowed spatial deformations. The exact details are not relevant here; what is relevant is that both are groupwise, but with a slight difference.

We hence would expect that the groupwise variants should be close together in performance, but that both should give significantly better registration results than the simple pairwise approach. These three algorithms present a suitable test of the discrimination ability of our proposed evaluation framework.

For these evaluation experiments, we limited ourselves to 2D images, which allows larger-scale experiments to be performed.

The raw dataset consisted of $\mathcal{N} = 104$ 3D MR images of normal brains. [1]. These were then affinely aligned, and a single slice extracted from each, at equivalent locations. This hence formed our training set of 104 2D slices.

[1]The age-matched normals in a dementia study generously provided by Neil Thacker and Paul Bromiley, Manchester.

This training set was then registered using the 3 registration algorithms detailed above. For each algorithm, an appearance model was then built from the found correspondence, with varying numbers of modes included in the model. An example of such a model is shown in Figure 12. The Specificity and Generalisation of each such model was then computed. The results as a function of the number of modes are shown in Figure 13.

A. Results

The first point to note about the results shown in Figure 13 is, that as we might have expected from the results shown in Figure 11, Generalisation G is not able to discriminant between the three NRR algorithms, having insufficient sensitivity. Specificity, however, as we might have expected from its superior sensitivity, can discriminate between the pairwise and groupwise methods; both groupwise registrations give lower values of the specificity measure than the pairwise algorithm. This difference persists as we vary the number of model modes and is statistically significant. We can conclude that either of these groupwise algorithms is superior to the pairwise algorithm.

There is possibly a slight difference when the two groupwise methods are compared, since one graph tends to lie lower than the other. However, when we compare this difference to the size of the error bars on the points, it is not large enough for us state that there is a statistically significant difference between the two groupwise variants.

VI. Discussion and Conclusions

We have described a model-based approach to assessing the accuracy of non-rigid registration of groups of images. The most important thing about this method is that it does not require any ground truth data, but depends only on the training data itself.

Validation experiments were conducted, based on perturbing correspondence obtained through registration. These show that our method is able to detect increasing mis-registration using just the registered image data. The results obtained for different sizes of shuffle neighbourhood show that the use of shuffle distance rather than Euclidean distance improves the range of mis-registration over which we can detect significant changes in registration accuracy.

More importantly, we have shown that what is being measured by our model-based approach varies monotonically with an overlap measure based on ground truth. And not only that, we have shown that in the case considered here, the model-based measure of specificity is in fact of greater sensitivity

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

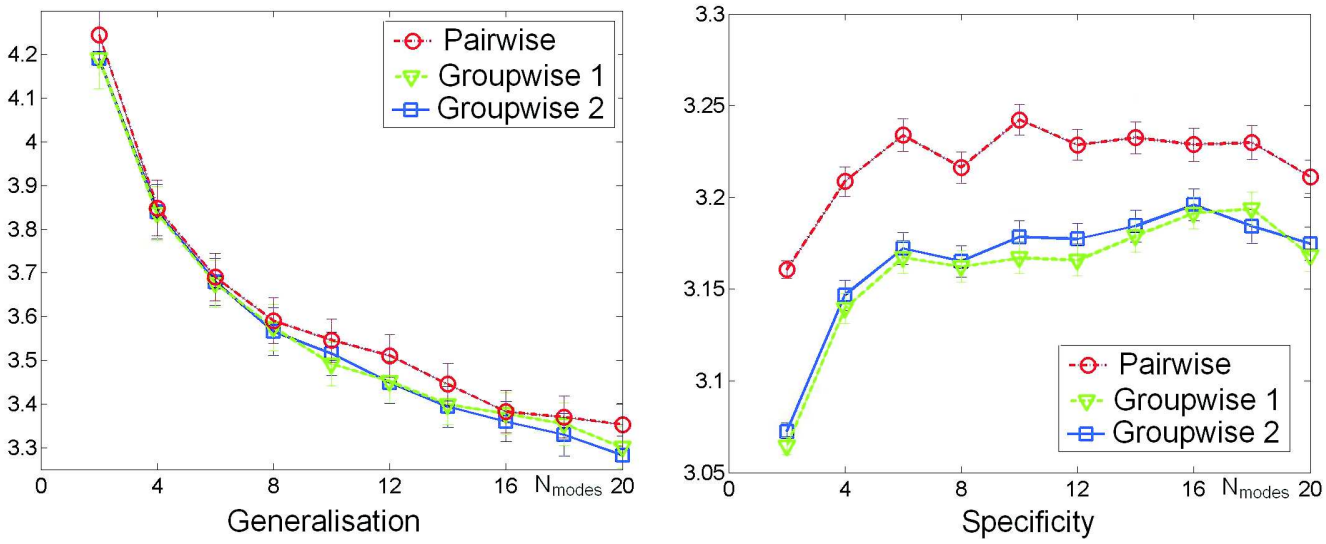


Fig. 13. Generalisation and Specificity of the three registration methods as a function of the number of modes included in the appearance model.

than the overlap measure based on ground truth, hence can reliably detect smaller differences in registration performance.

Finally, we have applied our model-based measure to assessing the quality of 3 different registration algorithms. The results obtained were in agreement with the results obtained during the validation phase as regards the relative sensitivity of the two model-based measures. We were able to show a quantitative improvement in performance of groupwise registration algorithms when compared to repeated pairwise registration.

We note that the experiments were conducted in 2D, which allowed larger-scale experiments to be conducted. However, the extension to 3D or higher is trivial, the only issue being that for higher-dimensional images, the calculation of shuffle distances (if used), will considerably increase the computational load.

In the above we used linear appearance modelling in our evaluation, but in principle, any generative model-building approach could be used. This method is totally general, and can be applied to the results of any registration algorithm.

This model-based method represents a significant advance as regards the important problem of evaluating non-rigid registration algorithms. It establishes an entirely objective basis for evaluation, since it is free from the requirement of ground truth data.

ACKNOWLEDGEMENT

The authors would like to thank David Kennedy of the Center for Morphometric Analysis at MGH, for providing the fully-annotated brain images. The work was done as part of the EPSRC/MRC funded MIAS-IRC grant, and the EPSRC IBIM grant (GR/S82503/01) . An EPSRC grant (GR/S48844/01) for Oscar Camara helped support studies that were based on ground truth.

REFERENCES

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

[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] 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.

[6] J. A. Schnabel, C. Tanner, A. C. Smith, M. O. Leach, C. Hayes, A. Degenhard, R. Hose, D. L. G. Hill, and D. J. Hawkes, "Validation of non-rigid registration using finite element methods," in *Lecture Notes in Computer Science*, vol. 2082. Springer, 2001, pp. 344–357.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] M. B. Stegmann, "Analysis of 4d cardiac magnetic resonance images," *Journal of The Danish Optical Society*, vol. 4, pp. 38–39, 2001.

[14] I. Joliffe, *Principal component analysis*. New York: Springer, 1986.

[15] 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 Informa-*

tion Processing in Medical Imaging (IPMI), Lecture Notes in Computer Science, vol. 3565. Springer, 2005, pp. 1–14.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. World Scientific Press, 1989.