Assessing the Accuracy of Non-Rigid Registration With and Without Ground Truth

R. S. Schestowitz[1], W. R. Crum[2], V. S. Petrovic[1],
C. J. Twining[1], T. F. Cootes[1] and C. J. Taylor[1]

[1]Imaging Science and Biomedical Engineering, University of Manchester
Stopford Building, Oxford Road, Manchester M13 9PT, United Kingdom

[2]Centre for Medical Image Computing, Department of Computer Science, University
College London, Gower Street, London WC1E 6BT, United Kingdom

Abstract

We present two methods for assessing the performance of non-rigid registration algorithms. We also show that assessment can be carried out with and without the need for some form of ground truth. One method utilizes a measure of overlap among data labels. The other method exploits the fact that, given a set of non-rigidly registered images, a generative statistical appearance model can be constructed. The quality of the model depends on the quality of the registration, and can be evaluated by comparing images sampled from it with the original image set. We derive indices of model specificity and generalisation, as well as introduce a formulation for overlap of anatimal label. We show that all of them demonstrate the loss of registration as a set of correctly registered images is progressively perturbed. Finally, we compare the sensitivities of these methods.

1. Introduction

Non-rigid registration (NRR) of both pairs and groups of images has in recent years increasingly been used as a basis for medical image analysis. Applications include structural analysis, atlas matching and change analysis [5]. The problem is highly under-constrained and a host of algorithms [4, 18] that have become available will, given a set of images to be registered, in general produce different results.

Various methods have been proposed for assessing the results of NRR [8, 10, 15, 14]. Most of these require access to some form of ground truth. One approach involves the construction of artificial test data, which limits application to 'off-line' evaluation. Other methods can be applied directly to real data, but require that anatomical ground truth be provided, typically involving annotation by an expert. This makes validation expensive and prone to subjective error.

We present two methods for assessing the performance of non-rigid registration algorithms. One of the methods requires ground truth to be provided *a priori*, whereas the other does not. We compare both methods on a registration of a set of 38 MR brain images and show them to provide a robust evaluation of registration success. Moreover, we demonstrate that both methods are in fact closely correlated if not interchangeable.

2. Method

data description: xxxxx

The first of the proposed methods assesses registration as the spatial overlap, 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 [1] is used to compute a single figure of merit for the overall overlap of all labels over all subjects.

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

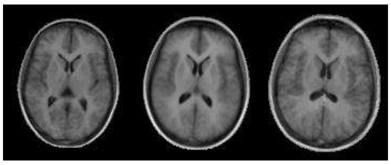

Fig. 1. The effect of varying the first model parameter of a brain appearance model by ± 2.5 standard deviations.

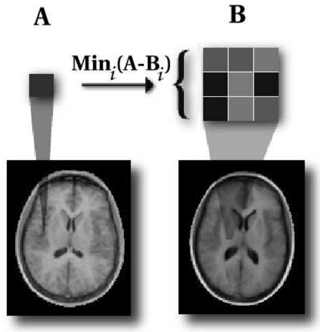

Fig. 4. The calculation of a shuffle difference image

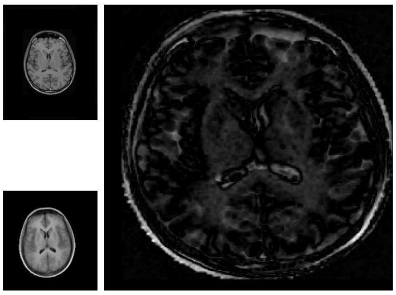

Fig. 5. An example of the shuffle difference image (right) when applied to two MR slices (left)

where i indexes voxels in the registered images, l indexes the label and k indexes the two images under consideration. A_{kli} and B_{kli} represent voxel label values in a pair of registered images and are in the range [0, 1]. The $MIN()$ and $MAX()$ operators are standard results for the intersection and union of a fuzzy set. This generalised overlap measures the consistency with which each set of labels partitions the image volume. The parameter α_l affects the relative weighting of different labels. With $\alpha_l = 1$, label contributions are implicitly volume weighted with respect to one another. We have also considered the cases where α_l weights for the inverse label volume (which makes the relative weighting of different labels equal), where α_l weights for the inverse label volume squared (which gives labels of smaller volume higher weighting) and where α_l weights for a measure of label complexity (which we define arbitrarily as the mean absolute voxel intensity gradient in the label).

The second method assesses registration as the quality of a generative, statistical appearance model, constructed from registered images. The idea is that a correct registration produces a true dense correspondence between the images, resulting in a better statistical appearance model of the images.

Registration is then evaluated through specificity and generalisation ability [17] of the model, or the ability of the model to i) generate realistic examples of the modelled entity and ii) represent well both seen and unseen examples of the modelled class. In practice these are evaluated by using generative properties of the model to produce a large number of synthetic examples (in this case brain images) that are then compared to real examples in the original set using some pre-defined image distance measure. Minimum distances of synthetic examples to examples in the original set and vice versa, give model specificity and generalisation respectively. Image distance is measured as a mean shuffle distance, or minimum Euclidean distance between a pixel in one image and a corresponding neighbourhood of pixels in the other.

To test the validity of the proposed methods, the brain images were annotated with 6 tissue classes including gray, white matter and CSF that provided the ground truth for image correspondence. Initially, the images were brought into alignment using an NRR algorithm based on the MDL optimisation. A test set of different registrations was then created by applying random perturbations to each image in the registered set using diffeomorphic clamped-plate splines. By choosing a different perturbation seed for each image and gradually increasing the magnitude of the perturbations, a series of image sets of progressively worse spatial correspondence and thus registration quality were obtained. By measuring the quality of the registration at each step, the proposed registration assessment measures can be validated.

3. Results

MIAS-IRC:

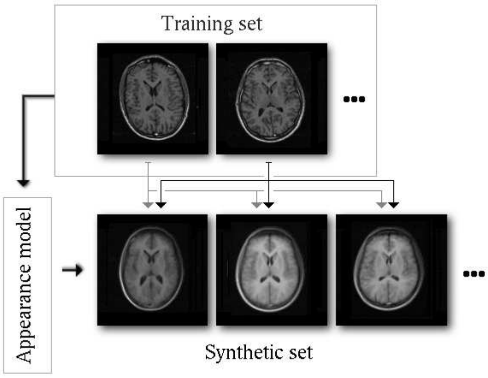

Fig. 3. The model evaluation framework. Each image in the training set is compared against every image generated by the model

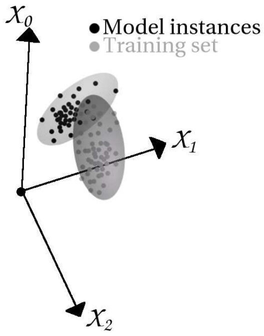

Fig. 2. Training set and model synthesis in hyperspace

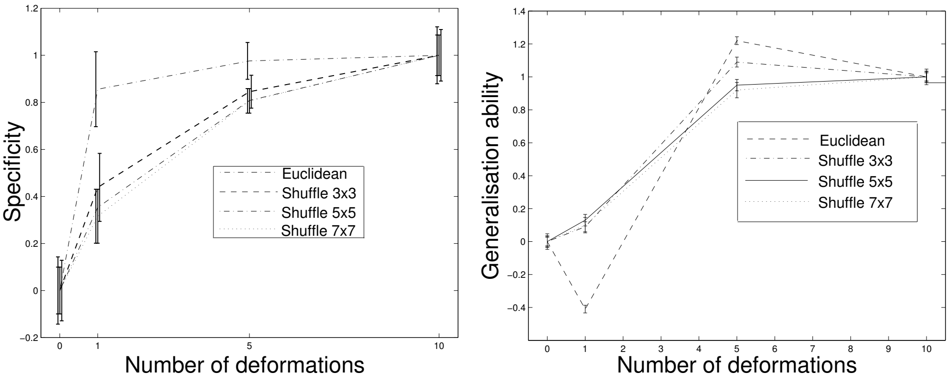

Fig. 6. Specificity and Generalisation for increasing mis-registration of different shuffle neighbourhood sizes.

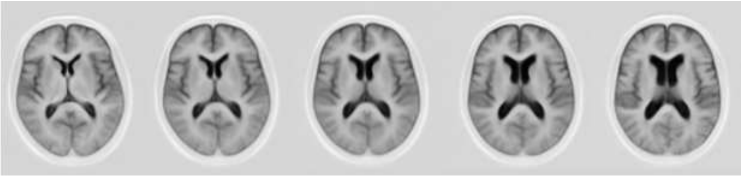

Fig. 8. Appearance model which was built automatically by group-wise registration. First mode is shown, ± 2.5 standard deviations.

Overall, the above approach was applied 10 times using 10 different perturbation seeds to ensure that both methods are consistent and results unbiased. Results of the proposed measures for increasing registration perturbation are shown in Figure 1. Note that Generalisation and Specificity plotted for different shuffle neighbourhood radius are in error form, i.e. they increase with decreasing performance. All metrics are generally well-behaved and show a monotonic decrease in registration performance. Such results directly validate the model-based metrics, which are shown be in agreement with the ground truth embodied in the region overlap based measure.

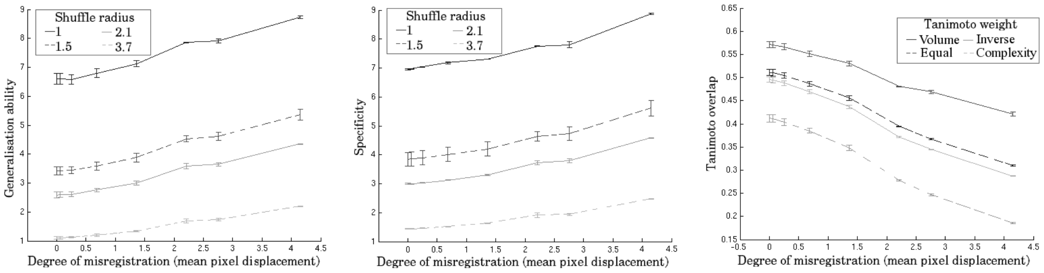

Figure 1. Behaviour of proposed metrics with increasing registration perturbation: a) Generalisation, b) Specificity and c) Tantimoto overlap

Finally, in order to obtain a quantitative comparison of the proposed algorithms we explore sensitivity of the proposed metrics, where the slighter the difference which can be detected reliably, the more sensitive the method. Sensitivity is in this case defined as the rate of change in the measure for a given perturbation range, normalised by the average uncertainty in the measurement over that range. More formally, sensitivity can be defined thus:

$$\frac{m - m_0}{d}/\overline{\sigma} \tag{2}$$

where m is the quality measured for a given value of displacement, m_0 is the measured quality at registration, d is the degree of deformation and $\overline{\sigma}$ is the mean over the error bars. Sensitivity is evaluated for all three of the proposed metrics and shown in Figure 2 with errors bars based on both an inter-instantiation error and a measure-specific error. The Specificity measure is the most sensitive for any radius of the shuffle distance followed by the overlap metric and Generalisation, with shuffle radii of 1.5 and 2.1 (equivalent to 3x3 and 5x5 neighbourhoods) giving optimal sensitivity.

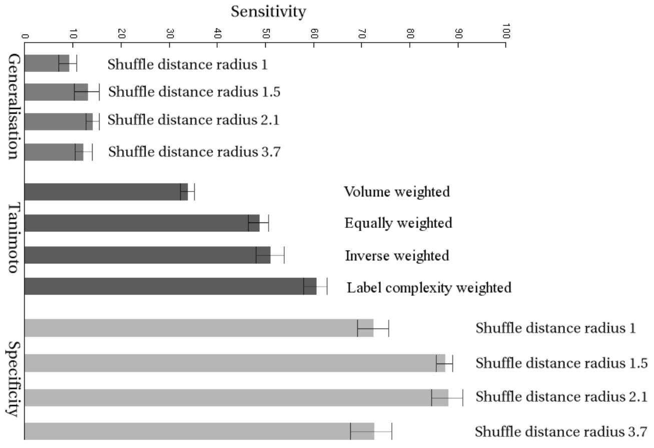

Figure 2. The sensitivity of registration assessment methods.

MICCAI:

The results of the experiment to test the effect of increasing mis-registration are shown in Fig. 6. These demonstrate that, for all sizes of shuffle neighbourhood, the specificity and generalisation values increase (get worse) with increasing mis-registration[1]. The results for different sizes of shuffle neighbourhood demonstrate that the range of mis-registration over which distinct values of specificity and generalisation are obtained increases as the neighbourhood size increases.

[1] Except that Generalisation is unstable for a 1x1 shuffle neighbourhood (Euclidean distance).

The results of the comparison between three different methods of NRR are shown in Fig. 9. These show that, particularly in terms of specificity, we can distinguish between the three approaches, with the fully groupwise method performing best, as anticipated. A model built using this approach is shown in Fig. 8.

4. Conclusions

MICCAI:

We have introduced a model-based approach to assessing the accuracy of non-rigid registration, without the need for ground truth. The validation experiments, based on perturbing correspondences obtained using ground truth, show that we are able to detect increasing mis-registration using just the registered image data. The results obtained for different sizes of shuffle neighbourhood show that the use of shuffle distance rather than Euclidean distance improves the range of mis-registration over which we can detect significant changes in registration accuracy. We have also shown that the approach is capable of detecting statistically significant differences in registration accuracy between three different (plausible) approaches to NRR.

We believe that this represents an important advance in the assessment of NRR, because it establishes an entirely objective basis for evaluating the reliability of NRR-based experiments, and for comparing the performance of different methods of NRR. The fact that no ground truth data is required means that the method can be applied routinely. Further work is needed to compare the results obtained using our new approach with those obtained using more sophisticated segmentation-based methods of evaluation.

MIAS-IRC:

The results shown in this abstract indicate that registration performance can be evaluated reliably both in the cases when ground truth information is available and when it is not. In particular, the methods based on generative statistical model evaluation are shown to be in agreement with the ground truth expressed through the true image region overlap metric based on the Tantimoto formulation. Proposed metrics are also shown to have sufficient sensitivity to detect very subtle changes in registration performance, on the level of perturbations measured in fractions of a pixel.

MICCAI 2005 SUBMISSION Appendage

5. Background

5.1. Assessing Non-Rigid Registration

One approach to assessing the results of NRR is to create a set of test images by taking original images and applying known spatial deformations. Evaluation involves comparing the deformation fields recovered by NRR to those known to have been applied [14, 15]. This approach can be used to test a given NRR method ’off-line’, but cannot be used to evaluate the results when the method is applied to real data as part of a registration-based analysis.

An alternative approach involves measuring the coincidence of anatomical annotations following registration. Variants of this approach include measuring the mis-registration of anatomical landmarks [8, 10], and the overlap between anatomically equivalent regions obtained using manual or semi-automatic segmentation [10, 14]. These methods are of general application, but are labour-intensive and error prone.

5.2. Statistical Models of Appearance

Statistical models of shape and appearance (combined appearance models) were introduced by Cootes, Edwards, Lanitis and Taylor [2, 3, 7], and have since been applied extensively in medical image analysis [9, 12, 16]. The construction of an appearance model depends on establishing a dense correspondence across a training set of images using a set of landmark points marked consistently on each training image.

Using the notation of Cootes [3], the shape (configuration of landmark points) can be represented as a vector $\mathbf{x}$ and the texture (intensity values) represented as a vector $\mathbf{g}$.
The shape and texture are controlled by statistical models of the form

$$\begin{aligned} \mathbf{x} &= \overline{\mathbf{x}} + \mathbf{P}_s\mathbf{b}_s \\ \mathbf{g} &= \overline{\mathbf{g}} + \mathbf{P}_g\mathbf{b}_g \end{aligned} \tag{3}$$

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

$$\begin{aligned} \mathbf{x} &= \bar{\mathbf{x}} + \mathbf{Q}_s\mathbf{c} \\ \mathbf{g} &= \overline{\mathbf{g}} + \mathbf{Q}_g\mathbf{c} \end{aligned} \tag{4}$$

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

6. Model-Based Evaluation

6.1. Specificity and Generalisationss

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

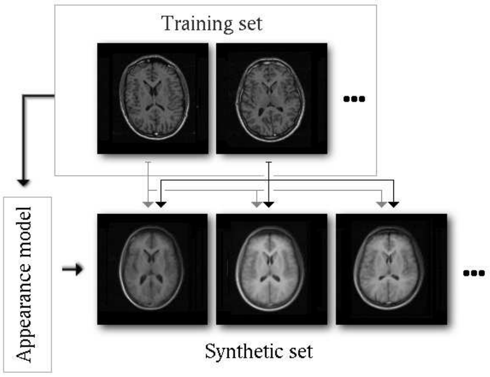

Fig. 3. The model evaluation framework. Each image in the training set is compared against every image generated by the model

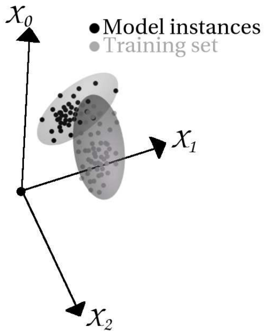

Fig. 2. Training set and model synthesis in hyperspace

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

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

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

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

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

Both values are low for a good model. Specificity measures the mean distance between images generated by the model and their closest neighbours in the training set, whilst Generalisation measures the mean distance between images in the training set and their closest neighbours in the synthesised set. The approach is illustrated diagrammatically in Fig. 3.

6.2. Measuring Distances in Between Images

The most straightforward way to measure the distance between images is to treat each image as a vector formed by concatenating the pixel/voxel intensity values, then take the Euclidean distance. Although this has the merit of simplicity, it does not provide a very well-behaved distance measure since it increases rapidly for quite small image misalignments. This observation led us to consider an alternative distance measure, based on the ’shuffle difference’, inspired by the ’shuffle transform’ [11]. The idea is illustrated in Fig. 4. Instead of taking the sum of squared differences between corresponding pixels, the minimum absolute difference between each pixel in one image and the values in a shuffle neighbourhood around the corresponding pixel is used. This is less sensitive to small misalignments, and provides a more well-behaved distance measure.

7. Validation of the Approach

7.1. Perturbing Ground-Truth

We conducted a series of experiments to test the hypothesis that reduced registration accuracy can be detected using model specificity and generalisation. An equivalent 2D mid-brain T1-weighted slice was obtained from each of 36 subjects using a 3D acquisition. A fixed number (167) of landmark points were positioned manually on the cortical surface, ventricles, caudate nucleus and lentiform nucleus, and used to establish a ground-truth dense correspondence over the set of images, using locally affine interpolation. A statistical appearance model was constructed using the methods described in 4.3, with the set of landmark coordinates forming

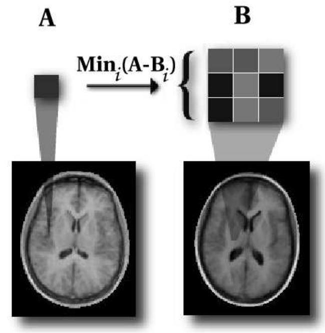

Fig. 4. The calculation of a shuffle difference image

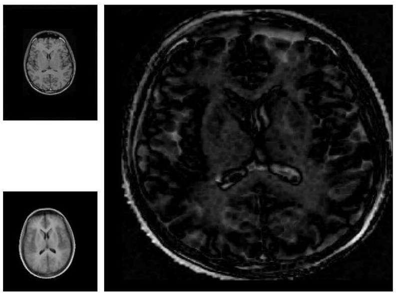

Fig. 5. An example of the shuffle difference image (right) when applied to two MR slices (left)

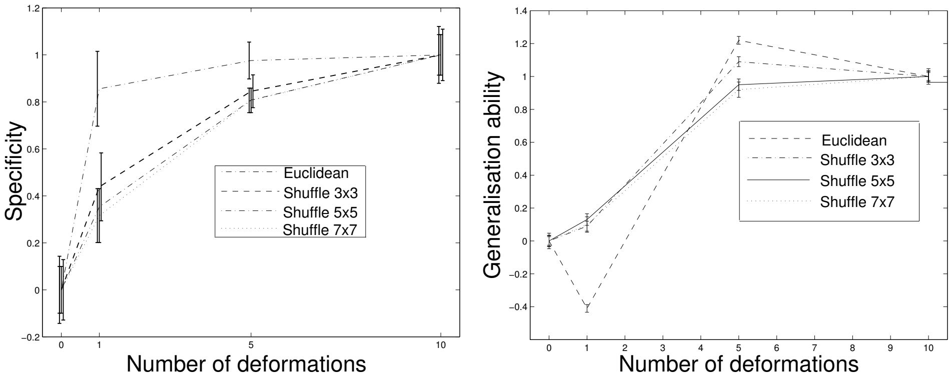

Fig. 6. Specificity and Generalisation for increasing mis-registration of different shuffle neighbourhood sizes.

the shape vector $\mathbf{x}$ for each image. Keeping the shape vectors fixed, we then applied a series of smooth pseudo-random spatial warps to the training images, resulting in successively increasing mis-registration. Each warp resulted in an average point displacement of between one and two pixels. Specificity and Generalisation results were obtained for 0, 1, 5, and 10 warps per image, using $m = 1000$.

7.2. Effects of the Shuffle Transform

The experiment described in the previous section was repeated for shuffle neighbourhoods of 1x1 (Euclidean distance), 3x3, 5x5, and 7x7, to test the hypothesis that this would extend the range over which different degrees of mis-registration could be discriminated.

7.3. Comparing Different Methods of NRR

A common task in medical image analysis is the estimation of correspondences across a group of images, to allow mapping of effects into a common co-ordinate frame when performing population studies. A widely used approach is to use a non-rigid registration algorithm to map a chosen reference image onto each example, defining the correspondence across the group [12]. However, it has been argued [4] that this *pairwise* approach does not take advantage of the full information in the group, and thus may lead to sub-optimal registration. We have been investigating *groupwise* methods of registration which aim to make the best use of the group as a whole when estimating the correspondence. We work within a minimum description length (MDL) framework. The aim is to construct a statistical appearance model which can exactly synthesize each example in the training set as efficiently as possible [17]. It has been observed that the more the compact the representation, the better the correspondences. The general approach is to define a deformation field between reference frame and each training image. For a given choice of sets of fields, one can compute the cost of encoding the images (a combination of the coding cost of the model, the cost of the parameters and the cost of residuals between the synthesized images and the training images). The effect on this total description length of modifying the deformation fields can be evaluated - the correspondence problem becomes a (very high dimensional) optimisation problem. Within this general framework we compare three different approaches (for details see [17]):

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

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

References

[1] 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. Proceedings of MICCAI 2005, LNCS 3749, pp 99-106.

[2] T. F. Cootes, C. Beeston, G. J. Edwards, and C. J. Taylor. A unified framework for atlas matching using active appearance models. In *Information Processing in Medical Imaging*, 1613:322-333, 1999.

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

[4] T. F. Cootes, S. Marsland, C. J. Twining, K. Smith, and C. J. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *European Conference on Computer Vision*, 2034:316-27, 2004.

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

[6] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor. A minimum description length approach to statistical shape modeling. *IEEE Transactions on Medical Imaging*, 21(5):525-537, 2002.

[7] G. J. Edwards, T. F. Cootes, and C. J. Taylor. Face recognition using active appearance models. In *European Conference on Computer Vision*, 2:581-595, 1998.

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

[9] A. F. Frangi, D. Rueckert, J. A. Schnabel, and W. J. Niessen. Automatic construction of multiple-object three-dimensional statistical shape models: application to cardiac modelling. *IEEE Transactions on Medical Imagin*g, 21:1151-66, 2002.

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

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

[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 Transactions on Medical Imaging*, 22(8)1014-1025, 2003.

[13] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, D. J. Hawkes. Non-rigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712-721, 1999.

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

[15] J. A. Schnabel, C. Tanner, A. Castellano-Smith, M. O. Leach, C. Hayes, A. Degenhard, R Hose, D. L. G. Hill, and D. J. Hawkes. Validation of non-rigid registration using finite element methods. In *Information Processing in Medical Imaging*, 2082:344-357, 2001.

[16] M. B. Stegmann, B. K. Ersboll, and R. Larsen. FAME - a flexible appearance modeling environment. *IEEE Transactions on Medical Imaging*, 22(10):1319-1331, 2003.

[17] C. J. Twining, T.F. Cootes, S. Marsland, S. V. Petrovic, R. S. Schestowitz, and C. J. Taylor. A unified information-theoretic approach to groupwise non-rigid registration and model building. To be presented in *Information Processing in Medical Imaging*, 2005.

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

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

ISBI Submission Plan/Structure

A. Introduction

— evaluating registration
— with or without ground truth
— paper validates and compares methods

LENGTH: 3/8 page

B. Background

B.1. Registration

— aim
— pair-wise, group-wise

B.2. Evaluating Registration

— need to evaluate it (rationale)

B.3. Label Overlap

— the concept

B.4. Appearance Models

— without ground truth
— construction using correspondences [ref]
— synthesis
— better registration ⇒ better model

LENGTH: 1 page

C. Method

C.1. Brain Data

describe - brain slices - IBIM

C.2. Overlap-based Assessment

— Tanimoto and Dice
— formulations

C.3. Specificity and Generalisation

— clouds in space
— necessary properties of a distance metric
— shuffle distance - illustrate, compare with Euclidean

C.4. Assessment Framework

— perturbation
— CPS, bending energy, warp directions (map), labels, randomisation

LENGTH: 1 1/2 page

D. Results

D.1. Overlap-based Measures

— curves

D.2. Model-based Measures

— more curves

D.3. Comparison of methods

— correlation, sensitivity plots

LENGTH: 7/8 page

E. Conclusions

— we have shown two valid methods for evaluating registration
— methods correlated in practice
— can evaluate registration
— can evaluate registration without ground truth