Tuning and Comparing Spatial Normalization Methods

Steven Robbins[1,2], Alan C. Evans[1], D. Louis Collins[1], and Sue Whitesides[2]

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

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 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 particular nonaffine deformation algorithm, ANIMAL.

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 [13].

A nonaffine transformation enables removal of anatomical variability to a greater extent. Many algorithms for nonaffine mapping have been proposed (e.g. [1,2,3,4,5,7,14,15,17]); these differ in the set of transformations searched, transformation parameterization, how the search is conducted, and the image feature(s) 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

$$\Phi(T) = \Phi_D(I, J \circ T) + a\Phi_M(T), \tag{1}$$

R.E. Ellis and T.M. Peters (Eds.): MICCAI 2003, LNCS 2879, pp. 910–917, 2003.

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 [11]. 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 [5] or borrowed from physics (e.g. elastic solids [2], viscous fluids [4], or diffusion [14]). These models include parameters corresponding to physical quantities such as "stiffness" or viscosity whose value is not determined by theory. The coefficient a in Equation 1, balancing the contribution of the data and model terms, is also undetermined by theory.

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 a particular algorithm, 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 [9], curves [17, 13], surfaces [10], or volumes [12,8]. These measures use a limited number of features, e.g. 128 landmark points per hemisphere [9], leaving them insensitive to the value of T at unlabelled points. We prefer a variability measure sensitive to each voxel of the standardized space.

A *segmentation* of an image is an assignment of a class label to each voxel. 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. [6] measured the label agreement between the ground truth and induced segmentations and used mean label agreement as the measure of anatomical variability. We can avoid requiring ground truth, and the attendant concerns about biasing the results if the ground truth is incorrect, by instead looking for consistency across the induced segmentations of standard space.

Consider voxel v in the standard space grid. This voxel maps to a certain point in subject i, which has a label that we denote l_{vi}. A spatial normalization method that achieves its goal of matching homologous points of each input will result in identical labels across the subjects ($l_{v1} = l_{v2} = \ldots$) for each voxel v. Warfield et al. [16] suggested to consider the label of voxel v as a random variable L_v, of which the set $\{l_{vi}\}$ is a sampling. The entropy of this distribution is

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

where p_l is the probability that L_v is assigned label l. The entropy measures the amount of uncertainty in the label (in bits, as we use base-2 logarithms), which we regard as the anatomical variability at voxel v. The uncertainty over the entire standardized space is bounded by the sum

$$H = \sum_v H(L_v). \tag{3}$$

We use H, which we term *total entropy*, as a measure of variability remaining after spatial normalization is applied.

2.2 ANIMAL

To illustrate the utility of tuning using total entropy, we use the ANIMAL algorithm [5] as a prototypical nonaffine registration method. 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 $\Delta(x) = T(x) - x$ 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 points arranged on a cubic 3D control grid. At non-grid points, the displacement is obtained using a cubic Catmull-Rom interpolating spline.

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

Outer Loop. The first iteration of the outer loop employs a control grid with step=8 mm. The feature used in the match is the two input images, each blurred using an isotropic Gaussian kernel with FWHM=8 mm. The next two iterations use a control grid with step=4 mm (FWHM=8 mm blurring) and step=2 mm (FWHM=4 mm blurring), respectively. Finally, a fourth iteration with grid step=2 mm is done using blurred (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.

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.

Inner Loop. Use v to index the control grid vertices, Δ_v is the current estimated displacement vector for vertex v, and δ_v is 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.

The objective function of Line 1 is composed of two terms for each control grid vertex. The first term, ϕ_v, is an image similarity measure (normalized cross correlation) evaluated on a small neighbourhood (a sphere of radius 1.5 times the control-grid step length) 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*.

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, the displacement in the x-, y-, and z-directions). However, the optimization at control vertex v is *not* performed 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 grid neighbourhood centered on v. The *stiffness* parameter $a_3 \in [0, 1]$ balances the two terms.

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 [5]. These values were obtained by trial-and-error using visual inspection of the displacements and resampled images to judge registration quality. The contribution of this paper is a more objective method to select these parameters.

3 Results

To investigate design choices of ANIMAL, we choose 40 T_1-weighted images from the ICBM data base (similar to images used in [5]). An arbitrary image is

selected to be the template and the other 39 are segmented into white matter, gray matter, cerebral spinal fluid, and background classes with non-brain voxels removed.

Consider first the outer loop. We normalize the data set using several choices for the numerical parameters and compute the entropy after each of the four iterations of the outer loop. The plot on the left of Figure 1 shows representative results using weight $a_2 = 1.0$, stiffness $a_3 = 0.9$, and choices of similarity cost ratio a_1 ranging from 0.1 to 1.0. The first item to note is that the total entropy

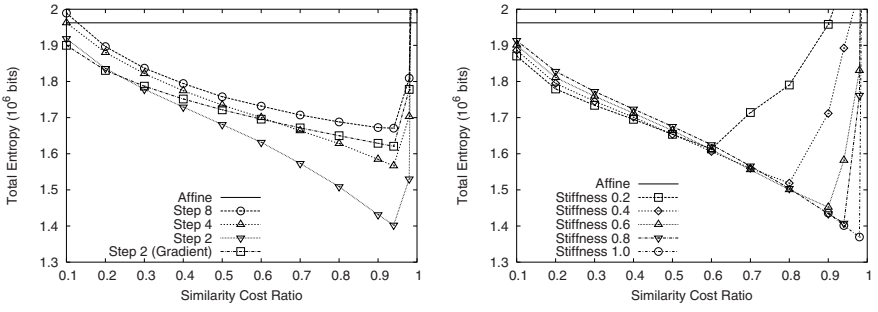

Fig. 1. Residual anatomical variability as measured by total entropy, H, for $N = 10$ individuals (first 10 of 39) after registration with ANIMAL, shown as a function of similarity cost ratio. Left plot shows results after each of the four iterations of the outer loop (weight=1.0, stiffness=0.9), along with the value for affine normalization, for reference. Right plot shows results using different stiffness values (weight=1.0) after three iterations of the outer loop.

is very large (even larger than obtained using the initial affine transformation) for similarity cost ratio of 1, indicating that the correction length penalty ψ in ANIMAL's objective function plays an important role in controlling the optimization. Transformations obtained with this ratio set to 1 (i.e. no contribution at all from ψ) contain much larger displacements, and are much less smooth and have more instances of folding (non-invertibility) than those obtained with similarity cost ratio < 1.

Secondly, it is clear that the anatomical variability is strictly reduced by each of the first three outer loop iterations, while the final iteration using gradient data degrades the result. Closer examination reveals the node thinning strategy as the culprit. For the iterations using intensity data, this heuristic retains nearly all the nodes lying in brain tissue, while skipping control-grid nodes 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 vertices (about 1/3 of the number of vertices in the previous outer iteration, which uses the same control grid), while all vertices participate in the smoothing. The effect is to smooth out the warp, degrading the data fit. Omitting the node thinning heuristic for the gradient data

fit does reduce the variability below that of the step 2 intensity fit. However, for this paper we focus on the results using the first three iterations of the outer loop.

We turn now to the three numerical parameters, running normalization experiments for weight values in the range 0.8-1.2, stiffness and similarity cost ratio values in the range 0-1. Figure 1 (right) shows typical plots of entropy for various stiffness and similarity cost ratios (weight 1.0). We can see that higher stiffness and similarity cost ratio values perform best, with a minimum near stiffness 1, similarity cost ratio 0.98. In fact, however, this is a local minimum, as we found that weight 0.8, stiffness 0.98, and similarity cost ratio 0.98 performs even better.

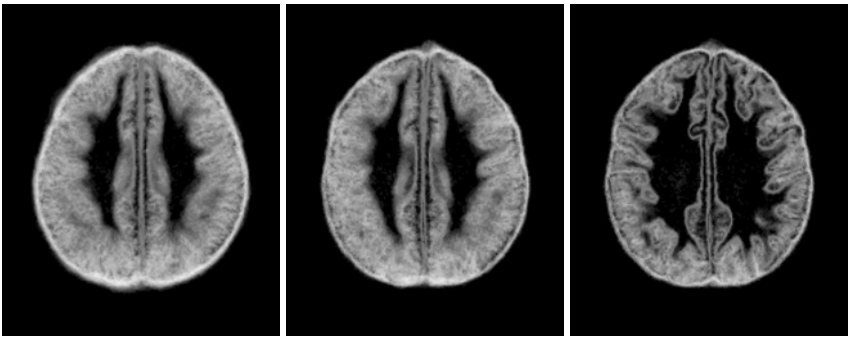

Fig. 2. Entropy maps ($N = 39$) from left-to-right: affine, classic ANIMAL, and tuned ANIMAL. Axial slice at $Z = 36$ shown (all registrations carried out in 3D). Voxels with more variability are brighter.

Figure 2 provides a visual illustration of the reduced anatomical variability in a set of 39 individuals, obtained using the tuned version of ANIMAL. The variability in the depth of many sulci is reduced, indicating that the sulci are better aligned.

Figure 3 shows intensity-averaged images which become sharper with tuning, a qualitative display of the improvement in aligning fine detail.

4 Discussion

As noted in Section 2.1, assessing variability using label consistency obviates the need for a ground truth segmentation. The variability of *any* label data can be measured using H, e.g. sulci, functional fields, etc. In order to assess the quality of anatomical normalization the labels should carry anatomical information. A labelling of sulci is a natural choice, which we have used in preliminary work and found broadly the same optimal parameters as those reported here. We choose instead to use tissue labels for two reasons: the labels can be obtained automatically, and they cover the entire brain. It is true that aligning gray and

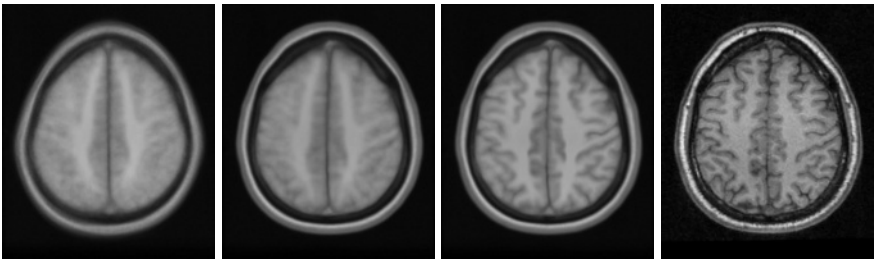

Fig. 3. Intensity-averaged images ($N = 39$) from left-to-right: affine, classic ANIMAL, and tuned ANIMAL. Axial slice at $Z = 46$ shown. The template is shown on the right, for comparison.

white matter does not guarantee the correct sulci are aligned. However, the converse holds: when the spatial normalization succeeds in aligning sulci, the surrounding gray and white matter tissues will be aligned and this will produce a lower variability measure. Our experiments demonstrate that variability (total entropy) of tissue labels is a useful performance measure.

We deliberately choose a performance measure based on a labelling rather than directly on the image intensity because the former allows inclusion of extra information in the labelling process. For example, the labels could be obtained manually or semi-automatically and automated procedures can bring in prior information such as the spatial distribution of tissues. Thus we are not simply measuring the same intensity correlations as the registration algorithm itself.

Another point to consider in assessing competing algorithms for spatial normalization is whether to have one or several measures of variability. As we have shown, a single measure enables optimization of the algorithm parameters. Prior works [16,6] have generated 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 that 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.

5 Conclusions

We have presented a strategy for evaluating the quality of a spatial normalization procedure on real data. 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, such as omitting the gradient data fit. We expect that our evaluation strategy would provide similar insights into other normalization methods.

This entropy measure can also be used to compare two competing methods of normalization, once each has been suitably tuned.

References

1. John Ashburner and Karl J. Friston. Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, 7(4):254–266, 1999.
2. Ruzena Bajcsy and Stane Kovačič. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46:1–21, 1989.
3. Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. and Mach. Int.*, 11(6):567–585, 1989.
4. G. E. Christensen, R. D. Rabbitt, and M. I. Miller. Deformable templates using large deformation kinematics. *IEEE Trans. Im. Proc.*, 5(10):1435–1447, 1996.
5. D. L. Collins and A. C. Evans. ANIMAL: Validation and applications of nonlinear registration-based segmentation. *Int. J. Pat. Rec. Art. Int.*, 11(8):1271–1294, 1997.
6. Fabrice Crivello, Thorsten Schormann, Nathalie Tzourio-Mazoyer, Per E. Roland, Karl Zilles, and Bernard M. Mazoyer. Comparison of spatial normalization procedures and their impact on functional maps. *Hum. Br. Map.*, 16:228–250, 2002.
7. Christos Davatzikos. Spatial normalization of 3D brain images using deformable models. *Journal of Computer Assisted Tomography*, 20(4):656–665, 1996.
8. B. Fischl, M. I. Sereno, R. B. H. Tootell, and A. M. Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Br. Map.*, 8(4):272–284, 1999.
9. I. D. Grachev, D. Berdichevsky, S. L. Rauch, S. Heckers, D. N. Kennedy, V. S. Caviness, and N. M. Alpert. A method for assessing the accuracy of intersubject registration of the human brain using anatomic landmarks. *NeuroImage*, 9:250–268, 1999.
10. P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, L. Collins, A. Evans, G. Malandain, and N. Ayache. Retrospective evaluation of inter-subject brain registration. *MICCAI*, LNCS v. 2208, pages 258–265, 2001.
11. Alexis Roche, Grégoire Malandain, and Nicholas Ayache. Unifying maximum likelihood approaches in medical image registration. *IJIST* 11:71–80, 2000.
12. P. E. Roland, Stefan Geyer, Katrin Amunts, Thorsten Schormann, Axel Schleicher, Aleksander Malikovic, and Karl Zilles. Cytoarchitectural maps of the human brain in standard anatomical space. *Human Brain Mapping*, 5:222–227, 1997.
13. Helmuth Steinmetz, Günter Fürst, and Hans-Joachim Freund. Cerebral cortical localization: Application and validation of the proportional grid system in MR imaging. *J. Computer Assisted Tomography*, 13(1):10–19, 1989.
14. J.-P. Thirion. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis* 2(3):243–260, 1998.
15. Paul Thompson and Arthur W. Toga. A surface-based technique for warping three-dimensional images of the brain. *IEEE Tran. Med. Im.*, 15(4):402–417, 1996.
16. S. K. Warfield, J. Rexilius, P. S. Huppi, T. E. Inder, E. G. Miller, W. M. Wells III, G. P. Zientara, F. A. Jolesz, and R. Kikinis. A binary entropy measure to assess nonrigid registration algorithms. *MICCAI*, LNCS v. 2208, pages 266–274, 2001.
17. Roger P. Woods, Scott T. Grafton, John D. G. Watson, Nancy L. Sicotte, and John C. Mazziotta. Automated image registration: II. intersubject validation of linear and nonlinear models. *J. Comp. Assist. Tom.*, 22(1):153–165, 1998.