Available online at www.sciencedirect.com
SCIENCE DIRECT®
Medical Image Analysis 8 (2004) 311–323
MEDICAL IMAGE ANALYSIS
www.elsevier.com/locate/media

Tuning and comparing spatial normalization methods

Steven Robbins [a,b,*], Alan C. Evans [a], D. Louis Collins [a], Sue Whitesides [b]

[a] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, 3801 University Street, Montreal, Que., Canada H3A 2B4
[b] School of Computer Science, McGill University, 3480 University Street, Montreal, Que., Canada H3A 2A7

Available online 7 August 2004

Abstract

Spatial normalization is a key process in cross-sectional studies of brain structure and function using MRI, fMRI, PET and other imaging techniques. A wide range of 2D surface and 3D image deformation algorithms have been developed, all of which involve design choices that are subject to debate. Moreover, most have numerical parameters whose value must be specified by the user. This paper proposes a principled method for evaluating design choices and choosing parameter values. This method can also be used to compare competing spatial normalization algorithms. We demonstrate the method through a performance analysis of a nonaffine registration algorithm for 3D images and a registration algorithm for 2D cortical surfaces.

Keywords: Anatomical variability measure; Registration performance measure; Image registration; Surface registration; Brain mapping

1. Introduction

The goal of spatial normalization in brain imaging is to remove, to the extent possible, the natural anatomical variability in a population by warping each individual's anatomy into a standardized space. Meaningful comparisons of spatially varying data (structural or functional) can then be made. The sensitivity of such comparisons is reduced by anatomical variability remaining after standardization. We wish to quantify this residual variability in order to choose the spatial normalization method for which it is the lowest.

The standardized system in widespread use today is a 3D Cartesian coordinate system into which each individual is mapped by an affine spatial transformation. Such a mapping procedure corrects only for location, orientation, and overall size of the input brain, leaving much variability (Steinmetz et al., 1989).

A nonaffine transformation enables removal of anatomical variability to a greater extent. Many algorithms for nonaffine mapping have been proposed (e.g., Ashburner and Friston, 1999; Bajcsy and Kovačič, 1989; Bookstein, 1989; Christensen et al., 1996; Collins and Evans, 1997; Davatzikos, 1996; Thirion, 1998; Thompson and Toga, 1996; Woods, 1998), these differ in the set of transformations searched, transformation parameterization, how the search is conducted, and the image feature used to drive the search. Such algorithms search for a spatial mapping T from input image I to image J by explicitly or implicitly minimizing some objective function of the form

$$\Phi(T) = \Phi_{\mathrm{D}}(I, J \circ T) + a\Phi_{\mathrm{M}}(T), \tag{1}$$

where Φ_{D} represents the data (image similarity) term and Φ_{M} represents the model term, also known as the regularizer as it embodies our "prior knowledge" of the transformation expected. The mathematical form for a data term has a theoretical basis in some instances (Roche et al., 2000). However, there is no biological theory to suggest a model term appropriate for transformation of one individual to another, so the models in use are either ad-hoc (Collins and Evans, 1997) or borrowed from physics (e.g., elastic solids (Bajcsy and Kovačič, 1989), viscous fluids (Christensen et al., 1996), or diffusion (Thirion, 1998)). These models include parameters corresponding to physical quantities such as "stiffness" or viscosity whose value is not determined by

* Corresponding author. Present address: Intelerad Medical Systems Inc., 460 Ste-Catherine St. W. Suite 210, Montreal, Que., Canada H3B 1A7. Tel.: +1-514-931-6222x7707; fax: +1-514-931-4653.
E-mail addresses: stever@bic.mni.mcgill.ca (S. Robbins), alan@bic.mni.mcgill.ca (A.C. Evans), louis@bic.mni.mcgill.ca (D.L. Collins), sue@cs.mcgill.ca (S. Whitesides).

1361-8415/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2004.06.009

theory. The coefficient a in Eq. (1), balancing the contribution of the data and model terms, is also undetermined by theory.

While spatial normalization is typically carried out using 3D transformations to match volumetric images, recently there has been a lot of interest in normalizing only the cerebral cortex treated as a 2D manifold (Van Essen et al., 1998; Fischl et al., 1999; Vaillant and Davatzikos, 1999; Thompson and Toga, 1996). As in the 3D image context, many parameter and design choices are not specified by theoretical reasoning.

An empirical performance measure is therefore required to evaluate design choices such as data and model terms, and to select parameter values. In the context of spatial normalization, residual anatomical variability is the natural choice for performance measure. In this paper we present such a measure of variability and demonstrate how it can be used to evaluate design choices and tune parameters of two different registration algorithms, dramatically improving the resulting registrations.

2. Methods

2.1. Anatomical variability measure

Anatomical variability is often visualized qualitatively in the "sharpness" of the mean intensity image after spatial normalization. The intensity values of a structural magnetic resonance (MR) image, while obviously carrying anatomical information, are affected by factors such as scanner settings, the partial volume effect, and the shading artifact. It is unclear how much the raw MR intensity value tells us about biological homology.

Instead, some anatomical "label" can be used which identifies a specific anatomical feature, as a dimensionless point landmark, a curve (1D), surface (2D) or volume (3D) label field. Anatomical variability can be quantified using some measure of the spatial distribution of corresponding points (Grachev, 1999), curves (Woods, 1998; Steinmetz et al., 1989), surfaces (Hellier et al., 2001), or volumes (Roland et al., 1997; Fischl et al., 1999). These measures use a limited number of features, e.g., 128 landmark points per hemisphere (Grachev, 1999), leaving them insensitive to the value of T at unlabelled points. We prefer a variability measure that is sensitive to each voxel of the standardized space.

A *segmentation* of an image is an assignment of a class label to each voxel. The labels can represent any relevant information. In this paper, labels of tissue type (gray matter, white matter, CSF, or background) are used and also labels of sulcal branches. While such labels represent structural anatomy, labels representing functionally defined regions could equally well be used.

Labels assigned to an input image can be carried along with a spatial transformation to induce a segmentation of a grid in the standard space. Using a "ground truth" segmentation of the standard space, Crivello et al. (2002) measure the label agreement between the ground truth and the induced segmentation of each individual. Given a population of individuals, mean label agreement is used as the measure of residual anatomical variability after spatial normalization. We can avoid requiring ground truth, and the attendant concerns about biasing the results if the ground truth is incorrect, by instead looking for label consistency across the population of at each location of standard space.

In this work each input brain has an associated segmentation. After spatial normalization is performed, each individual also has a segmentation in standard space. Each standard space voxel is thus associated with a set of labels, one label per individual. If a set of images were well-aligned after transforming each to the standard space, then any given location in the standard space would be consistently matched to similar tissue in each of the subjects. In the ideal (no variability) case, the labels at a given location in standard space would all be the same. In practice, this does not happen and we seek to measure how far we are from the ideal; i.e., to measure the variability at each location in standard space.

To measure the variability of labels at each standard space voxel v, we begin with the probability, $p_l(v)$, that voxel v takes label l. This probability is estimated as the fraction of inputs whose corresponding voxel is labelled l. The set of probabilities, $\{p_l(v) : l \in \mathscr{L}\}$, where $\mathscr{L}$ is the set of possible labels (e.g., tissue labels) is the *probability distribution* at voxel v. A standard measure of variability of this probability distribution is the *information entropy* (Cover and Thomas, 1991), defined as

$$H(v) = -\sum_l p_l(v) \log_2 p_l(v), \tag{2}$$

where $0 \cdot \log_2 0$ is defined to be zero. This entropy is a measure of the uncertainty in the label that should be assigned to voxel v. For example, a spatial normalization method that achieves its goal of matching homologous points of each input at voxel v will result in an identical label (say k) across the subjects, i.e., $p_k(v) = 1$ and $p_l(v) = 0$ for $l \neq k$, and thus have zero entropy. The other extreme is a distribution where each label is equally likely, i.e., $p_l(v) = 1/m$, where m is the size of the label set $\mathscr{L}$; in this case, the entropy is $\log_2 m$. The entropy of any other distribution falls between these two extremes (Cover and Thomas, 1991). The experiments presented later use either four classes or two classes, so the entropy values fall in the range 0–2, or 0–1, respectively.

The entropy $H(v)$ measures the amount of uncertainty (in bits, as we use base-2 logarithms) of the label

at v. We follow Warfield et al. (2001) in regarding $H(v)$ as the anatomical variability at voxel v. The sum

$$H = \sum_v H(v), \tag{3}$$

which we term *total entropy*, is used as an overall measure of variability remaining after spatial normalization is applied. We wish to tune a registration algorithm so that the total entropy is minimized.

2.2. ANIMAL: non-rigid registration of 3D images

To illustrate the utility of tuning using total entropy, we use the ANIMAL algorithm (Collins and Evans, 1997) as a prototypical nonaffine registration method for 3D images. This section briefly describes the algorithm, with attention to the numerical parameters the user must choose. The resulting transformation T is applied after an initial affine transformation. For convenience, ANIMAL works with the displacements $\Delta(x) \equiv T(x) - x$ rather than the transformation T itself. The displacement function Δ estimated by ANIMAL is parameterized as a *freeform deformation*, that is, the displacement vectors are stored for vertices arranged on a cubic 3D control mesh. At non-vertex points, the displacement is obtained using a cubic Catmull-Rom interpolating spline.

The transformation function is specified in a "world" coordinate system, which is defined independently of the source and target image voxel grids. Each image is endowed with the affine transformation function between world coordinates and its own voxel grid. Lengths in the world coordinate system, such as the control mesh vertex spacing, are given in units of millimeters.

ANIMAL is structured as two nested loops. The outer loop iterates over different control meshes in a coarse-to-fine manner, while the inner loop optimizes Δ on a fixed control mesh.

2.2.1. Outer loop

The first iteration of the outer loop employs a control mesh with a vertex spacing of 8 mm and is referred to as the 8 mm grid. The feature used in the match is a smoothed version of the input image, computed by convolution with an isotropic filter. The filter kernel is a Gaussian function whose full width at half maximum value (FWHM) is 8 mm. The next two iterations use a control mesh with a vertex spacing of 4 mm (4 mm grid) and 2 mm (2 mm grid). The source and target images are again smoothed with an isotropic Gaussian kernel: FWHM = 8 mm for the 4 mm grid and FWHM = 4 mm for the 2 mm grid. Finally, a fourth iteration with a vertex spacing of 2 mm is done using smoothed (FWHM = 4 mm) gradient magnitude images.

The initial iterate for the inner loop is interpolated from the result of the previous iteration of the outer loop, except the first iteration which starts with zero displacements.

2.2.2. Inner loop

Using v to index the control mesh vertices, let Δ_v be the current estimated displacement at vertex v and δ_v be the correction to Δ_v estimated at each iteration of the inner loop. We use $||\delta_v||$ to denote the magnitude of vector δ_v. The inner loop of ANIMAL is displayed in Algorithm 1.

Algorithm 1. Inner loop of ANIMAL.

(1) Optimize $\Phi(\{\delta_v\}) = \sum_v (a_1 \phi_v(\Delta_v + \delta_v) + (1 - a_1)\psi(||\delta_v||))$.
(2) Let $\Delta_v = \Delta_v + a_2 \delta_v$.
(3) Let $\overline{\Delta_v}$ be mean displacement of 26-neighbours of v. Set $\Delta_v = a_3 \overline{\Delta_v} + (1 - a_3)\Delta_v$.
(4) Loop over Steps 1–3 a fixed number of times.

The objective function of Line 1 is composed of two terms for each control mesh vertex. The first term, ϕ_v, is an image similarity measure (typically normalized cross correlation) evaluated on a small neighbourhood (a sphere of radius 1.5 times the control mesh vertex separation) around vertex v. The second term, ψ, is an increasing function that approaches ∞ at a finite value of $||\delta_v||$, thus limiting the size of the correction vector. The parameter $a_1 \in [0, 1]$ balances these two terms, and is known as the *similarity cost ratio*, or simply "similarity".

Each term of Φ is a function of exactly one correction vector δ_v so the optimization can be performed independently for each v, resulting in a large number of small optimization problems: each δ_v has three variables to optimize, namely the displacement in the x-, y- and z-directions. However, the optimization at control vertex v is *not* performed (and δ_v is set to zero) if the source image value at that location falls below 10% of the maximum source image value. Such locations are likely to be background and are skipped since there is nothing to be gained by fitting background regions that are dominated by noise. This heuristic is termed *node thinning*.

The update step of Line 2 employs a *weight* parameter a_2. The displacements are under-corrected if $a_2 < 1$ or over-corrected if $a_2 > 1$.

The displacement vector Δ_v is smoothed in Line 3 by taking a weighted sum of the current displacement estimate with the mean displacement of the 26 neighbours in a $3 \times 3 \times 3$ control mesh neighbourhood centered on v. The *stiffness* parameter $a_3 \in [0, 1]$ balances the two terms.

2.2.2.1. Parameters. The three parameters a_1, a_2 and a_3 need to be specified in order to complete the description of ANIMAL. Collins and Evans empirically chose values of 0.5, 0.6 and 0.5, respectively (Collins and Evans, 1997). These values were obtained by trial-and-error

using visual inspection of the displacements and resampled images to judge registration quality.

2.3. Non-rigid registration of 2D cortical surfaces

In order to demonstrate the general utility of using entropy for tuning, the method is applied to a second registration algorithm. We choose a non-rigid registration method recently developed (Robbins, 2003) for matching 2D cortical surfaces. In a pre-processing step, each surface is mapped to the unit sphere. The registration then searches for a mapping T that transforms the unit sphere to itself. The transformation T is specified using a triangulation of the source sphere. The value of the transformation is stored for each vertex of the triangulation and linearly interpolated at any non-vertex point.

The registration is structured as two nested loops, in a manner similar to ANIMAL. The outer loop iterates over different control meshes in a coarse-to-fine manner, while the inner loop optimizes T on a fixed control mesh.

2.3.1. Outer loop

The control mesh is obtained by repeated quadrisection [1] of an icosahedron. The first iteration of the outer loop employs the mesh obtained by threefold quadrisection of the regular icosahedron and thus has $4^3 \cdot 20 = 1280$ faces. There are three more iterations of the outer loop with control meshes containing $4^4 \cdot 20$, $4^5 \cdot 20$ and $4^6 \cdot 20$ faces, respectively.

The initial iterate for the inner loop is interpolated from the result of the previous iteration of the outer loop, except the first iteration which starts with the identity transformation.

2.3.2. Inner loop

The inner loop for surface registration first performs an optimization, starting at the current estimate T, for a transformation U that optimally matches the two surfaces. Then U is smoothed to produce the subsequent estimate of transformation T. In Algorithm 2, v is used to index the control mesh vertices.

Algorithm 2. Inner loop of Surface Registration.

(1) Minimize $\Phi(U) = \sum_v (\phi_v(U(v)) + a\psi(||U(v) - T(v)||))$.
(2) Let $C(v)$ be centroid of $\{U(u) : u \text{ is neighbour of } v\}$. Set $T(v) = U(v) + wC(v)$, projected to unit sphere.
(3) Loop over Steps 1–2 a fixed number of times.

The objective function in Line 1 is composed of two terms for each control mesh vertex. The feature used in the match is the geodesic distance transform from gyral crown vertices, denoted the *crown distance transform*. The first term, ϕ_v, measures the similarity of this feature using correlation coefficient evaluated on a small neighbourhood about vertex v. The size of this neighbourhood is controlled by a parameter denoted the *neighbourhood search radius*, r_n. The second term, ψ, is an increasing function of the change in the transformation at vertex v, $||U(v) - T(v)||$. This term is designed to limit the search for $U(v)$ to the hemisphere centred at $T(v)$ in order to use a Euclidean 2D parameterization of the search space. Therefore ψ is designed to approach value ∞ when $U(v)$ moves a finite distance from $T(v)$. A parameter known as the *search radius* (r_s) specifies the size of the search region, which must be smaller than a hemisphere. The *penalty ratio* parameter, a, balances these two terms.

The smoothing operation in Line 2, which is carried out in $\mathbb{R}^3$, is a simple weighted average of $U(v)$ and the centroid of its neighbourhood, where w is a user-specified smoothing weight.

Parameters. The user of this algorithm has four major parameters to specify: the search radius r_s, the neighbourhood radius r_n, the penalty ratio a, and the smoothing weight w. The search radius and neighbourhood radius are dimensionless quantities that multiply a length set by the coarseness of the control mesh. These radii can therefore be set to a fixed value for all iterations of the outer loop, as can the parameters a and w.

3. Results

Our approach is to spatially normalize a set of brains (either 3D images or 2D surfaces) and compute the total entropy of the result. We perform this measurement with varying design choices of the registration algorithm and choose the design that produces the lowest total entropy. It might seem more straightforward to simply register the brains using entropy itself, but this would be computationally costly as discussed in Section 4.

3.1. Animal

To investigate design choices of ANIMAL, 40 T_1-weighted images are selected arbitrarily from the ICBM data base (Mazziotta et al., 1995). An arbitrary image is selected to be the template and the other 39 images are segmented into white matter, gray matter, cerebral spinal fluid, and background classes (Kollokian, 1996) with non-brain voxels removed (Smith, 2002). Using the first 10 images of the 39, the total entropy is computed after registration using various choices for weight, stiffness, and similarity. The total entropy allows us to compare the impact of various design choices. Consider first the outer loop.

[1] Quadrisection is the operation that replaces each triangular face by four triangles obtained by joining the midpoints of each edge of the initial triangle.

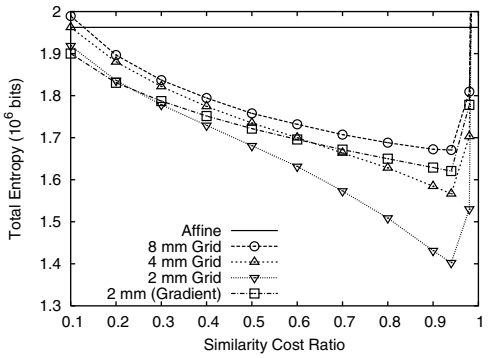

Fig. 1. Residual anatomical variability as measured by total entropy, H, on a sample of 10 individuals after registration with ANIMAL. Plot shows results after each of the four iterations of the outer loop (weight = 1.0, stiffness = 0.9) along with the value for 9-parameter affine normalization, for reference. Note that the variability is reduced for each of the first three iterations of the outer loop, but *increases* on the fourth iteration.

3.1.1. Outer loop

The expectation is that each iteration of the outer loop matches anatomy better than the previous iteration, and so the total entropy should decrease after each iteration. However, the results after the fourth iteration (matching using image gradient data) show an increase in total entropy. Fig. 1 shows representative results for weight = 1, stiffness = 0.9 and using a range of similarity values between 0.1 and 1. It is clear that the increase in entropy in the fourth level of the hierarchy occurs for a wide range of parameter values.

Close examination of the algorithm reveals that the node thinning strategy is the culprit. For the three iterations of the outer loop that use intensity data, this heuristic retains nearly all the control mesh vertices lying in brain tissue, while skipping control mesh vertices located outside of the head. In the gradient data iteration, however, only values on the scalp, ventricle, and superficial cortex edges are above the threshold. Displacements are therefore estimated on very few control mesh vertices (about 1/3 of the number of vertices in the previous outer iteration, which uses the same control mesh), while all vertices participate in the smoothing of the displacement vectors, Step 3 of Algorithm 1. The effect is to smooth out the warp, degrading the data fit.

Omitting the node thinning heuristic for the 2 mm grid gradient data fit brings the total entropy down below the value obtained using the 2 mm grid intensity fit. Omitting the heuristic for the intensity fits does not change the results appreciably, so no node thinning is done for any of the following.

3.1.2. Data term

In this section, we investigate changing the data similarity term from normalized cross-correlation to correlation coefficient.

The result of a number of tests using cross correlation with different parameter values is that the lowest entropy score is obtained using similarity = 0.98, weight = 0.8, stiffness = 0.98. Similar testing using correlation coefficient yields similarity = 0.3, weight = 1.0, stiffness = 1.0 as optimum. The entropy score in each case is approximately equal to 1.37×10^6 bits. However, the sensitivity to parameter variation about the optimal set is markedly different. The two plots in Fig. 2 compare the entropy scores as a function of similarity and of stiffness. Both plots show a much shallower curve with good performance (i.e., low entropy) over a broad range of parameter values, when using correlation coefficient.

The performance of ANIMAL using correlation coefficient is less sensitive to the parameter values than when cross correlation is used. This is desirable behaviour, so correlation coefficient is used for subsequent experiments.

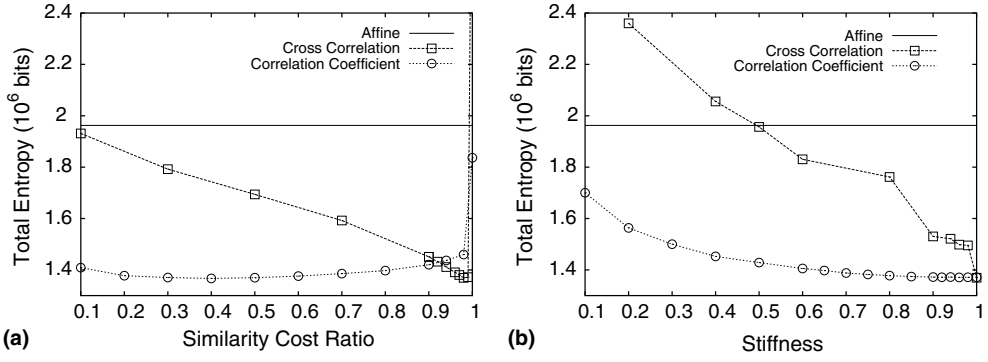

Fig. 2. A comparison of performance with two data similarity terms. Left: entropy as a function of similarity cost ratio for the cross correlation data term (using weight = 0.8, stiffness = 0.98) and the correlation coefficient data term (using weight = 1, stiffness = 1), along with the value for 9-parameter affine normalization for reference. Right: entropy as a function of stiffness for the cross correlation data term (using similarity = 0.98, weight = 1) and the correlation coefficient data term (using similarity = 0.3, weight = 1). For both plots, the data is taken at the 2 mm (intensity) level of ANIMAL outer loop. Note that the variability is much less sensitive to the parameter value when using the correlation coefficient data term.

3.1.3. Numerical parameters

Consider now the effect of the similarity cost ratio parameter (a_1) that controls the relative contributions of the data term, ϕ_v, and the displacement update penalty term, ψ, to the objective function. Fig. 2 shows that when using correlation coefficient, a value in the range 0.2-0.6 provides good performance.

Fig. 2 also shows that the entropy becomes large for similarity values near 1, even larger than obtained using the initial affine transformation. This phenomenon is observed using either of the two data terms. When the similarity parameter is set to 1, the registration is driven only by the data term ϕ_v with no control on the size of the correction vector, $||\delta_v||$. The transformations obtained contain much larger displacements, are much less smooth, and have more instances of folding (non-invertibility) than those obtained with similarity cost ratio <1. This confirms the importance of incorporating the regularization into the algorithm.

The right plot of Fig. 2 indicates that very high values of stiffness parameter (a_3) are best, so stiffness = 1 is used. The final numerical parameter of ANIMAL is the weight value (a_2), used in Line 2 of Algorithm 1. Values in the range 0.8–1.4 (using similarity = 0.3, stiffness = 1) show little change in entropy values. The weight is therefore generally set to 1.

3.1.4. Other experiments

The experiments presented all use the same set of 10 test subjects, which raises the question as to whether the parameters obtained are specific to the set of subjects used, or are generally applicable. To answer this, a second set of 10 subjects is registered using various similarity values with weight = 1 and stiffness = 1. The results (Fig. 3, left) show the same shallow curves, indicating that the same similarity = 0.3 value can be used for the new set.

While total entropy of the tissue classification gives a good measure of overall matching, it is also of interest to know how well the spatial normalization succeeds in aligning specific structures, such as a particular sulcus. To quantify this, a human anatomical expert manually identified 46 sulcal segments in the frontal lobes of the 40 ICBM images under study.

The total entropy of this set of labels (Fig. 3, right) shows the same slowly changing behaviour as a function of similarity cost ratio as obtained using the tissue labels (Fig. 2). However, the fourth outer loop iteration, which uses a gradient fit, produces a larger total entropy than that of the third iteration (the 2 mm grid intensity fit).

3.1.5. Tuned results

Fig. 4 provides a visual illustration of the reduced anatomical variability in the full set of 39 individuals, obtained using the tuned version of ANIMAL (only the first three steps of the outer loop are used). The variability in the depth of many sulci is reduced, indicating that the sulci are better aligned.

The improved alignment is most readily apparent in the large regions of homogeneous tissue such as the white matter and ventricles. The coronal view (second row in Fig. 4) clearly shows that the tuning is instrumental in aligning the white matter of many gyri. The gray matter is also often well-matched and shows up with low variability, such as in the circled regions of Fig. 4.

Boundaries between tissue types show high variability, some of which is due to misalignment and some of which is due to limitations of the classifier used to generate the segmentation. At a boundary, even a misalignment on the order of the voxel size is enough to change the labelling from an input image and hence the entropy value. On the other hand, some of the apparent variability is the result of imperfections in the tissue

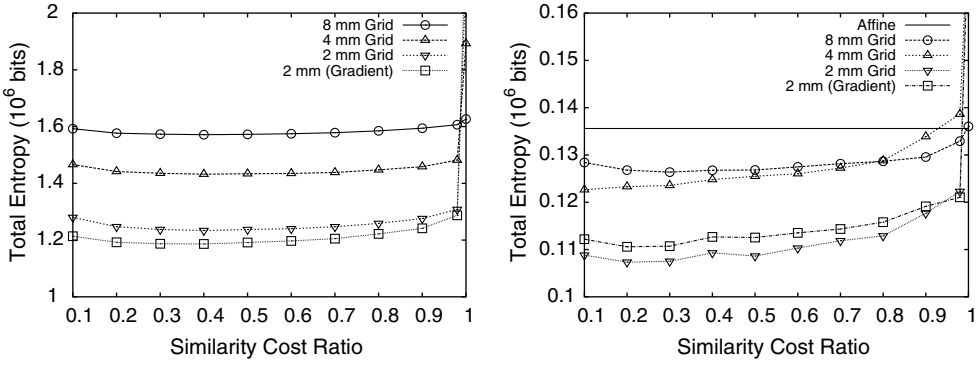

Fig. 3. Left: a second set of 10 subjects shows good performance in the same broad range of similarity values (using weight = 1, stiffness = 1) as exhibited in Fig. 2. The parameter values are thus not specific to the particular data set. Right: residual variability of segmentation of frontal sulci (46 labels). Plot shows results after each of the four iterations of the outer loop (weight = 1, stiffness = 1) along with the value for 9-parameter affine normalization, for reference. Note that similarity values in the range 0.2–0.6 produce good results, the same range as found using tissue class labelling.

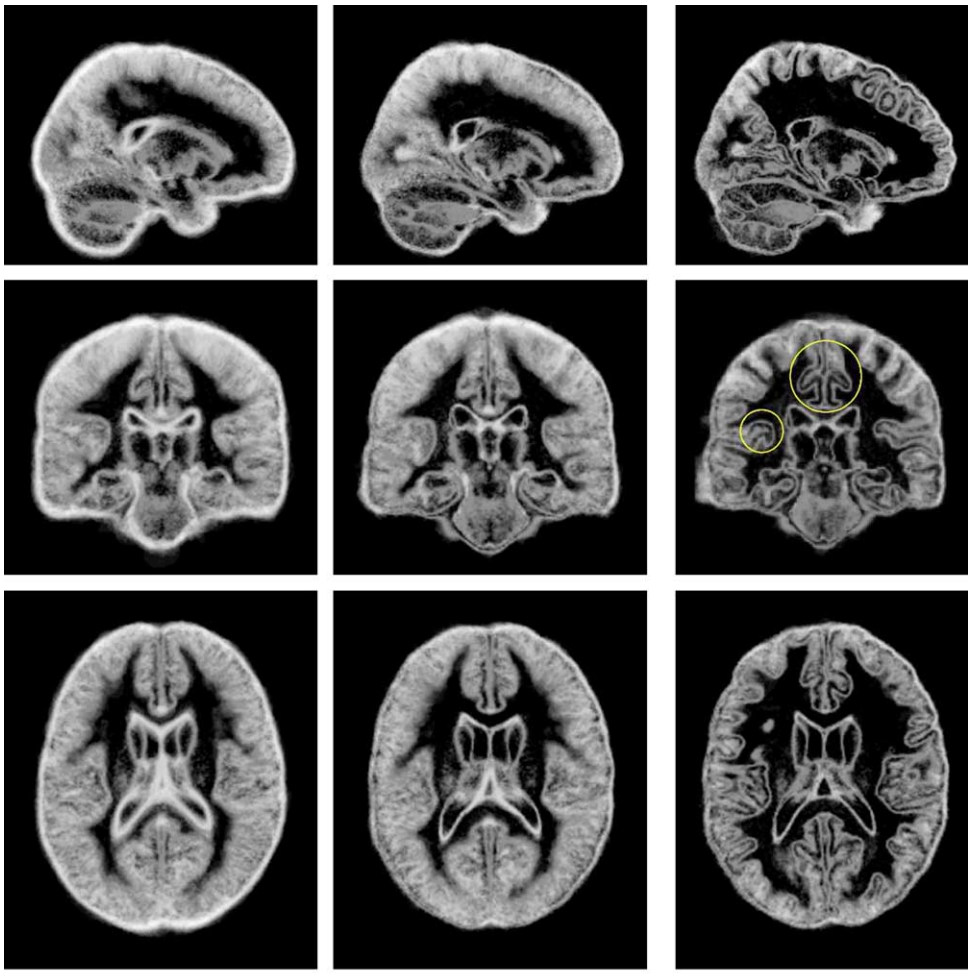

Fig. 4. Entropy maps of 39 individuals after spatial normalization using 9-parameter affine registration, ANIMAL with the default parameters (and cross correlation data term), and ANIMAL using correlation coefficient data term and optimal parameter values (similarity = 0.3, weight = 1, stiffness = 1). Voxels with more variability are brighter. Edges remain the most variable, both the white/gray interface and gray/CSF interface, producing an "outline" effect where the gray matter shows as less variable, bounded between two interfaces of high variability (examples are circled in the middle image of the third column). The third column shows a clear reduction in variability compared with the other two columns.

classification. Voxels in boundary regions frequent contain two (or more) tissue types, resulting in a signal intensity between the intensities of the two tissue types. Such partial volume voxels are more frequently misclassified. For example, the CSF is frequently misclassified as gray matter, leading to high entropy values in the CSF spaces of sulci.

Fig. 5 shows intensity-averaged images which become sharper with tuning, a qualitative display of the improvement in aligning fine detail.

3.2. Surface registration

The design choices of the surface registration algorithm can also be investigated using total entropy of a segmentation, as was done for ANIMAL. In this case, the label probabilities $p_l(v)$ are computed at every vertex v of a standard-space mesh. Eqs. (2) and (3) are used to compute the entropy, with the sums taken over all vertices rather than all voxels. The same template and ten test subjects that are used for the 3D work are used again to probe the choice of data term and to locate optimal parameter values.

In order to generate the segmentation of each test subject's surface, an automated vertex classification is required analogous to the tissue classification of voxels used in 3D. The cortical surface mesh is presumed to lie entirely on the boundary of two tissue types, specifically the interface between white matter and gray matter for the experiments presented here. Thus classification into tissue types is not an option. Instead, each vertex is classified as either gyral (lying on a gyral crown) or non-gyral. The classification is achieved by simply thresholding the crown distance transform values at distance

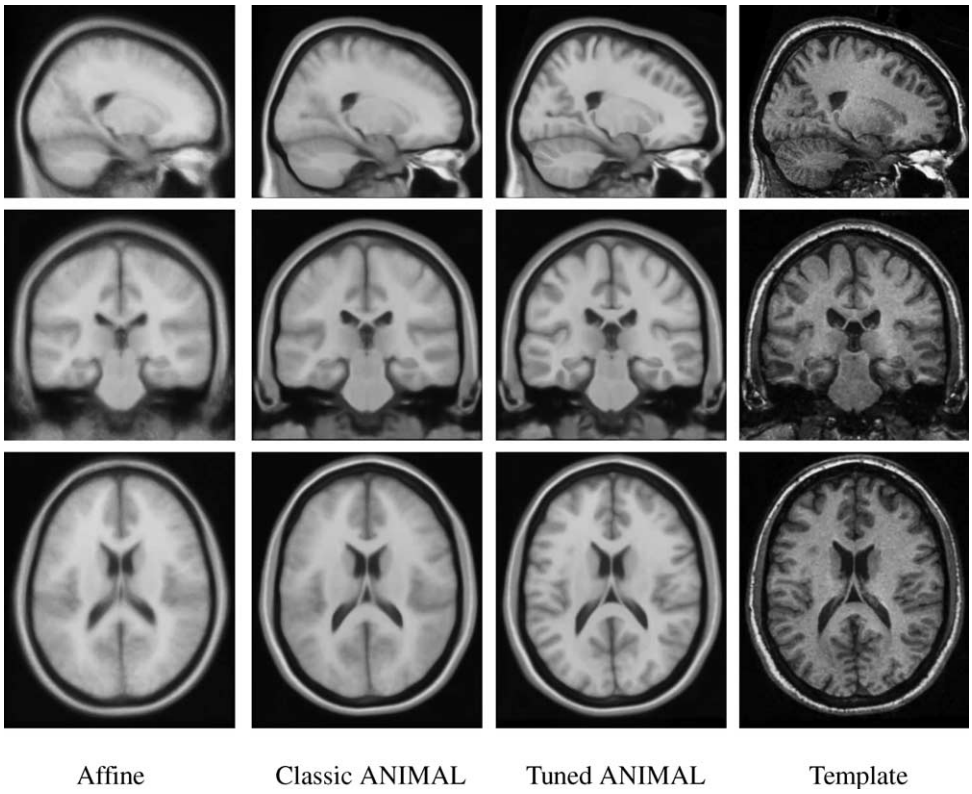

Fig. 5. Intensity-averaged images of 39 individuals after spatial normalization using 9-parameter affine registration, ANIMAL with the default parameters (and cross correlation data term), and ANIMAL using correlation coefficient data term and optimal parameter values (similarity = 0.3, weight = 1, stiffness = 1). The fourth column shows the template image. Note the sharpness of the third column compared to the first two columns, and the excellent matching of the third column with the template.

10 mm. In other words, all vertices lying with 10 mm (measured geodesically) of a gyral crown vertex are classified as ''gyral''.

3.2.1. Outer loop

The test data is normalized using several choices for the numerical parameters and the total entropy after each of the four iterations of the outer loop is computed. The left plot of Fig. 6 shows representative results using $r_s = 0.5$, $r_n = 2.8$, $w = 1$, and a range of penalty ratio values, a. The plot demonstrates that alignment improves at each finer resolution of control mesh.

3.2.2. Numerical parameters

The trade-off between the data match and the amount of displacement allowed during one step of the inner loop is controlled by the penalty ratio, a. Fig. 6 shows that the optimal value is $a = 0.05$. Setting the value of a

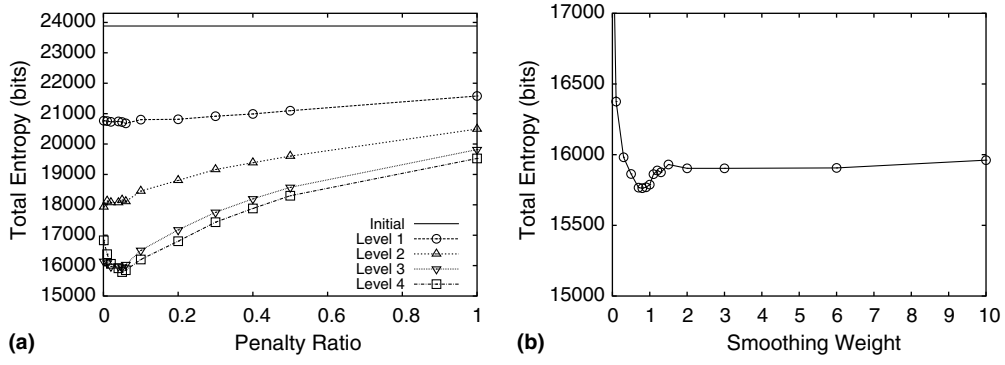

Fig. 6. Left: total entropy, H, on a sample of 10 individuals after surface registration, along with the initial (unregistered) value for comparison. The control mesh of each level is the quadrisection of the previous level; level 1 is the coarsest control mesh. Note the reduction of total entropy value with each mesh refinement level. Right: total entropy at level 4 as a function of smoothing weight.

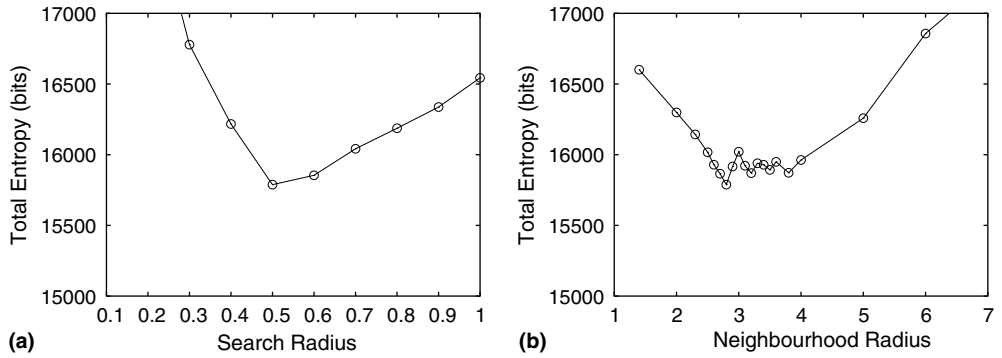

Fig. 7. Left: total entropy as a function of search radius, using neighbourhood radius = 2.8. Right: total entropy as a function of neighbourhood radius, using search radius = 0.5. Both plots use penalty ratio = 0.05 and smoothing weight = 1.

to zero eliminates the regularization and so, as is the case in 3D image matching, the warping is found to be less smooth.

The smoothing step of Algorithm 2 is controlled by the weight parameter, w. Fig. 6 (right) plots the entropy as a function of smoothing weight, showing weights near $w = 1$ provide the best performance. Note that large values for w will asymptotically set the smoothed mesh equal to the centroid value, $C(v)/||C(v)||$, and the performance does indeed level off in Fig. 6. For low values of smoothing weight, the plot shows a sharp increase in entropy. This happens because the lack of smoothing allows more folding of the mesh, which produces poor performance.

The other two major parameters for the algorithm are the search radius and the neighbourhood radius. Fig. 7 illustrates the optimal values $r_s = 0.5$ and $r_n = 2.8$, respectively. The behaviour obtained using too large or too small radius can be explained by looking at the details of the algorithm. For example, using a small search radius prevents the optimization from finding a good data match and results in a warping that is near the initial transformation and thus produces high total entropy. The poor performance produced by a large search radius, on the other hand, is because triangles are too easily able to reverse orientation. The range of radius values that give acceptable results, however, is not obtained from such considerations, so the performance measure is invaluable.

The experiments presented so far all use the same set of 10 test subjects. The variability as a function of penalty ratio is computed for a second set of 10 subjects. The results in Fig. 8 show the same qualitative behaviour and the same optimal value for the penalty ratio.

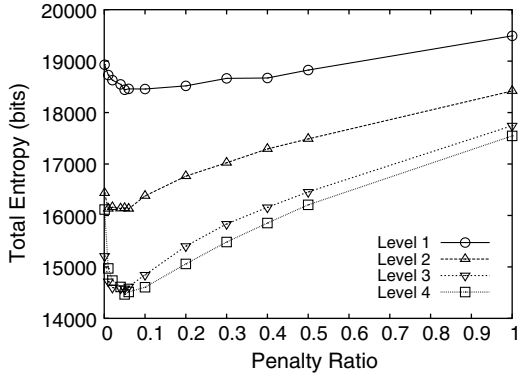

Fig. 8. A second set of 10 subjects show similar behaviour with respect to penalty ratio. Note that the optimal value of 0.05 is the same as in Fig. 6 (left).

3.2.3. Tuned results

Fig. 9 provides a visual illustration of the reduced anatomical variability using a set of 151 ICBM subjects registered using optimal parameter values $r_s = 0.5$, $r_n = 2.8$, $a = 0.05$ and $w = 1$. The variability is reduced in all areas of the cortex, indicating that the gyral patterns are better aligned. The variability that remains is concentrated on the edges of the gyral regions. As was the case in 3D, some of this variability is the result of imperfections in the vertex classification while some is due to misalignment. At a boundary, even a misalignment on the order of the spacing between control mesh vertices is enough to change the labelling from an input surface and hence the entropy value.

Fig. 10 shows images of the average crown distance transform feature value (analogous to the average intensity images of Fig. 5) which become sharper with tuning, a qualitative display of the improvement in aligning fine detail.

4. Discussion

Our use of entropy of the tissue labels raised a number of questions when this work appeared in an earlier form (Robbins et al., 2003). For example, why not work more directly with image intensity values rather than the

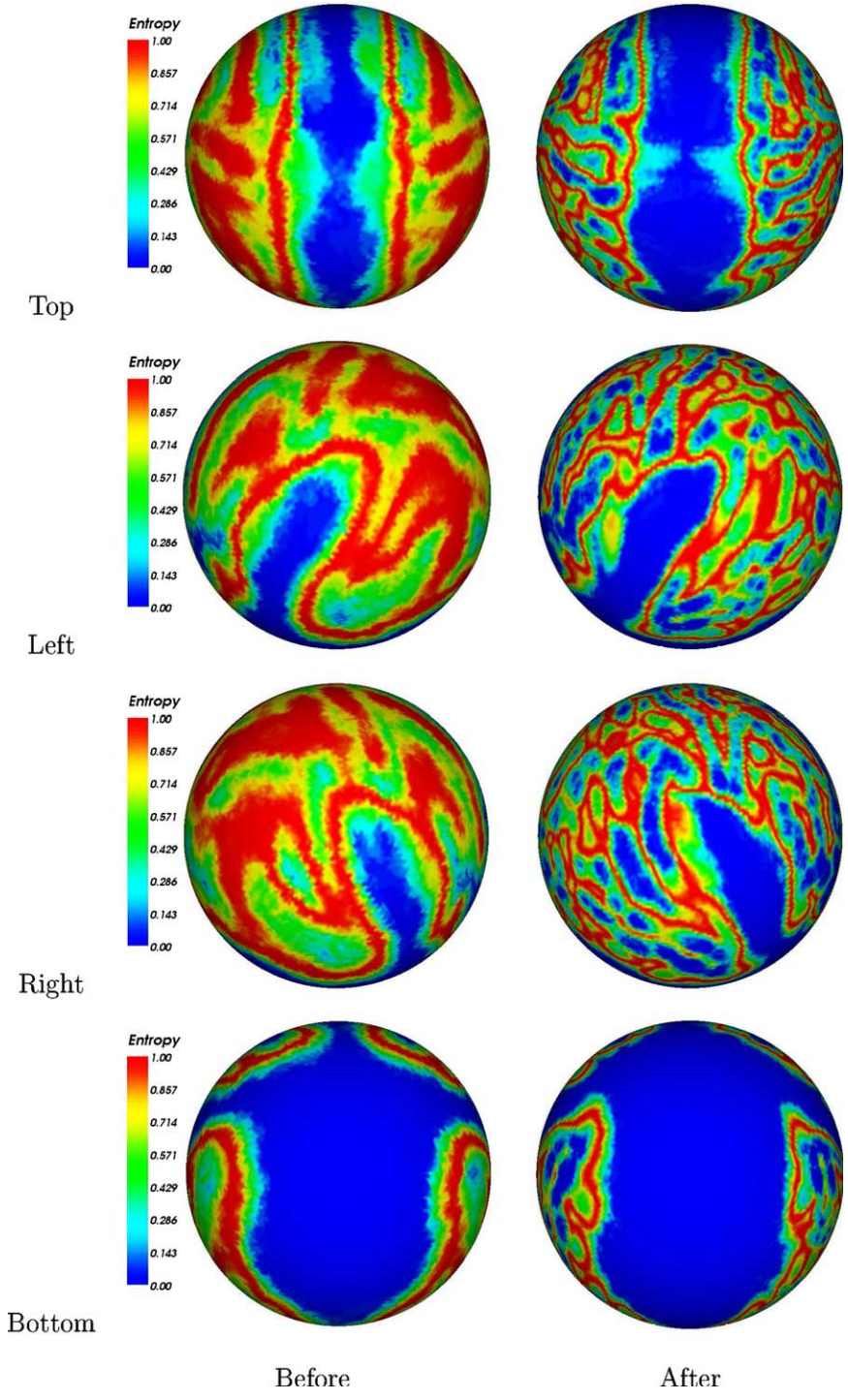

Fig. 9. Entropy maps before and after surface registration of 151 subjects. Notice that the entropy (variability) is reduced in all areas of the cortex after surface registration.

segmentation. We believe the segmentation is a more accurate indication of tissue type than is the image intensity. A typical MR image will exhibit a range of intensity values for a given tissue so while the matched pair voxels may both be white matter, the intensity value might well be different. This may give rise to a non-zero variability score even if the tissues are perfectly matched.

Another criticism is that a labelling with only white matter, gray matter, and CSF classes is crude and therefore so is the measure of variability. This is true to some extent, since in some cases gray matter belonging to a particular sulcus of one subject is being matched to gray matter contained in a neighbouring sulcus on a second subject. That does not negate the value of the

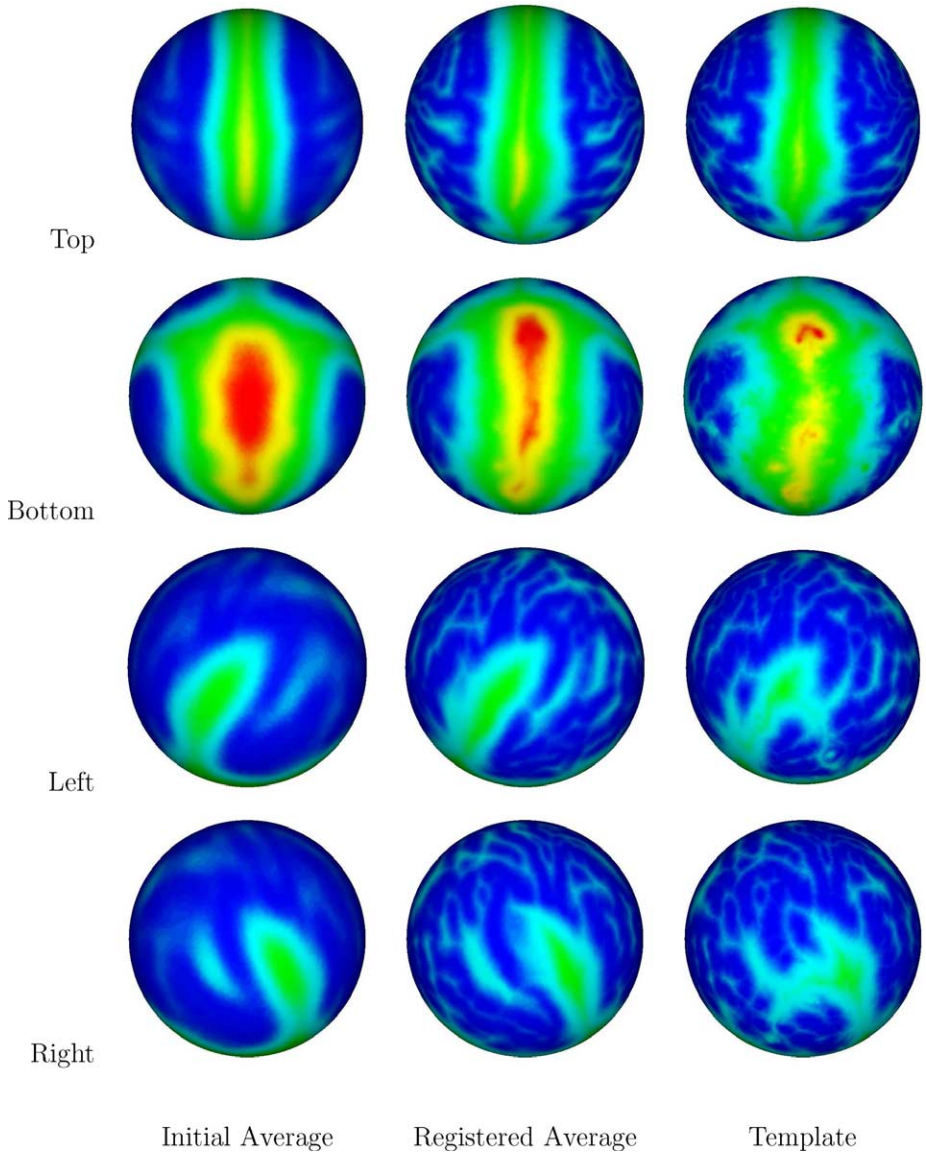

Fig. 10. Average of crown distance transform feature data of 151 subjects shown before and after surface registration with the template data shown for comparison. Note the appearance of smaller sulci after registration, and the agreement with the template data pattern.

general method, however, as the variability of *any* label data can be measured. The measure can be made more sensitive by using a finer labelling, e.g., functional fields or sulcal folds. The latter, in our experiments (see Fig. 3), shows broadly the same optimal parameters as those obtained using tissue labels. We choose to use simple tissue labels for two reasons. First, the labels can be obtained automatically using any one of a number of methods (Kollokian, 1996; Dale et al., 1999; Joshi et al., 1999). Second, the generated labels cover the brain, allowing a variability measure that is sensitive to label consistency across the entire brain.

Another point to consider in assessing competing algorithms for spatial normalization is whether to have one or several measures of variability. As we show, a single measure enables optimization of the algorithm parameters. Other authors (Warfield et al., 2001; Crivello et al., 2002) generate three or more measures of variability from tissue classification labels: the variability of CSF, of white matter, of gray matter, etc. This complicates the interpretation in the case that two normalization methods under comparison each score best in some measures but not for all measures; i.e., there may be no clear-cut winner. Though multiple measures would be useful in a situation in which performance tradeoffs were being evaluated, e.g., a tradeoff between residual variability and running time, it is not clear how one should trade off accuracy in normalizing different structures or tissue classes.

Finally, one may wonder whether it would be more straightforward to consider the entropy measure as a function of all N transformations and use total entropy

as the objective function for registration. Minimizing total entropy would thus co-register all N images simultaneously. Such an approach, which is feasible for affine registration of small 2D images (Miller et al., 2000), would be more direct than the method described here in which the registration optimization is done for each image (to the template) using fixed parameters selected using the entropy measure. However, given that the computational cost for (pairwise) 3D image registration using free-form deformations is already high, it is expected that attempting to simultaneously co-register a number of 3D images would not be feasible.

5. Conclusions

We have presented a strategy for evaluating the quality of a spatial normalization procedure on real data. The contribution of this paper is an objective method to assess the impact of any design choice such as the value of a numerical parameter or the choice of data similarity measure. As noted in Section 2.1, assessing variability using label consistency obviates the need for a ground truth segmentation. The evaluation procedure is fully automatic and can be applied to any spatial normalization method.

Our experiments on tuning pointed out several surprising features of the ANIMAL algorithm and allowed us to make modifications to it that improved its performance. We expect that our evaluation strategy would provide similar insights into other normalization methods, whether 3D or 2D. This entropy measure can also be used to compare two competing methods of normalization, once each has been suitably tuned.

References

Ashburner, J., Friston, K.J., 1999. Nonlinear spatial normalization using basis functions. Human Brain Mapping 7 (4), 254–266.

Bajcsy, R., Kovačič, S., 1989. Multiresolution elastic matching. Computer Vision Graphics and Image Processing 46, 1–21.

Bookstein, F.L., 1989. Principal warps: thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (6), 567–585.

Christensen, G.E., Rabbitt, R.D., Miller, M.I., 1996. Deformable templates using large deformation kinematics. IEEE Transactions on Image Processing 5 (10), 1435–1447.

Collins, D.L., Evans, A.C., 1997. ANIMAL: validation and applications of nonlinear registration-based segmentation. International Journal of Pattern Recognition and Artificial Intelligence 11 (8), 1271–1294.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. John Wiley & Sons, New York.

Crivello, F., Schormann, T., Tzourio-Mazoyer, N., Roland, P.E., Zilles, K., Mazoyer, B.M., 2002. Comparison of spatial normalization procedures and their impact on functional maps. Human Brain Mapping 16, 228–250.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis I: segmentation and surface reconstruction. NeuroImage 9 (9), 179–194.

Davatzikos, C., 1996. Spatial normalization of 3D brain images using deformable models. Journal of Computer Assisted Tomography 20 (4), 656–665.

Fischl, B., Sereno, M.I., Tootell, R.B.H., Dale, A.M., 1999. High-resolution inter-subject averaging and a coordinate system for the cortical surface. Human Brain Mapping 8 (4), 272–284.

Grachev, I.D., Berdichevsky, D., Rauch, S.L., Heckers, S., Kennedy, D.N., Caviness, V.S., Alpert, N.M., 1999. A method for assessing the accuracy of intersubject registration of the human brain using anatomic landmarks. NeuroImage 9, 250–268.

Hellier, P., Barillot, C., Corouge, I., Gibaud, B., LeGoualher, G., Collins, L., Evans, A., Malandain, G., Ayache, N., 2001. Retrospective evaluation of inter-subject brain registration. In: Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science, vol. 2208, pp. 258–265.

Joshi, M., Cui, J., Doolittle, K., Joshi, S., Van Essen, D., Wang, L., Miller, M.I., 1999. Brain segmentation and the generation of cortical surfaces. NeuroImage 9, 461–476.

Kollokian, V., 1996. Performance analysis of automatic techniques for tissue classification in magnetic resonance images of the human brain. Master's thesis, Computer Science, Concordia University, Montreal, Que., Canada.

Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 1995. A probabilistic atlas of the human brain: theory and rationale for its development. NeuroImage 2, 89–101.

Miller, E.G., Matsakis, N.E., Viola, P.A., 2000. Learning from one example through shared densities on transforms. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 464–471.

Robbins, S., Evans, A.C., Collins, D.L., Whitesides, S., 2003. Tuning and comparing spatial normalization methods. In: Ellis, R.E., Peters, T.M. (Eds.), Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science, vol. 2879. Springer, Montreal, Canada, pp. 910–917.

Robbins, S.M., 2003. Anatomical Standardization of the Human Brain in Euclidean 3-Space and on the Cortical 2-Manifold. Ph.D. Thesis, School of Computer Science, McGill University, Montreal, Que., Canada.

Roche, A., Malandain, G., Ayache, N., 2000. Unifying maximum likelihood approaches in medical image registration. International Journal of Imaging Systems and Technology 11, 71–80.

Roland, P.E., Geyer, S., Amunts, K., Schormann, T., Schleicher, A., Malikovic, A., Zilles, K., 1997. Cytoarchitectural maps of the human brain in standard anatomical space. Human Brain Mapping 5, 222–227.

Smith, S.M., 2002. Fast robust automated brain extraction. Human Brain Mapping 17 (3), 143–155.

Steinmetz, H., Fürst, G., Freund, H.-J., 1989. Cerebral cortical localization: application and validation of the proportional grid system in MR imaging. Journal of Computer Assisted Tomography 13 (1), 10–19.

Thirion, J.-P., 1998. Image matching as a diffusion process: an analogy with Maxwell's demons. Medical Image Analysis 2 (3), 243–260.

Thompson, P., Toga, A.W., 1996. A surface-based technique for warping three-dimensional images of the brain. IEEE Transactions of Medical Imaging 15 (4), 402–417.

Vaillant, M., Davatzikos, C., 1999. Hierarchical matching of cortical features for deformable brain image registration. In: Information Processing in Medical Imaging, Lecture Notes in Computer Science, vol. 1613, pp. 182–195.

Van Essen, D.C., Drury, H.A., Joshi, S., Miller, M.I., 1998. Functional and structural mapping of human cerebral cortex:

solutions are in the surfaces. Proceedings of the National Academy of Sciences USA 95 (February), 788–795.

Warfield, S.K., Rexilius, J., Huppi, P.S., Inder, T.E., Miller, E.G., Wells III, W.M., Zientara, G.P., Jolesz, F.A., Kikinis, R., 2001. A binary entropy measure to assess nonrigid registration algorithms. In: Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science, vol. 2208, pp. 266–274.

Woods, R.P., Grafton, S.T., Watson, J.D.G., Sicotte, N.L., Mazziotta, J.C., 1998. Automated image registration. II. Intersubject validation of linear and nonlinear models. Journal of Computer Assisted Tomgraphy 22 (1), 153–165.