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

Abstract

The paper presents a generic approach, which us used assess the quality of appearance models. Moreover, it is capable of evaluating different non-rigid registration algorithms without any notion of ground truth. We base this approach on the observation that a statistical appearance model can be constructed from a set of non-rigidly registered images. Models can be evaluated by comparing images generated by it with the image set from which it was constructed. The quality of the model depends on the quality of its seminal registration, which means that registration can be evaluated by constructing and evaluating models. Indices are derived which reflect on model specificity and generalisation. It is then shown that are negatively affected as a set of correctly registered images is progressively perturbed. To demonstrate the practicality of these methods, different registration algorithms are compared in terms of performance.

1. Introduction

Non-rigid registration (NRR) is ubiquitously used as a basis for medical image analysis. Its applications include atlas matching, analysis of change [6], and structural analysis. A variety of approaches to NRR exist and they differ in terms of the objective function that defines mis-registration, the representation of spatial deformation fields, and the approach used to minimize the mis-registration by selecting good deformations. Most commonly, pairs of images are being registered [21], though groups can be considered too [5]. This under-constrained problem suffers from subjectivity of its solution. For any set of images to be registered, the different approaches are likely to produce different results. One obvious way to assessing solutions is by making use of the ground truth solution. Several methods have been demonstrated which work along these lines [9, 11, 18, 16]. These methods 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. As The correct solution is indeed hard to obtain, assessment without ground truth appears highly valuable.

Appearance model have been extensively used as the basis for interpretation by synthesis. Such models are derived from sets of training images and they capture statistics about variability in these sets. A set of images, which is used to construct an appearance model, is directly related to its quality. When the images are properly correspondent, the model is improved. As NRR aims to bring sets of images to correspondence, the output of a good NRR algorithm builds a good model.

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

The remainder of this paper covers background on models and registration, explains the methods, and presents results which support ideas being the method. An example is then shown where models are advertently degraded, by mis-registering their training set. Lastly, several registration algorithms are compared to demonstrate one main application of our approach.

2. Background

2.1. Assessment of Non-Rigid Registration

A common approach to assessment of the results of NRR involves the generation of test images. Such images are created by taking the original images and then applying known deformations to them. The process of evaluation is based on comparison between the deformation fields recovered by NRR and those which have originally been applied [16, 18]. This type of approach can be used to test NRR methods 'off-line'. It cannot, however, be used to evaluate the results when the method is applied to real data as part of a registration-based analysis.

Another approach involves measuring the overlap between of anatomical annotations before and after registration. Similar approaches involve measurement of the mis-registration of anatomical regions of significance [9, 11], and the overlap between anatomically equivalent regions obtained using segmentation, which is either manual or semi-automatic [11, 16]. Although these methods cover a general range of applications, they are labour-intensive and are often prone to errors.

2.2. Statistical Models of Appearance

Statistical models of shape and appearance (combined appearance models) were introduced by Cootes, Edwards, Lanitis and Taylor [2, 3, 8], and have since been applied extensively in medical image analysis [10, 14, 19]. 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

$$\begin{aligned} \mathbf{x} &= \overline{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s \\ \mathbf{g} &= \overline{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \end{aligned} \tag{1}$$

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

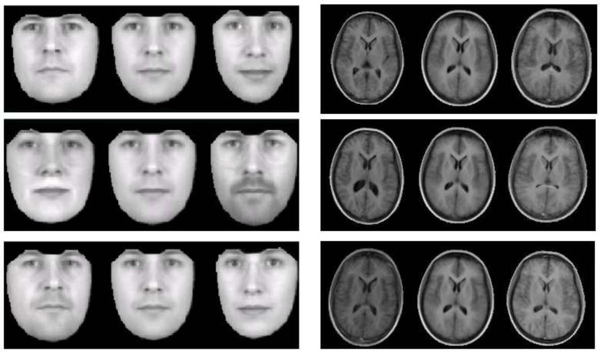

Fig. 1. The effect of varying the first, second, and third model parameters of a face and brain appearance models by ±2.5 standard deviations.

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

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.

2.3. The Correspondence Problem

A very key step in construction of combined appearance models is that of identifying dense correspondence across a given set of training images. This is often achieved by marking up the training set by hand, simply identifying significant points in the images and interpolating between these points. In recent years, automation of this process was a problem of great interest. One approach to solving this problem is to use NRR and bring the images to alignment by optimising a similarity measure [10, 14]. A different approach refines initial estimates of the correspondence so as to code the set of images in the most efficient way [1]. We have recently outlined an approach which is based on optimising the total description length of the training set, using its model [20].

In Section XX our approach is validated by deliberately perturbing the correspondence in models, i.e. decreasing the registration. Such models were built using manual annotation that establishes a reliable correspondence. In Section XX our approach is used to compare common registrations methods [10, 14], as well as our minimum description length approach.

3. Evaluation Method

This section presents the evaluation method which can assess registration in a model-based fashion and, more broadly, it explains the use of the approach when evaluating models of appearance.

3.1. Specificity and Generalisation

Our approach to model evaluation is based on directly measuring key properties of a given model. An effective model is one which is able to generate a broad range of example of the class of modelled images. This property is referred to as

Generalisation ability. This property is not sufficient since the model must also generate examples that are *consistent* with the class of modelled images. This property is referred to as *Specificity.*
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 [7]. 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 σ_S and σ_G can be defined as follows:

$$G = \frac{1}{n} \sum_{i=1}^{n} min_j \, |I_i - I_j|, \tag{3}$$

$$S = \frac{1}{m} \sum_{j=1}^{m} min_i \, |I_i - I_j|. \tag{4}$$

$$\sigma_G = \frac{SD(min_j \, |I_i - I_j|)}{\sqrt{n-1}}, \tag{5}$$

$$\sigma_S = \frac{SD(min_j \, |I_i - I_j|)}{\sqrt{m-1}}. \tag{6}$$

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.

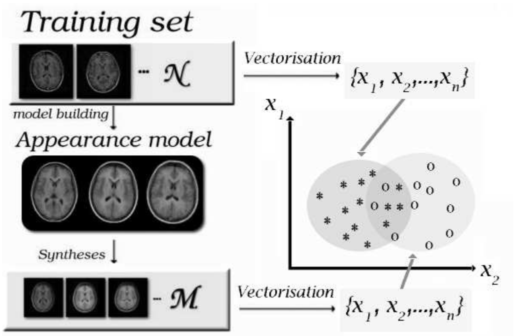

Fig. 2. The model evaluation framework. A model is constructed from the training and images are generated from the model. Each image is vectorised and can be visualised as a cloud in hyperscape.

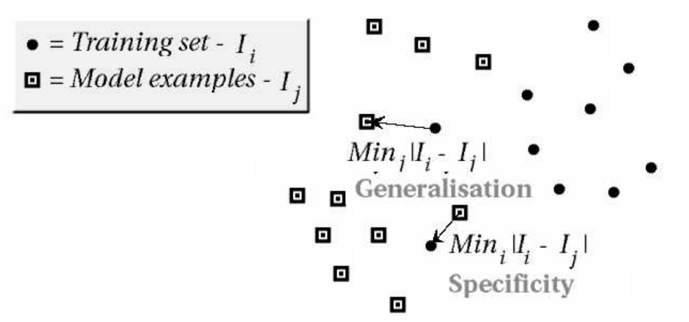

Fig. 3. Representation in hyperspace of the model metrics calculation method

3.2. 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' [12]. 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.

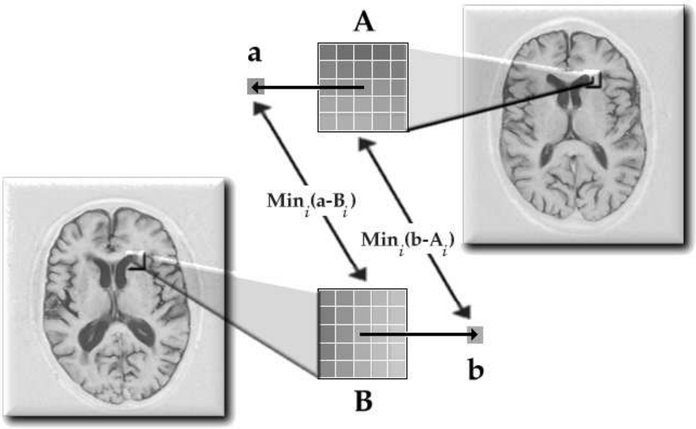

Fig. 4. The calculation of a shuffle difference image

4. Validation of the Approach

Each warps will on average shift a pixel a distance of 2-3(XX) pixels.

4.1. 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

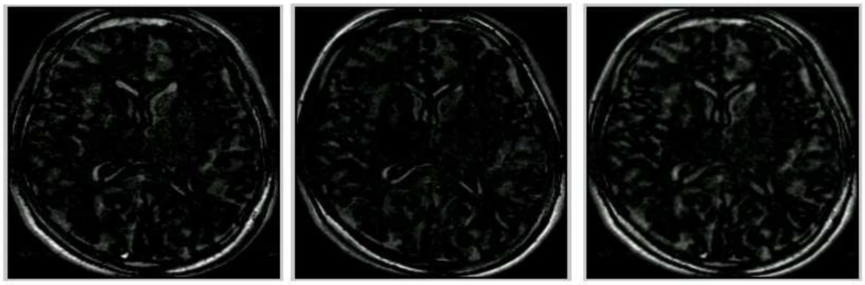

Fig. 5. An example of the shuffle difference image from one image to a second image (left), from the second image to the first (centre), and the symmetrical shuffle distance image (right)

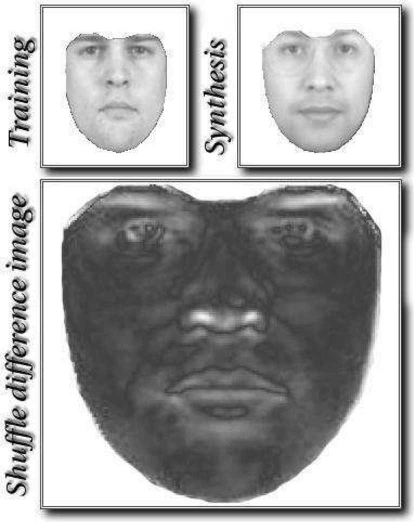

Fig. x. An example of a shuffle distance image, computed from a training set face image and an arbitrary model synthesis.

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 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 $m = 1000$.

4.2. 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

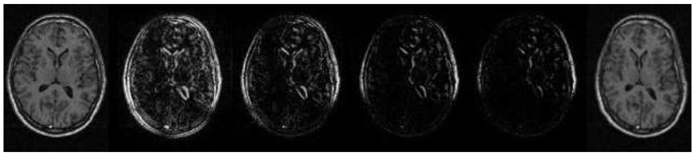

Fig. x. A comparison between shuffle distance evaluation types. On the left: original image; on the right: warped image; in the centre (from left): shuffle distance with $r = 0$(absolute difference), $1.5, 2.9$ and 3.7.

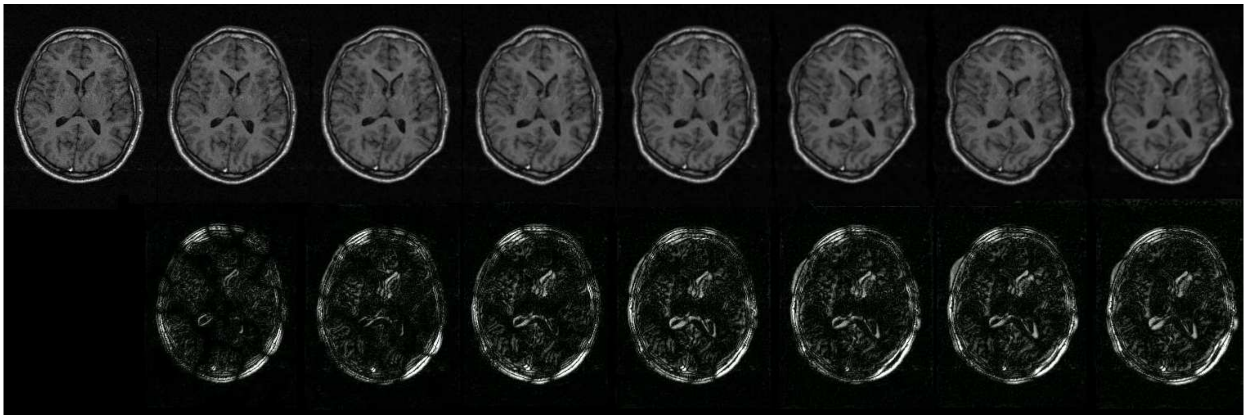

Fig. x. Examples of registration degradation for 0 to 7 concatenated amooth CPS warps. Euclidean distance to the original image is shown below.

that this would extend the range over which different degrees of mis-registration could be discriminated.

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

4.3. 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 [14]. However, it has been argued [5] that this *pairwise* approach does not take advantage of the full information in the group, and thus may lead to sub-optimal registration. We have been investigating *groupwise* methods of registration which aim to make the best use of the group as a whole when estimating the correspondence. We work within a minimum description length (MDL) framework. The aim is to construct a statistical appearance model which can exactly synthesize each example in the training set as efficiently as possible [20]. 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 [20]):

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

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

5. Results

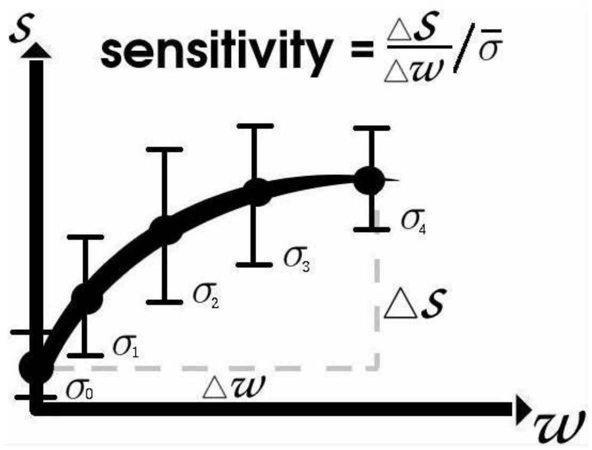

Fig. 7. The calculation of sensitivity for measures of Generalisation and Specificity.

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[2]. The results for different sizes of shuffle neighbourhood demonstrate that the range of mis-registration over which distinct values of specificity and generalisation are obtained increases as the neighbourhood size increases.

The results of the comparison between three different methods of NRR are shown in Fig. 9. These show that, particularly in terms of specificity, we can

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

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

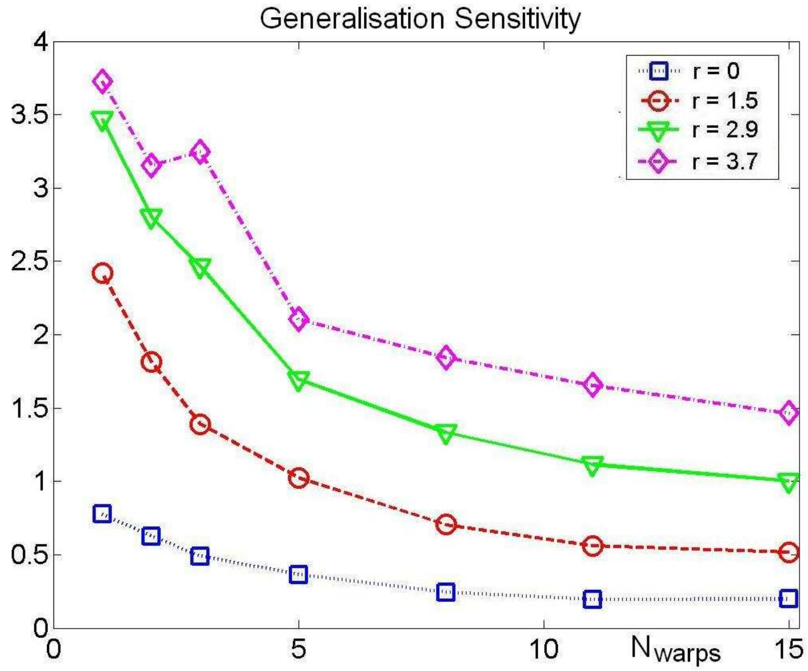

Fig. 8. Sensitivity of Generalisation for face data where registration degrades.

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.

6. Discussion and Conclusions

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

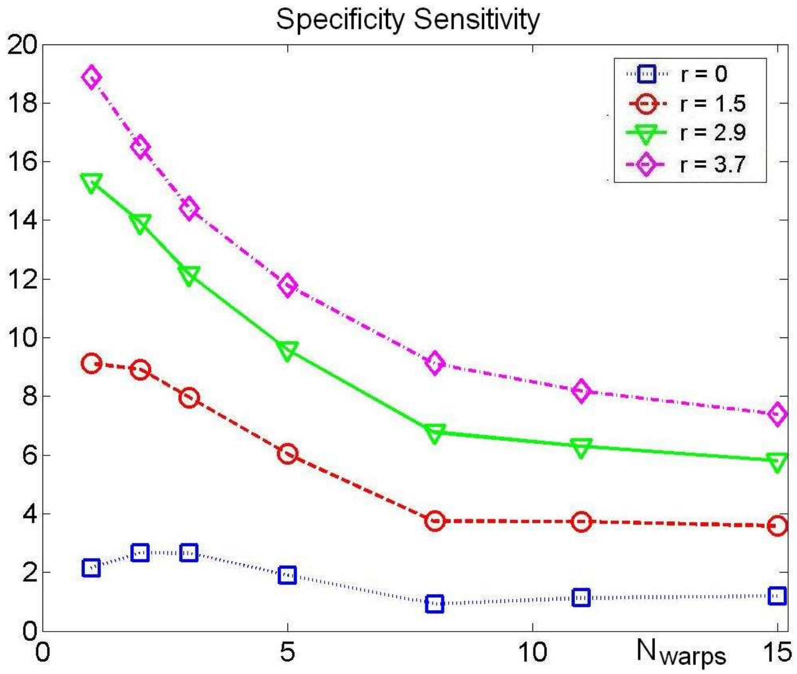

Fig. 9. Sensitivity of Specificity for face data where registration degrades.

References

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

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

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

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

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

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

[7] 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.

[8] 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.

[9] 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.

[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 Transactions on Medical Imaging*, 21:1151-66, 2002.

[11] P. Hellier, C. Barillot, I. Corouge, B. Giraud, G. Le Goualher, L. Collins, A.

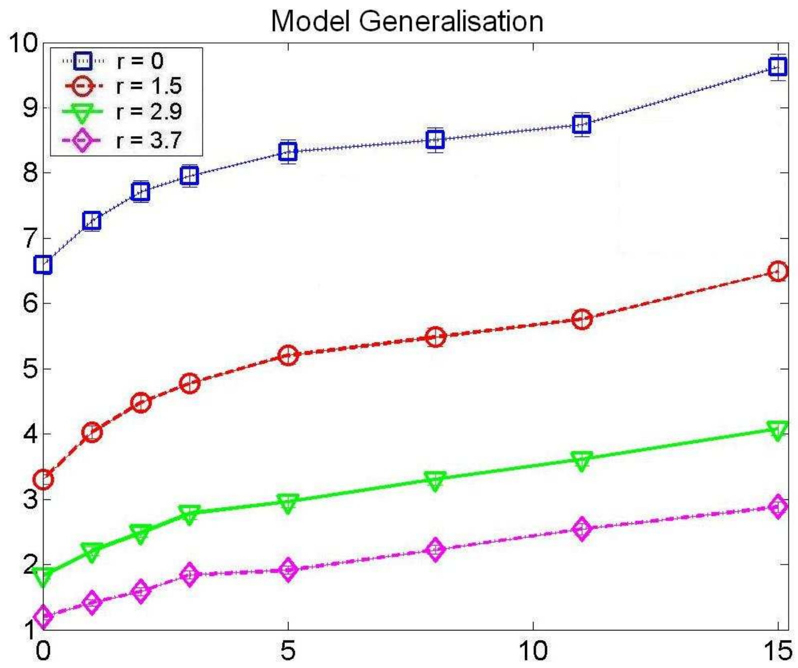

Fig. 10. Generalisation (with corresponding error bars) of brains as their registration degrades

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.

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

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

[14] 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.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] 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

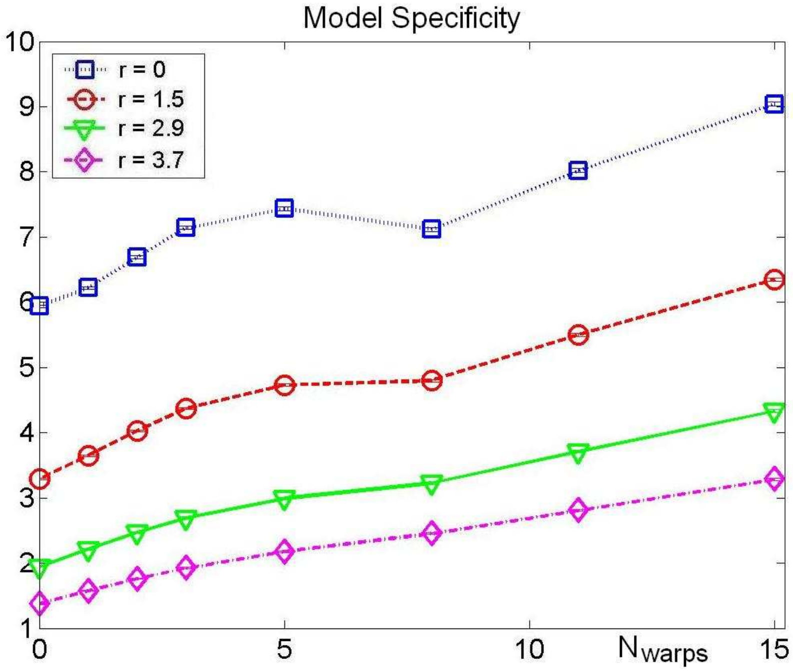

Fig. 11. Specificity (with corresponding error bars) of brains as their registration degrades

Fig. 12. The first mode of an appearance model of the brain whose training set was subjected to deformation. ±2.5 standard deviations are shown.

and model building. To be presented in *Information Processing in Medical Imaging*, 2005.

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

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

Fig. 15 Registration evaluation which compares 3 different registration algorithms. Specificity is shown on the left and generalisation ability on the right. Values are the mean over a wide range of modes in the model. SEE BELOW – MERGE?

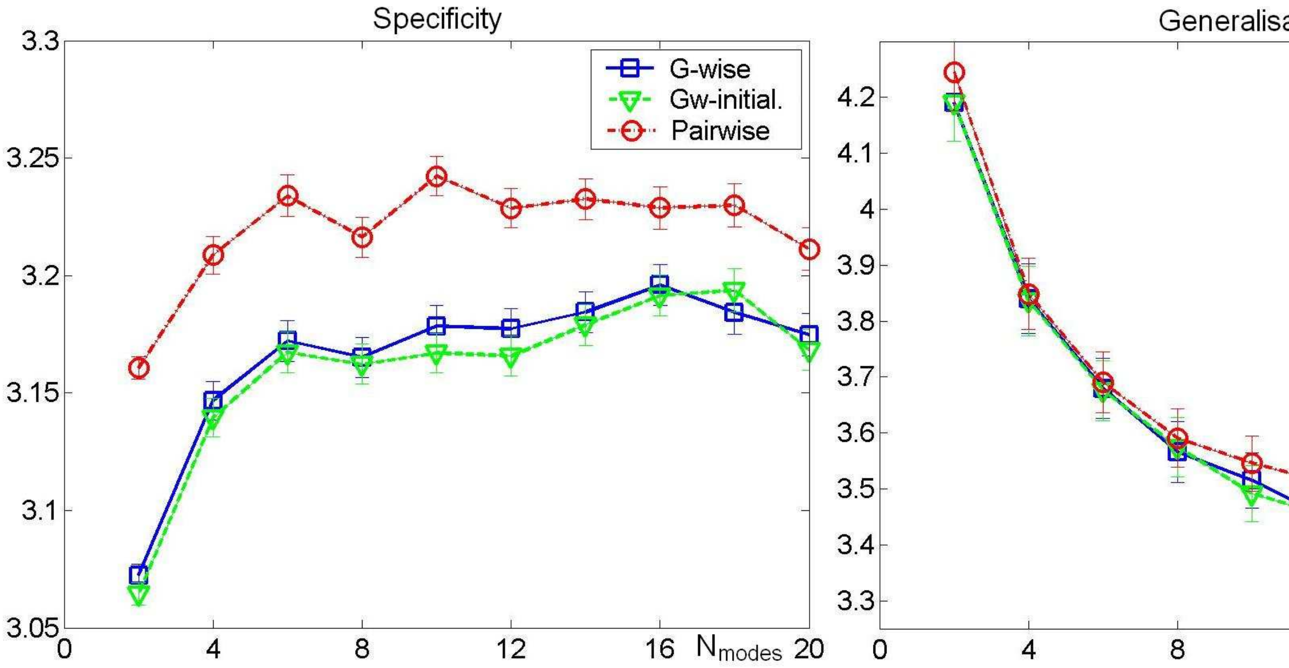

Fig. 16. Specificity and generalisation of the three registration methods.