NeuroImage 51 (2010) 214–220

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Evaluation of volume-based and surface-based brain image registration methods

Arno Klein [a,*], Satrajit S. Ghosh [b], Brian Avants [c], B.T.T. Yeo [d], Bruce Fischl [e,f,g], Babak Ardekani [h,i], James C. Gee [c], J.J. Mann [a], Ramin V. Parsey [a]

[a] New York State Psychiatric Institute, Columbia University, NY, NY 10032, USA
[b] Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA
[c] Penn Image Computing and Science Laboratory, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104–2644, USA
[d] Cognitive Neuroscience Lab, Harvard University, USA
[e] Athinoula A Martinos Center, Massachusetts General Hospital, USA
[f] Department of Radiology, Harvard Medical School, USA
[g] CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA
[h] Nathan Kline Institute, Orangeburg, NY 10962, USA
[i] New York University School of Medicine, NY, NY 10016, USA

ARTICLE INFO

Article history:
Received 23 November 2009
Revised 13 January 2010
Accepted 26 January 2010
Available online 1 February 2010

ABSTRACT

Establishing correspondences across brains for the purposes of comparison and group analysis is almost universally done by registering images to one another either directly or via a template. However, there are many registration algorithms to choose from. A recent evaluation of fully automated nonlinear deformation methods applied to brain image registration was restricted to volume-based methods. The present study is the first that directly compares some of the most accurate of these volume registration methods with surface registration methods, as well as the first study to compare registrations of whole-head and brain-only (de-skulled) images. We used permutation tests to compare the overlap or Hausdorff distance performance for more than 16,000 registrations between 80 manually labeled brain images. We compared every combination of volume-based and surface-based labels, registration, and evaluation. Our primary findings are the following: 1. de-skulling aids volume registration methods; 2. custom-made optimal average templates improve registration over direct pairwise registration; and 3. resampling volume labels on surfaces or converting surface labels to volumes introduces distortions that preclude a fair comparison between the highest ranking volume and surface registration methods using present resampling methods. From the results of this study, we recommend constructing a custom template from a limited sample drawn from the same or a similar representative population, using the same algorithm used for registering brains to the template.

Introduction

Brain images are registered to one another (or to an average template) to establish correspondences of all kinds, such as across structures, patterns of functional activity, physiological data, and connectivity data. These correspondences enable comparison across time, task, and population. Brain image registrations are performed either on image volumes in their "native" space, or on surface representations of the brain.

Surface registration methods require computationally intensive extraction of a cortical surface, and may not be accurate for topologically different brains (such as lesioned and other pathological cases), but have been demonstrated to perform accurately under many conditions, including recent studies comparing surface features with cytoarchitectonic data (Rademacher et al., 1993; Hinds et al., 2008; Fischl et al., 2008). Significant advantages of performing registrations on a surface compared with in a volume include computational efficiency (less to register, one less degree of freedom), and distances along the cortical surface are more faithfully represented as geodesic distances along a surface rather than Euclidean distances across, for example, banks of a sulcus.

Most prior attempts to compare volume and surface registration methods have used function to gauge registration accuracy and compared a nonlinear surface-based spherical registration method with affine, Talairach-based linear or piecewise linear registration (Anticevic et al., 2008, Desai et al., 2005; Fischl et al., 1999; Thompson and Toga, 1996). Hinds et al. (2008) compared the quality of two atlases, one constructed using FreeSurfer spherical registration and the other using a single volume-based nonlinear registration method, according to the cumulative probability of a region (V1); this atlas comparison is a more indirect comparison than evaluating pairwise registrations.

* Corresponding author.
E-mail address: arno@binarybottle.com (A. Klein).

1053-8119/$ – see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.neuroimage.2010.01.091

In this study, we attempt to directly compare surface with volume registration methods. For the volume registration methods, we selected SyN (Avants et al., 2008) and ART (Ardekani et al., 2005a, b), the only methods that attained top rank for all tests and for all four label sets and populations used in a recent, large-scale evaluation of brain image registration methods (Klein et al., 2009).

For the surface registration methods, we selected FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) (Fischl et al., 1999) and Spherical Demons (Yeo et al., 2010). FreeSurfer is the most widely used of the fully automated, surface-based, brain image analysis software packages that perform registration without requiring landmarks. FreeSurfer was recently shown to outperform affine registration and SPM Normalize (Ashburner and Friston, 1999) when applied to a pediatric population (Ghosh, personal communication). Spherical Demons is a recently introduced extension of the Demons algorithm (Thirion, 1998) to the sphere, which is reported to have comparable performance to FreeSurfer but runs at least an order of magnitude faster (Yeo et al., 2010). Other popular freely available surface registration methods include Caret (http://brainmap.wustl.edu) (Van Essen et al., 2001) and BrainVisa (http://brainvisa.info) (Cointepas et al., 2001). They were not included in this study because Caret still requires manually assigned landmarks (personal communication with Van Essen and Dierker) and we are not aware of any means to apply a nonlinear transform to an arbitrary set of labels within BrainVisa.

We conducted more than 16,000 registrations between 40 brain images, either directly to one another or via templates, with the brains represented as either volumes or as surfaces. The registration transforms were then applied to manually labeled versions of these images (on volumes and on surfaces) to evaluate registration accuracy. We used permutation tests to compare registration performance to first select the top-ranking volume and surface registration methods, and then we compared these selections with one another. The initial set from which we made our selection was the following: SyN and ART on brain images with and without skulls, SyN, FreeSurfer, and Spherical Demons via custom templates, and FreeSurfer via its default atlas.

Materials and methods

In this section, we describe the brain image and label data, custom template construction, selection of image pairs to be registered to one another, and our evaluation measures and analysis method. We performed these latter steps on the LPBA40 data (see below) using an OSX system (Mac Pro 2-Quad-Core (8-processor) Intel Xeon, 3 GHz, 6 GB RAM) with a 10.5 operating system, and on the FS40 data (see below) using a computer cluster at the Martinos Center at the Massachusetts General Hospital (MGH), where each node runs CentOS 64-bit and has two quad-core Xeon 5472 3.0 GHz CPUs and 32GB of RAM. The FS40 data are not publicly available and therefore all operations on these data were performed on the MGH computing cluster. Custom Python (http://www.python.org) and Matlab (http://www.mathworks.com) software programs called the different registration programs to process thousands of pairs of images, evaluated the results, and produced the graphics in the Results section.

MRI data and manual labels

LPBA40 data

We used the magnetic resonance image (MRI) volumes and corresponding manually labeled volumes for 40 normal subjects (20 male, 20 female, ages 19–40 years, mean 29.2 years) used to construct the LONI Probabilistic Brain Atlas (LPBA40) at the Laboratory of Neuro Imaging (LONI) at UCLA (Shattuck et al., 2008) (available online at http://www.loni.ucla.edu/Atlases/LPBA40). The T1-weighted images ($256\times256\times124$ voxels (volume elements); $0.86\times0.86\times1.5$ mm^3/voxel for 38 subjects, $0.78\times0.78\times1.5$ mm^3/voxel for 2 subjects; 10–12.5 ms TR; 4.2–4.5 ms TE; 20° flip angle) were preprocessed according to existing LONI protocols to produce skull-stripped brain volumes. These volumes were aligned to the MNI305 atlas (Evans et al., 1993) using rigid-body transformation to correct for head tilt and reduce bias in the manual labeling process. This produced a transform from native space to labeling space and an associated inverse transform. In each of the 40 subjects, 56 structures were manually labeled according to custom protocols (http://www.loni.ucla.edu/Protocols/LPBA40) using the BrainSuite software (http://brainsuite.usc.edu). Brain masks were constructed from the manual labels and projected back to the native space to produce brain-only MRI volumes. These volumes were then corrected for non-uniformity using BrainSuite's Bias Field Corrector. Sulci were used as label boundaries; both gray and white matter voxels between these boundaries were labeled in the original dataset.

We used a pre-released update of the label volumes that corrected for errors in two of the subjects, particularly with the right putamen. We removed non-gray matter regions for each brain using a gray matter mask for that brain constructed from the probabilistic tissue-class image of that brain provided in the LPBA40 set (Fig. 1, upper panel).

FS40 data

We also used manually labeled surfaces from MGH for two reasons. First, we discovered some consistent labeling errors in the LPBA40 data (unlabeled gray matter at the depths of sulcal folds) due to difficulties in labeling convoluted sulci and gyri in the 3-D volume. Second, we wanted to compare the results obtained using volume-labeled manual data (LPBA40) with results obtained using surface-labeled manual data (FS40). The FS40 data consist of MRI volumes and corresponding manually labeled cortical surfaces (generated by FreeSurfer) for 40 subjects (10 young, 10 middle age, 10 elderly, 10 diagnosed with Alzheimer's-related dementia; 14 male, 26 female, ages 18–87 years, mean 56 years). These are the same subjects used to construct FreeSurfer's default "aparc" cortical parcellation atlas (http://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation). The original T1-weighted images (four MPRAGE-protocol sessions with 9.7 ms TR[1]; 4 ms TE; 10° flip angle) or their FreeSurfer-extracted surfaces were used by the registration algorithms. In each of the 40 subjects, 35 structures were manually labeled according to custom protocols (34 described by Desikan et al. (2006) and the insula labeled by Sarah Green of Boston University; see Fig. 1, lower panel).

Custom template construction

We divided the LPBA40 data into two random sets of 20 images each, from which we constructed two custom templates using SyN, FreeSurfer, and Spherical Demons. The custom templates constructed by these different software packages all attempt to iteratively refine nonlinear coregistrations of a group of individual brain images to create a group average, sometimes referred to as an optimal average template. We repeated this for the FS40 data, resulting in a total of 12 custom templates. The ANTS toolkit (containing SyN) comes with a program for building optimal average templates with distributed or serial computing (buildtemplateparallel.sh (Avants et al., 2010)). The algorithm builds an average shaped brain within the diffeomorphic space, and is also capable of building an average shape and appearance brain as in this work, to reduce dependence on the topological idiosyncracy of any individual brain. The program calls SyN to register a group of brain images to one another in multiple stages to refine co-registration of the constituent images. For example, in the first stage, all of the images are affine-registered to one brain, and this collection is intensity-averaged to create a single, blurry brain

[1] According to Siemen's pre-2000 definition of TR: time between RF excitations.

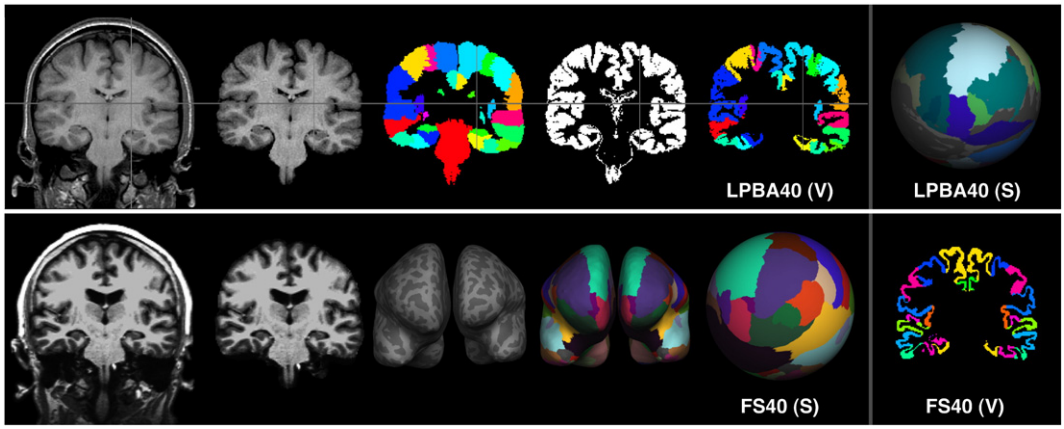

Fig. 1. Volume-labeled and surface-labeled brain image data. *Upper panel (LPBA40 sample)*: For the volume-labeled brain image data, we used updated versions of the 40 brains used to construct the LONI Probabilistic Brain Atlases (Shattuck et al. 2008, 56 labeled regions). Left to right: T1-weighted MRI coronal slice, extracted brain, manual labels, gray matter mask, cortical gray matter label volume (V), and the left hemisphere labels resampled onto the unit sphere (S, with different colors). Notice that some of the surface is missing labels (gray). *Lower panel (FS40 sample)*: For the surface-labeled brain image data, we used the 40 brains used to construct FreeSurfer's default cortical parcellation atlas (Desikan et al. 2006, 35 labeled regions). Left to right: T1-weighted MRI coronal slice, extracted brain, FreeSurfer-generated surfaces of the left and right hemispheres seen from the front, the same surfaces with manual labels, the left hemisphere labels projected on the unit sphere, and the labels resampled in the brain volume (with different colors). (Note: images are not at the same scale, the colors in the upper and lower panels do not correspond, and neither do the colors in the rightmost images correspond with those to the left.)

image. Then each of the original brain images is nonlinearly registered to this average to create a new average. This second step is repeated multiple (in our case, three) times, and the result is an optimal average template. Spherical Demons also comes with a script for constructing templates in an analogous manner. FreeSurfer does not come with such a script, but a custom Python script calling FreeSurfer commands for constructing such a template may be downloaded from http://www.mindboggle.info/papers/.

Registration pairs

For each of the registration methods, we registered a source image to a target image. The source and target images were drawn from the same set used to construct one of the two custom templates. For example, each of the 20 images that SyN used to construct one template would be registered using SyN to each one of the other 19 images, either directly or via the second template, also generated by SyN. Each method was run on 760 pairs.

Selection Test 1 (V-V)

Test 1 was conducted to select the top-ranking volume registration method and the top-ranking surface registration method for comparison against each other in subsequent tests. For volume registration, we used SyN and ART to register each of the 760 pairs of (brain-only as well as whole-head) LPBA40 images directly to one another, and SyN to register each pair via the appropriate custom template (see above and Supplementary Fig. 1). For surface registration, we used FreeSurfer and Spherical Demons to register each pair via the appropriate custom template (see above and Supplementary Fig. 1). FreeSurfer was intended to be used with its own atlas, so we performed the same registrations via the FreeSurfer atlas as well. We conducted all evaluations for Test 1 on transformed LPBA40 volume data, as in the study of Klein et al. (2009) (see below):

Test 1 (V-V): volume labels, volume evaluations

Comparison Tests 2–4 (S-S, S-V, V-S)

For Tests 2 through 4, we selected the top-ranking volume and top-ranking surface registration method as determined by the Selection Test 1 (V-V): SyN, FreeSurfer, and Spherical Demons, each with its custom templates. We ran these three methods again to account for every combination of surface and volume labels and evaluations (see Table 1):

Test 2 (S-S): surface labels, surface evaluations
Test 3 (S-V): surface labels, volume evaluations
Test 4 (V-S): volume labels, surface evaluations

Resampling Test 5

To evaluate how well the individual anatomical regions register to one another, we applied each source-to-target transformation to the corresponding source labels (with nearest-neighbor interpolation) and compared the resulting transformed source labels to the manual labels of the target[2].

Volume-based evaluation is straightforward for volume registration methods (SyN and ART), but in order to apply it to surface registration methods, the volume labels had to be resampled to the surface, and after surface registration, they had to be resampled again to the target volume. The FreeSurfer commands (see Appendix A) that resample labels to the target volume (for FreeSurfer and Spherical Demons) automatically propagate the labels through a target gray matter mask created by FreeSurfer. To emulate FreeSurfer's propagation of labels through its target gray matter mask, we propagate (SyN- and ART-transformed) labels through the same mask using a fast marching level set algorithm supplied by the ANTS toolbox. The fast marching method geometrically propagates labels to the closest unlabeled points along the cortical surface, rather than through three-dimensional space. Experiments verified that this approach yields superior results when compared to labeling based on a Euclidean distance transform, which–in some cases–will not respect distance along the surface. For instance, labels across sulcal boundaries may be nearby under the Euclidean distance transform but not at all proximal when distances are measured across the cortex.

Surface-based evaluation is straightforward for surface registration methods but not for volume registration methods because of double resampling in the latter case, as above. Therefore, Test 1 (V-V) is useful for comparing volume registration methods with one another

[2] Note that some structures, such as the cerebellum in the LPBA40 data, are not included in the de-skulled brain images but are included in the label images; they therefore do not influence the computation of the transformations but their labels are transformed and are included in the evaluation.

Table 1
Overview of tests: volume and surface labels, registration, and evaluation.

	Labels	Registration	Evaluation
Test 1	V	V*	V
	V	S	V
Test 2	S	S*	S
	S	V	S
Test 3	S	S*	V
	S	V	V
Test 4	V	V*	S
	V	S	S

We conducted four tests in this study. We evaluated volume (V) and surface (S) registration methods using either manually labeled volumes (LPBA40) or surfaces (FS40), and volume or surface evaluation measures. Test 1 is useful for comparing volume registration methods with one another (VVV) and Test 2 is useful for comparing surface registration methods with one another (SSS). Tests 1 and 2 are not useful for comparing volume and surface registration methods, because they compare methods that perform all operations in the same space (VVV and SSS) with methods that require double resampling to register in a space different from the label space and evaluation space (VSV and SVS). Tests 3 and 4 correct for this by admitting a single resampling from one space to another (such as Test 4: VVS and VSS).
Higher ranks were obtained for the cases where there is no resampling prior to registration (V-V- and S-S-).
An asterisk is placed next to the registration type that achieved a higher rank according to permutation tests.

and Test 2 (S-S) is useful for comparing surface registration methods with one another. Tests 1 and 2 are not useful for comparing volume with surface registration methods, however, because they compare methods that perform all operations in the same space with methods that require double resampling to register in a space different from the label space and evaluation space. Tests 3 (S-V) and 4 (V-S) attempt to correct for this by admitting a single resampling from one space to another for all methods.

For the final test, Test 5, we wanted to determine an upper bound for the resampling error resulting when resampling labels from a volume to a surface or the reverse. We used the same FreeSurfer commands used in the previous tests (see Appendix A) to resample each LPBA40 brain's volume labels to a surface then back to its volume, and compared the resulting double-resampled labels against the original volume labels. We also resampled each FS40 brain's surface labels to a volume then back to its surface, and compared the resulting double-resampled labels against the original surface labels.

Evaluation measures

We evaluated source-to-target registration accuracy by comparing the transformed source labels with the corresponding manual labels in the target. These evaluation measures assume that the manual label sets are correct, or "silver standards." Overlap quantifies some fraction of transformed source S and target T points (voxels or pixels) where their labels agree. For information on overlap measures, including cases for multiple and fractional labels, see Crum et al. (2005). Our overlap measure is the mean overlap, MO, also called the Dice coefficient. It is the intersection between two similarly labeled regions r in S and T divided by the mean volume of the two regions, where || indicates volume computed as the number of voxels. MO can be summed over a set of multiple labeled regions:

$$\mathrm{MO} = \frac{2\sum_r |S_r \cap T_r|}{\sum_r (|S_r| + |T_r|)}$$

For our surface evaluations, we also computed a modified Hausdorff distance (Dubuisson and Jain, 1994) between corresponding label boundaries (on a surface representation of the cortex):

$$\mathrm{DE}(S,T) = \sum_r \mathrm{mean}(\mathrm{mean}_{s\in Sr}(\inf_{t\in Tr}(d(s,t))), \mathrm{mean}_{t\in Tr}(\inf_{s\in Sr}(d(s,t))))$$

where inf represents infimum and d represents the shortest distance on the cortical surface mesh between two points.

Permutation tests

Testing for significant differences in the performance of the registration methods is not trivial because each brain is reused multiple times, and independence of observations cannot be assumed. For example, each of the 20 brain images in each set was registered to the 19 others in its set, and was also the registration target of these 19 others. To get around this issue of non-independency, we ensured that a brain does not appear in more than one brain pair by random selection without replacement. We conducted a permutation test on a small sample of 20 such brain pairs (10 pairs per set), and repeated this test on multiple such samples for all four tests. We performed permutation tests to determine if the means of the evaluation measures (overlap or distance values averaged across labeled regions) are the same for the different registration methods. These tests are constructed after Menke and Martinez (2004) and according to the following permutation algorithm:

1. Select a (new) random sample of P brain pairs with non-repeating brains.
2. Select a pair of registration methods (two vectors of P evaluation values).
3. Subtract the two vectors; compute the mean difference D.
4. Select a subset of the elements from one of the vectors.
5. Swap this subset across the two vectors.
6. Subtract the vectors; compute the mean difference Dp.
7. Repeat steps 4–6 N times.
8. Count the number of times n where $\mathrm{abs}(Dp) \geq \mathrm{abs}(D)$.
9. Compute the exact p-value: $p = n / N$.
10. Repeat steps 1–9 T times; compute the fraction of times $p \leq 0.05$.

For each of $T = 10{,}000$ iterations, we selected a new random sample of $P = 20$ brain pairs (each brain was used only once), ran $N = 1000$ permutations for that sample (to keep the duration of the tests under 24 h), and computed a single exact p-value. The fraction of the resulting 10,000 p-values less than or equal to 0.05 between any two registration methods is then compared with the fractions computed for all other pairs of methods to determine relative performance across the methods. Registration methods within ranks 1, 2, and 3 have positive mean p-value percentages lying within 1, 2, and 3 standard deviations of the highest mean p-value percentage, respectively.

Results

Table 1 presents whether a volume or surface registration method obtained a higher ranking for each of the four tests, according to permutation tests. Higher ranks were obtained for the cases where there is no resampling prior to registration (see Table 1).

Custom template consistency

For custom template-based registrations in Test 1 (V-V) (SyN, FreeSurfer, and Spherical Demons), the average overlap obtained with one template was within 1 standard deviation of the average obtained with the second template. This result assured us that the custom templates were not idiosyncratic representations of the subject pool and that their results may be pooled together for the evaluation. All

Table 2
Permutation test ranking of the registration methods (Test 1 with LPBA40 data).

	Volume registration methods	μ (SD)
Rank 1	SyN with custom templates	0.25 (0.27)
Rank 2	SyN	−0.38 (0.18)
	ART	−0.22 (0.28)
	Surface registration methods	**μ (SD)**
Rank 1	FreeSurfer with custom templates	0.09 (0.13)
	Spherical Demons with custom templates	0.01 (0.05)
Rank 2	FreeSurfer with default atlas	−0.10 (0.12)

This table lists the registration methods that attained the top three ranks according to permutation tests on volume overlap values (averaged across labeled regions) for Test 1 (V-V) with LPBA40 data and FreeSurfer-generated target masks (whole-head volume registrations are not included here). Registration methods within ranks 1, 2, and 3 have positive mean *p*-value percentages (μ) lying within 1, 2, and 3 standard deviations (SD) of the highest mean *p*-value percentage, respectively (see text). Tests 2 through 4 were run only on these top-ranking registration methods (SyN, FreeSurfer, and Spherical Demons, all with custom templates). Refer to Table 3 for a more sensitive ranking of the surface registration methods, since for the Test 1 results, the surface registration methods underwent a double resampling to register volume labels as surfaces and then evaluate them as volumes (see text and Supplementary Table 1).

the following results represent all 760 registrations using both templates for a given method.

Results of Selection Test 1 (V-V)

Table 2 lists the volume and surface registration methods that attained the top three ranks according to permutation tests. For the volume registration methods, SyN performed better with its custom templates than without the templates, and for direct pairwise registration (without a template), both SyN and ART performed within the same rank and their registrations were more accurate for brain-only images than for whole-head images (Supplementary Figs. 2 and 3). For the surface registration methods, FreeSurfer registration with custom templates (constructed as part of this study using 20 of the FS40 brains) performed better on average than with its default atlas (previously constructed using all 40 of the FS40 brains). To determine whether the default FreeSurfer template is smoother than the custom templates, which might allow it to accommodate larger registration errors, we sampled the curvature files for the subjects used to construct a given template into an atlas space (using the FreeSurfer command: "mris_preproc –target fsaverage –meas curv…"), then estimated the FWHM on the spherical surface (using FreeSurfer's "mris_fwhm" command). When we used all 40 FS40 subjects, corresponding to the default template, we obtained an average FWHM of 3.03, whereas when we used the 20 FS40 subjects in each of the custom templates, we obtained a slightly lower average of 2.96, suggesting that the default template is slightly smoother (see Supplementary Table 5). The results of the comparison between FreeSurfer and Spherical Demons are consistent with that of Yeo et al. (2010) in that Spherical Demons (with custom template) compared favorably with FreeSurfer (with default template) in that study on a similar data set.

Results of Comparison Test 2 (S-S)

Test 2 (S-S) provides a more sensitive ranking of the surface registration methods, since it does not involve a double resampling of the surface labels, as in Test 1 (V-V). Table 3 presents permutation test results of Test 2's *surface* evaluations of *surface* registration methods (FreeSurfer and Spherical Demons) with custom templates conducted on *surface* (FS40) labels. FreeSurfer with a template ranked higher than Spherical Demons with a template according to mean surface overlap and modified Hausdorff distance measures (Table 3).

Comparing results of Tests 1–4 with results of the Resampling Test 5

Supplementary Tables 1, 2, 3, and 4 present some of the results of Tests 1, 2, 3, and 4, respectively. Supplementary Tables 1 and 2 also compute Δoverlap or Δdistance, the maximum absolute difference between average overlap or distance calculated for SyN and that for the surface registration methods. These values are compared with the average resampling errors computed for the corresponding label set. The values for Tests 3 (S-V) and 4 (V-S) are not compared with resampling errors, because the space in which the evaluations were conducted are different in the two cases (volume vs. surface or the reverse).

For Test 1 (V-V), using LPBA40 volume labels, Δoverlap is less than the resampling error for every label, which means that the resampling error is too great to distinguish between the performance of the top-ranking volume and surface registration methods. Likewise, for Test 2 (S-S), using FS40 surface labels, Δoverlap and Δdistance are less than the corresponding resampling errors for almost every label, which again means that for almost every region the resampling error is still too great to distinguish between the performance of volume and surface registration methods. The only exceptions are pericalcarine and superior parietal regions, whose values are greater by at least 1 standard deviation of the resampling values.

Discussion

After performing thousands of registrations between brain images (as surfaces and as volumes), we confirmed that removing non-brain matter aids brain volume registration, custom-made optimal average templates improve registration over direct pairwise registration, and resampling errors introduced by converting volume labels to surfaces or surface labels to volumes can be used to make a fair comparison between volume and surface registration methods using present resampling methods. Evaluation software, data, and updated information will be made publicly available via the http://www.mindboggle.info/papers/ Web site.

Caveats

The general caveats that must be taken into account when evaluating registration data are the same as those discussed in the study of Klein et al. (2009); briefly, they are the following:

1. Image correspondence is often mistaken for anatomic correspondence (Crum et al., 2005).
2. A given brain may not be well represented by the majority of the set of brains to which it is being compared.
3. Label overlap and boundary distance measures ignore misregistration within a labeled region.
4. Our evaluation measures rely on information (labels) which is not directly included in the images, which is good for evaluating the registrations, but they do not inform us about the intrinsic properties of the spatial transformations.

Table 3
Permutation test ranking of the *surface* registration methods (Test 2 with FS40 data).

	Surface overlap test	Hausdorff distance test
Rank 1	FreeSurfer with templates	FreeSurfer with templates
Rank 2	Spherical Demons with templates	Spherical Demons with templates

This table is constructed in the same manner as Table 2, but ranks surface overlap values (left) and modified Hausdorff distances (right) between *surface*-warped source labels and their corresponding target labels, again averaged across labeled regions, but using the *FS40* data. The volume registration methods are not included in this table because for Test 2 (S-S), these methods underwent a double resampling to register surface labels as volumes and evaluate them as surfaces (see text and Supplementary Table 2).

5. Nonlinear registration to average templates affects different brain regions in different ways that lead to relative distortions in volume that are difficult to predict (Allen et al., 2008).

There are three primary caveats specific to both the study of Klein et al. (2009) and the present study. First, each brain image was labeled only once, and had to be treated as a "silver standard." Second, we could only evaluate each algorithm using the software parameters that were recommended in the software documentation or by their authors. We fully expect that each registration algorithm could perform better given the opportunity to experiment with these settings. Third, most of our tests were conducted on healthy subjects and should not be generalized to pathological cases. Surface registration methods that seek to preserve topology might not be suitable in cases such as lesions, which introduce topological differences. With regard to preprocessing, surface registration methods require cortical surface extraction, and volume registration methods are more accurate after extracting brains from the images; many researchers routinely run both of these procedures without evaluation.

There is an additional caveat concerning prior training. FreeSurfer has been trained on the FS40 data, and because these data are private, this introduces some bias in the evaluation study because other methods do not train on these data.

Recommendations

Given that SyN performed better with its custom templates than without the templates, and FreeSurfer performed better with its custom templates than with its default atlas, we recommend where possible to construct a custom template from a limited sample drawn from the same or a similar representative population, using the same algorithm used for registering brains to the template. We do not have evidence to advocate the use of general-purpose brain atlases for nonlinear registration over direct source-to-target registration for normal brains. And it is clear from the volume registrations that registering brain-only images is more accurate than registering whole-head images. Brain-only registration not only excludes extraneous features such as eyes from the optimization but also focuses the per-iteration effort on regions of the brain that are relevant for this evaluation.

The resampling tests demonstrate that, for almost every region, the resampling error is too great to distinguish between the performance of top-ranking volume and surface registration methods (SyN, FreeSurfer, and Spherical Demons, all using customized optimal average templates). Based on these results, it may not be possible to directly compare evaluations of these surface and volume registration methods using the present resampling methods, when considering the full surface or full volume or the full extent of their label boundaries.

The performance differences in this evaluation are determined by how well the boundaries of labeled structures correspond to one another after registration. These boundaries are exactly where both uncertainties lie and where interpolation effects are most prominent. However, by restricting an evaluation to appropriate non-boundary landmarks, it might be possible to circumvent resampling altogether and directly compare volume-based and surface-based registration methods. Landmark selection may bias a volume-based or surface-based registration approach, just as, for example, the landmarks (label boundaries) used in a volume-based labeling protocol could favor a volume-based registration method. One impartial selection of landmarks could simply be a uniform sampling of points in a brain volume or on the unit sphere. Such a selection and tracking of points might be relatively immune to a given labeling protocol since the vast majority of points will lie inside regions and not on their (disputed) boundaries. This last point cannot be overemphasized, because the space in which brains are labeled bias, to an unknown degree, operations performed on brain images using these labels, and evaluations conducted on the results of performing these operations.

Acknowledgments

We would like to thank the reviewers for their comments, David Shattuck for making available an early second release of the publicly available LPBA40 whole-head MR and label data, and Douglas Greve for advice on estimating smoothness of the FreeSurfer spherical templates. The first author is grateful to his colleagues in the Division of Molecular Imaging and Neuropathology, and, as always, to his two closest colleagues Deepanjana and Ellora. This work was partially funded by the National Institutes of Health through NIH grant R01 MH084029. Satrajit Ghosh was partially supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) grant R03 EB008673. The contributions to this paper by Babak A. Ardekani were supported by grant R03 EB008201 from the NIBIB and the National Institute of Neurological Disorders and Stroke (NINDS).

Bruce Fischl would like to acknowledge support provided in part by the National Center for Research Resources (P41-RR14075, and the NCRR BIRN Morphometric Project BIRN002, U24 RR021382), the National Institute for Biomedical Imaging and Bioengineering (R01 EB006758), the National Institute on Aging (AG02238), the National Institute for Neurological Disorders and Stroke (R01 NS052585-01) as well as the Mental Illness and Neuroscience Discovery (MIND) Institute, and is part of the National Alliance for Medical Image Computing (NAMIC), funded by the National Institutes of Health through the NIH Roadmap for Medical Research, grant U54 EB005149. Additional support was provided by The Autism & Dyslexia Project funded by the Ellison Medical Foundation.

Appendix A

All of the software packages under comparison are freely available via the Internet, are easy to install, and are under active development. ANTS (SyN) requires ITK (http://www.itk.org) and Spherical Demons requires Matlab. FreeSurfer has the most extensive documentation and an online wiki (http://surfer.nmr.mgh.harvard.edu/fswiki), ANTS has a growing set of online tutorials (http://www.picsl.upenn.edu/ANTS/), and ART and Spherical Demons have very limited documentation. SyN and FreeSurfer provide the most command-line options. If time is a constraint, ART is the fastest of the four (about 15 minutes), followed by SyN (about an hour), followed by FreeSurfer and Spherical Demons (about a day, including all preprocessing leading to cortical surface extraction). Python and Matlab programs containing all command-line arguments (including those below) may be downloaded from this article's supplementary Web site (http://www.mindboggle.info/papers/).

ART was developed by Ardekani et al. (2005) at the Nathan Kline Institute, NY. The executables can be downloaded at http://www.nitrc.org/projects/art/.

ART uses local normalized cross-correlation between the source and target images as its similarity measure. It determines a displacement vector field defined for each grid point on the target image using a non-parametric free-form multi-resolution approach. The displacement vector field obtained at each resolution level is regularized by median and low-pass filtering. Prior to nonlinear registration, there are options for the program to determine a linear rigid-body (6-parameter) registration between the target and subject images followed by a linear affine (12-parameter) registration (Ardekani et al., 2005a,b).

Warp: 3dwarper -trg <target>.img -sub <source> -acpc -A -sd 8.0 -v
Reslice: applywarp3d -nn -w <transform>.nii
(In the present version, we were not able to output files with desired names, so we renamed files as they were generated.)

SyN was developed by Avants et al. (2008) at the University of Pennsylvania. Source code and executables are available as part of the

ANTS toolkit (http://www.picsl.upenn.edu/ANTS/). The symmetric normalization (SyN) methodology uses a symmetric parameterization of the shortest path of diffeomorphisms connecting two neuroanatomical configurations. The SyN formulation uses a bidirectional gradient descent optimization which gives results that are unbiased with respect to the input images. SyN also provides forward and inverse continuum mappings that are consistent within the discrete domain and enables both large and subtle deformations to be captured. Specific performance characteristics depend upon the range of similarity metrics chosen for the study and the velocity field regularization.

Warp: ANTS 3 -m PR[<target>.nii, <source>.nii, 1, 2] -o <output transform>.nii -r Gauss[2,0] -t SyN[0.5] -i 30x99x11 -use-Histogram-Matching

Reslice: WarpImageMultiTransform 3 <labeled source>.nii <output labels>.nii -R <target>.nii <transform>Warp.nii <transform>Affine.txt -use-NN

FreeSurfer was developed by Fischl et al. (1999) at the Massachusetts General Hospital. Source code and executables are available from http://surfer.nmr.mgh.harvard.edu/fswiki. The registration algorithm in FreeSurfer uses the weighted sum of squared differences dissimilarity measure. FreeSurfer represents deformation by a displacement field on the sphere. The degrees of freedom are twice the number of vertices since the vertices are constrained to be on the sphere. For the regularization, FreeSurfer penalizes metric distortion (changes in triangular edge lengths) and areal distortion (changes in triangular area). It also has an areal expansion term that penalizes small triangles and encourages an invertible transformation. Optimization occurs in a multiscale fashion, where coarser features are used, followed by finer features. In particular, the coarser scale uses average convexity, while the finer scale uses the mean curvature of the surface. Surface registrations were performed by mris_register and mri_surf2surf. For volume registrations, the volume labels were sampled on surfaces using mri_convert and mris_sample_parc, and the surface registration results were sampled in the target volumes using mri_aparc2aseg and mri_convert.

Spherical Demons was developed by Yeo et al. (2010) at the Massachusetts General Hospital as an extension to FreeSurfer. Matlab source code (run in place of mris_register above) is available from http://yeoyeo02.googlepages.com/sphericaldemonsrelease and developments are underway to create a C++ version for ITK. Spherical Demons uses the weighted sum of squared differences dissimilarity measure. Spherical Demons represents deformation by a diffeomorphic displacement field on the sphere. The degrees of freedom are twice the number of vertices since the vertices are constrained to be on the sphere. For the regularization, Spherical Demons uses the “Demons” hierarchical regularization (Cachier et al., 2003). The resulting regularization can be shown to be a Sobolev type of regularization of the displacement field, and is in practice approximated by iterative smoothing of the displacement field. Optimization occurs in a multiscale fashion, where coarser features are used, followed by finer features. In particular, the coarser scale uses curvature of the inflated surface, followed by average convexity and finally the mean curvature of the surface.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2010.01.091.

References

Allen, J.S., Bruss, J., Mehta, S., Grabowski, T., Kice Brown, C., Damasio, H., 2008. Effects of spatial transformation on regional brain volume estimates. NeuroImage 42 (2), 535–547. doi:10.1016/j.neuroimage.2008.05.047 (August 15).

Anticevic, A., Dierker, D.L., Gillespie, S.K., Repovs, G., Csernansky, J.G., Van Essen, D.C., Barchm, D.M., 2008. Comparing surface-based and volume-based analyses of functional neuroimaging data in patients with schizophrenia. Neuroimage 41 (3), 835–848.

Ardekani, B., Braun, M., Hutton, B.F., Kanno, I., Iida, H., 1995. A fully automatic multimodality image registration algorithm. J. Comput. Assist. Tomogr. 19 (4), 615–623 (August).

Ardekani, B.A., Guckemus, S., Bachman, A., Hoptman, M.J., Wojtaszek, M., Nierenberg, J., 2005. Quantitative comparison of algorithms for inter-subject registration of 3D volumetric brain MRI scans. J. Neurosci. Methods 142 (1), 67–76. doi:10.1016/j.jneumeth.2004.07.014 (March 15).

Ashburner, J., Friston, K.J., 1999. Nonlinear spatial normalization using basis functions. Hum. Brain Mapp. 7 (4), 254–266 10.1002/(SICI)1097-0193(1999)7:4<254::AID-HBM4>3.0.CO;2-G.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41.

Avants, B.B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J.C., 2010. The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49 (3), 2457–2466.

Cachier, P., Bardinet, E., Dormont, D., Pennec, X., Ayache, N., 2003. Iconic feature based nonrigid registration: the PASHA algorithm. Comput. Vis. Image Underst. 89 (2-3), 272–298. doi:10.1016/S1077-3142(03)00002-X.

Cointepas, Y., Mangin, J.-F., Garnero, L., Poline, J.-B., Benali, H., 2001. BrainVISA: software platform for visualization and analysis of multi-modality brain data. NeuroImage 13 (6), 98 (June).

Crum, W.R., Camara, O., Rueckert, D., Bhatia, K.K., Jenkinson, M., Hill, D.L.G., 2005. Generalised overlap measures for assessment of pairwise and groupwise image registration and segmentation. Med. Image Comput. Comput.-Assist. Interv. : MICCAI 8 (Pt. 1), 99–106.

Desai, R., Liebenthal, E., Possing, E.T., Waldron, E., Binder, J.R., 2005. Volumetric vs. surface-based alignment for localization of auditory cortex activation. NeuroImage 26 (4), 1019–1029. doi:10.1016/j.neuroimage.2005.03.024 (July 15).

Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31 (3), 968–980 (July 1).

Dubuisson, M.-P., and Jain, A.K., 1994. A modified Hausdorff distance for object matching. In Pattern Recognition, 1994. Vol. 1—Conference A: Computer Vision & Image Processing. Proceedings of the 12th IAPR International Conference on, 1:566–568 vol. 1. doi:10.1109/ICPR.1994.576361.

Evans, A.C., Collins, D.L., Mills, S.R., Brown, E.D., Kelly, R.L., and Peters, T.M., 1993. 3D statistical neuroanatomical models from 305 MRI volumes. In Nuclear Science Symposium & Medical Imaging Conference: 1993 IEEE Conference Record, ed. Lowell Klaisner, 1813–1817. IEEE, October. http://ieeexplore.ieee.org/xpl/abs_free.jsp%3FarNumber%3D373602.

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

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Thomas Yeo, B.T., Mohlberg, H., Amunts, K., Zilles, K., 2008. Cortical folding patterns and predicting cytoarchitecture. Cereb. Cortex (New York, N.Y.: 1991) 18 (8), 1973–1980. doi:10.1093/cercor/bhm225 (August).

Hinds, O.P., Rajendran, N., Polimeni, J.R., Augustinack, J.C., Wiggins, G., Wald, L.L., Diana Rosas, H., Potthast, A., Schwartz, E.L., Fischl, B., 2008. Accurate prediction of V1 location from cortical folds in a surface coordinate system. NeuroImage 39 (4), 1585–1599. doi:10.1016/j.neuroimage.2007.10.033 (February 15).

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G.E., et al., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46 (3), 786–802. doi:10.1016/j.neuroimage.2008.12.037 (July 1).

Menke, J., and Martinez, T.R., 2004. Using permutations instead of Student's *t* distribution for *p*-values in paired-difference algorithm comparisons. In Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, 2:1331–1335 vol.2. Vol. 2. doi:10.1109/IJCNN.2004.1380138.

Rademacher, J., Caviness, V.S., Steinmetz, H., Galaburda, A.M., 1993. Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology. Cereb. Cortex 3 (4), 313–329. doi:10.1093/cercor/3.4.313 (July 1).

Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W., 2008. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage 39 (3), 1064–1080 (February 1).

Thirion, J.-P., 1998. Image matching as a diffusion process: an analogy with Maxwell's demons. Med. Image Anal. 2 (3), 243–260. doi:10.1016/S1361-8415(98)80022-4 (September).

Thompson, P., Toga, A.W., 1996. A surface-based technique for warping 3-dimensional images of the brain. IEEE Trans. Med. Imag. 15 (4), 402–417.

Van Essen, D.C., Drury, H.A., Dickson, J., Harwell, J., Hanlon, D., Anderson, C.H., 2001. An integrated software suite for surface-based analyses of cerebral cortex. J. Am. Med. Inform. Assoc. : JAMIA 8 (5), 443–459.

Yeo, B.T.T., Sabuncu, M.R., Vercauteren, T., Ayache, N., Fischl, B., Golland, P., 2010. Spherical Demons: Fast Diffeomorphic Landmark-Free Surface Registration. IEEE Transactions Medical Imaging 29 (3), 650–668.