Accurate and robust brain image alignment using boundary-based registration

Douglas N. Greve\textsuperscript{a,⁎}, Bruce Fischl\textsuperscript{a,b}

\textsuperscript{a} Martinos Center for Biomedical Imaging, 143 13th Street, Charlestown, MA, USA
\textsuperscript{b} MIT Computer Science and AI Lab/Division of Health Sciences and Technology, USA

\textbf{ARTICLE INFO}

Article history:
Received 18 March 2009
Revised 26 May 2009
Accepted 17 June 2009
Available online 30 June 2009

\textbf{ABSTRACT}

The fine spatial scales of the structures in the human brain represent an enormous challenge to the successful integration of information from different images for both within- and between-subject analysis. While many algorithms to register image pairs from the same subject exist, visual inspection shows that their accuracy and robustness to be suspect, particularly when there are strong intensity gradients and/or only part of the brain is imaged. This paper introduces a new algorithm called Boundary-Based Registration, or BBR. The novelty of BBR is that it treats the two images very differently. The reference image must be of sufficient resolution and quality to extract surfaces that separate tissue types. The input image is then aligned to the reference by maximizing the intensity gradient across tissue boundaries. Several lower quality images can be aligned through their alignment with the reference. Visual inspection and fMRI results show that BBR is more accurate than correlation ratio or normalized mutual information and is considerably more robust to even strong intensity inhomogeneities. BBR also excels at aligning partial-brain images to whole-brain images, a domain in which existing registration algorithms frequently fail. Even in the limit of registering a single slice, we show the BBR results to be robust and accurate.

© 2009 Elsevier Inc. All rights reserved.

\textbf{Introduction}

The fine spatial scales of the structures in the human brain represent an enormous challenge to the successful integration of information from different images. Cortex has a typical thickness of between 1 and 5 mm (Fischl and Dale, 2000). Thalamic nuclei have spatial extents on the order of a few millimeters (Kandel et al., 2000). Cortex is also highly folded, meaning that points that are 20 mm from each other as measured along the cortical surface can be within a few millimeters in three-dimensional space. This places stringent requirements on the accuracy and precision of algorithms for aligning images. Even a few millimeters of alignment error can cause a voxel in a tissue type in one image to be assigned to the wrong tissue type in another image. This challenge extends to several domains, including within-subject integration of images from a wide range of imaging modalities such as structural MRI, functional MRI (fMRI), arterial spin labeling (ASL), diffusion weighted imaging (DWI), positron emission tomography (PET), within-subject single-mode longitudinal analysis, and surface-based group analysis (Fischl et al., 1999).

Automatic within- and cross-modal registration has a long history of research in neuroimaging. The most prominent and widely used algorithms currently extant are Cross-Correlation (Collins et al., 1995), Mutual Information (MI; Maes et al., 1997; Maes et al., 1999; Wells et al., 1996), Normalized Mutual Information (NMI), and Correlation Ratio (CR; Roche et al., 1998). Surface-to-surface and shape registration (Borgefors, 1988) have also been used for multimodal registration; however, West et al. (1999) found that these were not as accurate as intensity-based techniques. The basic model used in CR is that an intensity value in one mode will have one, and only one, matching intensity value in the other mode. MI and NMI are similar to CR but less restrictive in that they attempt to sharpen the intensity joint histogram. For both methods, intensity mismatches are evidence for misregistration. Correspondence can be achieved by adjusting the registration parameters (i.e., translations, rotations, scalings, shears) until the best match (i.e., minimum cost) is achieved. Unfortunately, intensity inconsistencies can exist for other reasons, which may be mode-specific. For example, in a Blood-Oxygen-Dependent (BOLD) weighted image, the tissue outside of the brain often appears quite dark whereas it will be bright on an anatomical. The brain can be extracted from the anatomical image (Segonne et al., 2004), but the quality of the resulting registration will then be sensitive to the aggressiveness and quality of the extraction. Echo planar images (EPI) are also subject to B0 distortion in the form of non-linear metric distortion and intensity “drop out” (Jezzard and Balaban, 1995). In addition, coil sensitivity profiles and B1 inhomogeneity can create spatial intensity fluctuations. Coil sensitivity fluctuations can become extreme when surface coils are used as is often done when studying retinotopy with fMRI (Sereno et al., 1995). All these effects create

\textsuperscript{⁎} Corresponding author.
E-mail addresses: greve@nmr.mgh.harvard.edu (D.N. Greve), fischl@nmr.mgh.harvard.edu (B. Fischl).

1053-8119/5 – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.06.060
inconsistencies in input intensity matching that are unrelated to the
quality of alignment, but will be indistinguishable from alignment
erors and thus can drive an alignment away from the true
optimum.

In longitudinal analysis, in which the same subject is imaged
over time in different scan sessions, cross-session alignment of the
brain is well described by a rigid transformation, but there are
significant non-rigid effects such as differences in jaw and tongue
placement or head-neck angle. In theory, these can be accommo-
dated by brain extraction, but, again, the resulting registration will
be sensitive to differences in the extraction of the brain at different
time points. Finally, we point out that some imaging modalities may
not have full-brain coverage. For example, the brain coverage in EPI
must be reduced in order to reduce the slice thickness and maintain
temporal resolution. Partial field-of-view (FoV) brain coverage
creates enormous problems when attempting to register to a
whole head.

Recently, Saad et al. (2009) performed an extensive analysis
comparing CR and MI and found that they had errors that could easily
exceed 3 mm. This led them to propose a new cost function optimized
for T2⁎–T1 registration using a local Pearson correlation (LPC). LPC is a
local method, meaning that the cost at each voxel is only dependent
on the nearby voxels (roughly the 36 nearest neighbors). As part of
the computation of the Pearson correlation, the mean intensity over a
neighborhood is subtracted from the intensity at each voxel in the
neighborhood. This is effectively a spatial high-pass filter and so
enhances edges in both the input and reference images; the
correlation is then computed over the neighborhood. Areas away
from tissue boundaries should contribute little to the overall cost since
these areas tend to have locally homogeneous intensities in both
images which will be suppressed by the local mean removal. This
means that the areas near the tissue boundaries will drive the cost
function; in this way it is similar to BBR.

Like Saad, et al., we have also observed inaccuracies and
sensitivities in the registrations found by CR and MI. In this paper,
we propose a new algorithm based on the principle that the most
salient registration cue is the contrast across a tissue boundary, and so
we refer to it as Boundary-Based Registration, or BBR. Unlike the
methods reviewed above, BBR does not treat the two images as equal.
One of the images (the “reference image”) must be a high-quality
anatomical volume sufficient for extracting surfaces that separate
brain structures and tissue types. The second image (the “input
image”) can be any modality as long as it has tissue contrast.
Alignment is achieved by maximizing the gradient of the input image
intensity across the surface boundary. Intensity values in the
anatomical image are not part of the cost function. If two or more
input images need to be aligned, they can each be separately aligned to
the reference with BBR. We emphasize that this is not a surface-to-
surface registration like those described above. A surface is only
extracted from one image, the high-quality anatomical; intensities are
used from the second image. It is similar to LPC in that a local cost is
computed from tissue boundaries. However, there are some differ-
ences in that BBR operates over a much smaller neighborhood (a few
millimeters) and uses percent contrast instead of correlation
coefficient. At the time this manuscript was being prepared, LPC had
only recently been published, and so we have not had the chance to
compare it to BBR directly.

We show that BBR yields superior accuracy compared to CR and
NMI using blinded human raters as well as improved FMRI results. We
then show that BBR is extremely robust to variations in its parameters
and initialization, major fluctuations in image intensity inhomogene-
ity, and to partial-brain images, even to the extent of accurately
registering single slices, something for which most current registra-
tion methods fail. The software that implements BBR is publicly
distributed as part of the FreeSurfer (surfer.nmr.mgh.harvard.edu)
software package.

Methods

Cost function derivation

Our fundamental hypothesis in this research is that the quality of
registration is best assessed by the magnitude and direction of the
change of intensity across a tissue boundary. This is essentially what is
done by a human when registering visually (see Fig. 2), which
motivates us to quantify goodness of alignment using boundary
intensity gradients. In this work, the boundary we use is derived from
an anatomical volume using a surface model. An anatomical surface
model is one where a mesh of points, triangles and edges is used to
define the boundary between different tissue types or structures. The
mesh consists of a set of vertices and the neighborhood relations of
each vertex. Each vertex has a coordinate in the 3D anatomical space
that allows us to compute various geometric features of the boundary
such as curvature and surface normal vectors. Though this method can
be applied using the surface between any structures, we focus on the
surface between cortex and adjoining white matter. We believe that
cortex is the best target for two reasons: (1) it is highly folded,
allowing dense sampling of 3D space, and, (2) it is very thin (2.5 mm
on average; Fischl and Dale, 2000), implying that small alignment
errors will generate large changes in the cost function. Despite the fact
that we use the cortex as a target, we emphasize that the goal is to
achieve good registration between the entire input volume and the
entire anatomical volume, not just the locations along the cortex.
However, for affine transformations, good cortical registration is
sufficient to achieve good registration everywhere (assuming no non-
linear distortion).

Several software programs exist to construct computational
models of the cortical surface such as FreeSurfer (Dale et al., 1999;
surfer.nmr.mgh.harvard.edu), Caret (Van Essen et al., 2001), Brain-
VISA (Rivière et al., 2003) and Brain Voyager (www.brainvoyager.
com). In FreeSurfer, the cortical surface models include a representa-
 tion of the surface between cortex and white matter (hereafter
referred to as the “white surface”) as well as the outer surface of the
brain (hereafter referred to as the “pial surface”); this allows an
estimate of cortical thickness at each vertex. FreeSurfer also provides
detailed labeling of both cortical and subcortical structures. This can
be useful for excluding certain areas known to have problems in
certain modes (e.g., orbital frontal cortex in EPI).

Let A be the 3D anatomical volume indexed by r (i.e., the intensity
of A at point r is A(r)). Let E be the 3D input volume of the same subject
indexed by p. Then some transform T exists such that p = T(r,θ)
will bring the two volumes into registration where θ is the set of registra-
tion parameters. The input intensity value at anatomical place-
r r for parameter set θ can be computed as E(T(r,θ)). We
model T as an affine (12 degree-of-freedom (DOF) transform (i.e., 3
translations and 3 rotations, 3 scales, and 3 shears). The coordinate
system is centered at the center of the anatomical volume, which
puts it near the center of the brain. Transformations are applied in
the order of translation, rotation, scaling, and shear, with translations
measured in mm and rotations measured in degrees. By default, we
use a rigid transform (6 DOFs) because we assume that both volumes
are from the same subject; however, our software can handle as
many as 12 DOFs. The problem, then, is to find T given all the issues
discussed above.

A vertex V on the white surface will have a 3D location rV and
surface normal nV (see Fig. 1A). We can compute a location just inside
the surface (i.e., in the white matter) as T
w = rV − Δw nV, and just
outside the surface (i.e., in the cortical gray matter) as T
c = rV + Δc nV,
where Δw and Δc are the white and gray matter projection
distances, respectively, and may be vertex-dependent. For a given θ,
these points can then be transformed to the input volume through
T, i.e., p
w = T(r
w) and p
c = T(r
c) in order to sample the
intensity at these points given the set of registration parameters,
One of the advantages to this method is that $D_V$ is mostly insensitive to spatial intensity inhomogeneities due to coil profiles or B1 inhomogeneities. These fluctuations are slowly varying over space, whereas $D_V$ is computed from points that are only a few millimeters apart. We also use a percent contrast measure at each vertex:

$$Q_v(\theta) = \frac{100}{0.5(g_v(\theta) + w_v(\theta))} \frac{g_v(\theta) - w_v(\theta)}{g_v(\theta)}.$$

Using a percent contrast assures that bright regions do not get a larger weight than dark regions. We pass the percent contrast through a non-linearity (see Fig. 1B) and sum over the surface to compute the final cost as given in the equation below:

$$J(\theta) = \frac{1}{N} \sum_{v \in B} h_v(1 + \tanh(M_v \cdot (Q_v - Q_0))).$$

where $Q_0$ is an offset parameter, $M_v$ is a slope parameter the sign of which is dependent upon the expected direction of contrast in the input image (positive for gray matter brighter than white), and $h_v$ is a weight for vertex $V$. Allowing the slope parameter to change with vertex can allow the modeling of contrast changes with anatomy. The $\tanh(\cdot)$ function is motivated by the desire to have a smooth function with saturating non-linearity that reduces the weight for vertices that have a large contrast in the unexpected direction. In this way it is similar to the saturating cost functions used in robust statistics (Hampel et al., 1986). It also assures that vertices with a large direction of contrast in the expected direction have a low cost. We could have used a binarization (i.e., setting the cost to 0 if the direction of contrast is in the expected direction or 1 otherwise), but this would have made the cost function less smooth. We expect that most any function that meets these criteria would work, and we show that the final results are quite insensitive to parametric variations in the cost function. The total cost is summed over a (possible) subset $B$ of vertices that fall within the 3D FoV of $E$. A subset allows for speed increases during an initialization phase and/or avoidance of problem regions. $N$ is the total number of vertices participating; this might be less than the number of vertices in $B$ if some vertices fell out of the 3D FoV of $E$.

By default, we use the following parameter settings: $Q_0 = 0$; $M_v = +0.5$ for modalities where we expect gray matter to be brighter than white matter (e.g., a T2-weighted MRI) and $-0.5$ for the reversed contrast; $h_v = 1$ for all vertices; Subset $B$: only vertices in cortex; the second and third optimization stages (below) sample every 100th vertex; the fourth and fifth sample every vertex. $\Delta_g$ and $\Delta_w$ are the distances from the gray/white junction that we sample along the surface normal ($n_v$) into the white matter. The percent contrast at a vertex ($Q_v$, Eq. (1)) is then computed as the relative difference between the gray and white matter intensities. (B) Function that converts contrast into a cost, with large contrasts in the expected direction (positive for BOLD-weighted) having a low cost and those in the unexpected direction (negative) having a large but saturating cost.

---

**Handling B0 distortion**

B0 distortion is the geometric and intensity distortion caused by inhomogeneities in the main magnetic (B0) field caused predominantly by air–tissue interfaces in the scanner (Jezzard and Balaban, 1995). Acquisitions with long k-space readouts (such EPI) are most susceptible, and the B0 distortion tends to increase with an increase in resolution as this will usually increase the readout time. The metric distortion is non-linear and can be corrected through the use of a B0...
map (Jezzard and Balaban, 1995). The intensity distortion (also called “dropout”) cannot be corrected. The metric and intensity distortions tend to occur in known anatomical locations (e.g., orbital frontal, medial temporal gyrus, etc.), and these areas are automatically labeled in FreeSurfer (Desikan et al., 2006). This gives us the ability to exclude them from the cost function. Note that these regions will still be aligned properly if the B0 distortion has been corrected, though the signal may be less useful due to dropout. If the B0 distortion has not been corrected, then these areas will be misregistered. We have built into our BBR implementation the ability to mask out the following regions: middle temporal gyrus, inferior temporal gyrus, temporal pole, fusiform gyrus, entorhinal, medial orbital frontal gyrus, caudal anterior cingulate gyrus, and rostral anterior cingulate gyrus. Below, we explore the effect that this has.

**Optimization**

Optimization is the process of finding the global cost minimum over the parameter space. No method except a costly, and possibly intractable, brute-force search can guarantee that the global optimum has been found. Registration methods typically rely on a course search over a wide range of parameter space followed by gradient descent to avoid this problem. A local minimum may also be encountered because discontinuities exist near the global optimum (Jenkinson et al., 2002). This type of local minimum is avoided because, as we show, the BBR cost function is very smooth. The BBR optimization stages are:

1. **Initialization.** In this step we compute an initial registration that falls into the capture region, roughly within 7 mm and 7° of the optimum (Fig. 3). This can be accomplished in several ways:
   a. The anatomical and input volumes were acquired in the same session, in which case the initial registration can be performed in scanner coordinates based purely on the geometry as long as the subject has not moved excessively between acquisitions.
   b. The input volume is whole-brain, in which case we use existing software that implements a global method (e.g., FSL/FLIRT or SPM/spm_coreg) to obtain an initial registration that will fall into the BBR capture region.
   c. The input volume is partial-brain or surface coil but a whole-brain volume was also acquired, in which case methods (a) and (b) can be combined. Note that any whole-brain volume is sufficient; it does not need to be an anatomical. If a surface coil is used, then the whole-brain should be acquired with the body coil.
   d. If the above methods fail, the initial registration can be determined manually.

2. **Coarse Search.** Here we sample the cost function at 3 points along each parameter dimension to form a 6 dimensional grid in parameter space. The translations are sampled at $-4, 0,$ and $+4$ mm and the rotations are sampled at $-4, 0,$ and $+4°$ for a total of $3^6 = 729$ points; the point with the smallest cost is fed to the next step. This is done in order to extend the effective capture range of the algorithm, allowing datasets with large translations and rotations to be accurately registered. The surface is sampled every 100th vertex to speed computations (our typical surfaces have on the order of 150,000 vertices, so this still results in a dramatically over-determined problem with 1500 data points and 6 unknowns). For problem data sets, the FoV and sampling resolution of this grid can be expanded at the cost of more computation time.

3. **Gradient Descent I.** The tolerance (see below) is set to $10^{-4}$ and the surface is sampled every 100th vertex to speed computations. The result is fed to the next stage.

4. **Fine Search.** Again, we sample the cost function at 3 points along each parameter dimension to form a 6 dimensional grid in parameter space. The translations are sampled at $-0.1, 0,$ and $+0.1$ mm and the rotations are sampled at $-0.1, 0,$ and $+0.1°$ for a total of $3^6 = 729$ points; the point with the smallest cost is fed to the next step. This is done for cases where the topography around the minimum is very flat. This step assures that the bowl will be approached from the best nearby position. The surface is sampled every vertex.

5. **Gradient Descent II.** The tolerance (see below) is set to $10^{-8}$ and the surface is sampled every vertex. If the registration is being done with more than 6 DOF, then the extra DOFs are optimized here. The previous steps use only 6 DOF.

The gradient descent is performed using a series of 1D minimizations (i.e., Powell’s method (Press et al., 1988) as implemented in the VXL library (vxl.sourceforge.net)). The gradient descent has one parameter, the tolerance used to terminate the descent. The tolerance is the absolute difference in costs between successive steps divided by the average cost at those steps. The minimal tolerance achievable is constrained by the machine precision. In our tests, we found that setting the tolerance below $10^{-8}$ had no effect on a 64 bit machine.

**Evaluation criteria**

In order to evaluate the accuracy of the BBR, we compare it to two other popular registration algorithms: CR and NMI. Except where noted, all algorithms are constrained to 6 DOF. Many neuroimaging software packages have implementations of CR and NMI, e.g., FSL (www.fmrib.ox.ac.uk/fsl); SPM (www.fil.ion.ucl.ac.uk/spm), AFNI (afni.nimh.nih.gov/afni), AIR (bishopw.loni.ucla.edu/AIR), and MINC (www.bic.mni.mcgill.ca/software). For our evaluations, we will use the FSL implementation of CR (the FLIRT program), and the SPM implementation NMI (the spm_coreg program). These will be evaluated based on accuracy and robustness. Accuracy is judged based on visual inspection in which the raters are blinded to the registration algorithm identity, and ability to map fMRI activation to the surface. Robustness is assessed based on the ability of an algorithm to maintain a registration when the input image or the algorithm itself is changed in some way. The accuracy and robustness tests are performed on the same data set. Using the accuracy tests, we show that the accuracy of BBR equals or exceeds that of CR and NMI on whole-brain, artifact-free data. Using the robustness tests, we show that the BBR registration deviates little from the whole-brain result even under substantial changes to the data or to the implementation parameters. This implies that BBR remains accurate under these conditions. One way to measure accuracy is to know the ground truth under simulated conditions. While, it is not a simple matter to simulate metric and intensity distortion in MRI, some progress has been made in this area (Drobnjak et al., 2006; Xu et al., 2007). However, we have chosen to use a combination of visual inspection and consistency tests on real images.

**Visual inspection**

We use the FreeSurfer interactive tool called “tkregister2” for visually inspecting (and possibly changing) a registration. The tool allows the anatomical and functional images to be displayed. The cortical surface can also be displayed on both images. Screen shots are shown in Fig. 2. Panel A shows the T1-weighted anatomical with the white surface shown in red; notice how the white surface closely follows the gray/white boundary. Panel B shows a BOLD-weighted EPI image. While the cortical anatomy is blurry, it can clearly be seen.
**JfMRI activation**

Activation in an fMRI experiment should appear in gray matter and so the amount of activation measured on the cortical surface can be used as a measure of the quality of the registration between the anatomical and the fMRI volumes. For a given registration, we sample the fMRI activation onto the surface and simply count the number of voxels above a threshold \( p < .001 \); with the idea that the more voxels, the better the registration. Of course, some of those voxels will be false positives; however, the fMRI analysis we used (see below) includes temporal whitening so that the actual false positive rate should be close to the nominal rate of 0.1%. More importantly, we apply the registration after the analysis, so each registration will see the same false positive rate, and so false positives should not bias our ranking of methods. For more details, see fMRI analysis methods below.

**Robustness**

As mentioned above “robustness” is the ability of an algorithm to maintain a registration when the input image or the algorithm itself is changed in some way. The basic idea here is that there is only one optimum registration, and that the optimum should be invariant to initialization, image corruptions and a wide range of algorithm parameters. Note this makes no statement about whether the optimum is a good registration. To quantify the robustness, we need a measure of the difference between two registrations. For this we have chosen to use a measure we call the Average Absolute Distance (AAD); this is the distance that the cortical surface moves between the two sets of registrations parameters \( \theta_1 \) and \( \theta_2 \). The AAD is defined to be:

\[
AAD(\theta_1, \theta_2) = \frac{\sum_{v} |T(r_v, \theta_1) - T(r_v, \theta_2)|}{N_v}.
\]

The AAD is measured in mm and gives us an easily interpretable metric for registration consistency. Note that this is not biased by the fact that we generate the cost function from the surface. The surface is simply a convenient way to sparsely sample the registration differences over the entire brain. This measure is similar to the measure introduced in Jenkinson et al. (2002) in which a closed-form expression was used to compute the average RMS deviation between two registrations averaged over a sphere. However, AAD is computed over the actual cortical surface instead of a sphere and uses the absolute value instead of the square.

We propose several robustness tests:

1. **Cost Function Parameter Manipulation** (BBR only). The purpose of this test is to determine how sensitive the registration is to changes in the parameters of the cost function and the way algorithm is implemented. The parameters we perturbed are \( Q_0, M_v, \Delta_v, \) and \( \Delta_w \). We generate the first set of registration parameters \( \theta_1 \) using the default parameters. We then change the cost function parameters and re-register the original volume, using the first registration as initialization, to generate the second set of parameters \( \theta_2 \). We then compute the AAD between the two sets of registrations. Ideally, the AAD would be 0.

2. **BO Masking** (BBR only): in this test, the first registration is generated using all cortical vertices on the surface. This registration seeds a second registration in which the vertices in the BO regions described above are masked out (i.e., do no contribute to the cost function). The AAD is then computed between the two registrations.

3. **Initialization Test** (BBR only): in this test, we register a data set, then modify the resulting registration by translating and rotating it, and then use this modified registration as the initial starting point for Stages 2–5 to see how closely we return to the original registration. This simulates the case were Stage 1 is inaccurate. For each subject,
we randomly selected 30 translations and rotations uniformly in the range of \(+/−10\) mm and \(+/−10^\circ\) (i.e., all 6 registration parameters were simultaneously changed). The size of the “perturbation” was defined as the maximum of these translations and rotations (measured in mm or degrees). The re-registration was deemed to have failed if it did not return to within 0.100 mm AAD of its original location.

4. **Reduced Field-of-View (FoV) Test:** the first registration is generated from a whole-brain fMRI volume. The second registration is generated from the same volume after removing some number of slices from the top and bottom and using the first registration as initialization. These should generate exactly the same set of registration parameters. Thus we can measure the AAD as a function of the number of remaining slices. This simulates the case where data of interest is acquired with a reduced FoV, and a full FoV data set is acquired in the same scanning session to assist in registration.

5. **Intensity Inhomogeneity Test:** the first registration is generated from the original fMRI volume. The bottom half of the slices are then discarded while the top half have their intensity scaled on a slice-by-slice basis. This situation simulates the case where a surface coil is used (which induces a sharp fall-off in intensity away from the coil). The initial whole-brain volume largely free from intensity inhomogeneity can be obtained using the body coil for reception (this is noisy but adequate). The coil profile is implemented as a half-cosine function: \(F = \frac{1}{2} (\cos(\alpha) + 2 - \alpha)\) where \(S\) is the slice number, \(N\) is the number of slices, and \(\alpha\) is a control parameter (see Fig. 6). For slices nearest the top of the brain (i.e., low values of \(S\)), \(F\) is nearly 1 (no intensity distortion). For slices further away from the top, \(F\) drops off and reaches 0 when \(S = N\). When \(\alpha\) is 0, \(F = 1\) for all values of \(S\) (i.e., no distortion), as \(\alpha\) increases to 1, the amount of drop off increases. Thus, we can measure the AAD as a function of \(\alpha\).

6. **Surface Inaccuracy Test:** in this test we evaluate how the BBR results change when a less accurate surface is used. For this we used the FMRIB’s Automated Segmentation Tool (FAST; Zhang et al., 2001) distributed with FSL. We used this program to segment the anatomical images into gray, white, and CSF, and then used FreeSurfer to generate a surface between white matter and the other tissue types. We then ran BBR with this surface, initializing with the registration from when the FreeSurfer surface was used. The AAD was then computed between these two registrations. Since this application only has one surface, we used an absolute gray matter projection distance of 2 mm. Note that we are not trying to do a quantitative comparison between FreeSurfer and FAST; we simply want to generate reasonable surfaces that we do not expect to be as accurate as those generated by FreeSurfer.

7. **Degree-of-Freedom Test:** in this test we evaluate how the BBR registration changes when higher DOFs are used. We use the 6 DOF registration to seed the 9 (+scale) and 12 (+scale+shear) DOF optimizations, then measure the AAD between the pairs for all 18 subjects.

**Evaluation data set, MRI methods, and fMRI analysis**

The data set used to evaluate the registration method was collected as part of the Functional Biomedical Informatics Network (fBIRN, www.nbirn.net) East Coast Traveling Subjects (ECTS) pilot study. Eighteen subjects were each scanned at four sites undergoing both functional and anatomical protocols. This evaluation only deals with the data collected at MGH (Siemens 3 T Tim-Trio with 12-channel head coil). A T1-weighted MP-RAGE volume was collected for each subject and analyzed in FreeSurfer to create the white and pial cortical surfaces and compute thickness and surface normal vectors at each vertex. Eight whole-brain fMRI data sets (TR=2000 ms, TE=30 ms, flip=77°, bandwidth=2298 Hz/pixel, 30 slices, slice thickness 4 mm skip 1 mm, matrix size 64×64 at 220×220 mm, 142 time points, sequential slice order) were acquired for each subject. Subjects performed a working memory task that consisted of four block types (1) Scrambled images, (2) Encode, (3) Distractor, (4) Probe. During Encode, subjects passively viewed line drawings which they were asked to remember. During the Distractor block, subjects were shown complex pictures and asked to respond with a button press as to whether there was or was not a human face. During the Probe phase, subjects were shown pairs of line drawings, one of which they had seen during the Encode phase and asked to respond with a button press as to which drawing they had seen before. Each block lasted 16 s. The middle time point of each run was used as the fMRI template for both registration and motion correction. The fMRI data were motion corrected using MCFLIRT and slice-time corrected using the FSL slicetimer program. The fMRI data were not spatially smoothed or otherwise interpolated as this would have made the results less sensitive to registration errors. The fMRI data were not corrected for B0 distortion. The fMRI time series analysis was performed using FEAT (Woolrich et al., 2001), modeling each block as a separate explanatory variable using the default gamma hemodynamic response model. The contrast between Probe and Scrambled blocks was thresholded at a voxel-wise threshold of \(p<.001\) (uncorrected). This contrast yields robust activation across wide areas of the brain. Note that at this point, the thresholded map is still in the native functional space. The fMRI template was then registered to the anatomical using CR (FLIRT), NMI (spm_coreg), and BBR (with CR registration as initialization), all constrained to 6 DOF. For each registration method, the binarized activation map was sampled onto the cortical surface midway between the white and pial surfaces using nearest-neighbor interpolation (no attempt was made to exclude regions of high B0 distortion). The activation percentage was computed as the total number of positives divided by the number of vertices. This yielded 8 values for each subject. These were averaged to give 18 values for each method, and a paired permutation test was used to assess the significance of the differences between BBR, CR, and NMI.

**Results**

**Cost function smoothness, optimality, and capture region**

The cost function for 1D translations and rotations are shown in Fig. 3 for a single subject; the 0-point is the location of the optimal solution (visually, the alignment looked very accurate). The cost function was evaluated by translating (Fig. 3A) in each axis by \(+/- 100\) mm in 0.010 mm steps; this 200 mm FoV was enough to encompass the entire brain in all dimensions. It was also evaluated by rotating about each axis by \(+/- 100^\circ\) in steps of 0.01\(^\circ\) (Fig. 3B). The three panels in Figs. 3A and B show the behavior of the cost function at different scales around the optimum. There are two things to note about these results. First, the optimum is sharp and global — while there are many local minima, none come close in cost at the true optimum; this is expected given the thin and highly folded nature of cortex. The actual cost at the optimum will depend on the average contrast between gray and white matter and so dependent on the modality and pulse sequence preparation. As this contrast drops, the cost at the optimum will rise. However, the cost far away from the optimum will always be around 1.0 regardless of the nature of the image. This is because, away from the optimum, the surface cuts randomly through the volume, and so the average contrast should be about 0, making the cost about 1.0 (see Fig. 1, panel B). While this suggests we can detect inaccurate registrations as those with costs “close” to 1.0, we cannot select a single cost threshold that will be good for all possible input images because of the dependence of the optimal cost on the actual contrast. However, once we do know what the expected contrast is, we can determine such a threshold. Note that costs near
to or far from the optimum will not change much with intensity shading or partial-brain FoV because we used a locally computed percent contrast. We explore this further below. The second thing to note about Fig. 3 is that the cost function is extremely smooth, even at the scale of tens of micrometers and 100ths of a degree. The implication of this is that once the gradient descent starts inside the capture region, it is very likely to find the optimum. The capture range is quite sharp, about $+/-5$ mm and $+/-5°$, though we expect this to be expanded by about 4 mm and 4° due to the course search (Stage 2).

**Visual inspection test**

Fig. 2 shows typical results for the three registration techniques. As discussed above, the accuracy is judged by how well the red line (white surface) approximates the boundary of the bright patches. Inaccuracies in registration are indicated by bright patches outside of the red line or dark patches inside the red line. Both CR (panel 2C) and NMI (panel 2D) have some problem areas indicated by the green arrows (these are in the same place relative to the red line in all three panels). There are extremely few locations in the BBR (panel 2E) registration that are problematic. For reference, the AAD between BBR and CR for this data set was 3.74 mm; the AAD was 2.86 mm between BBR and NMI. Authors DNG and BRF rated the three registrations for each of the 18 subjects (blinded to registration algorithm, as noted above). The best registration was found to be “Acceptable” for all subjects. For both raters (DNG and BRF), BBR was rated best for all subjects; NMI received 17 second place rankings; CR received 1 second place. The only difference between the raters was the one CR case receiving a second place rating was different. The difference between BBR and CR was $4.895$ mm AAD; the difference between BBR and NMI was $2.686$ mm AAD. These are similar to the errors for CR and MI found by the raters in Saad et al. (2009).

**fMRI activation comparison**

All methods achieved activation in excess of 13% of the cortical surface, averaged over all 18 subjects and 8 runs (BBR: 15.5%, NMI: 15.2%, CR: 13.9%). As discussed above, false positives will add about 0.1% and should not systematically bias any method since each method sees the same false positive rate. BBR increased the activation percentage over CR by about 12% ($p < 10^{-6}$). Each subject individually showed improvement, making the difference systematic and highly significant. BBR increased the activation percentage over NMI by only about 2% ($p < .01$). While modest, this difference was also systematic with 15/18 subject showing improvement. The significances quoted above were determined using a permutation test (Nichols and Holmes, 2001) with 1,000,000 iterations on the subject-wise difference between the activation percentages.

**BBR sensitivity to starting point**

The starting point perturbation results are shown in Fig. 4A. For perturbations below 5 (mm or degrees), all registrations were successful (i.e., AAD $< 0.100$ mm). At a perturbation of 7, fewer than 5% failed (i.e., were caught in a local minimum). Fig. 4B shows the mean cost of the successes and of the failures. One can see that there is a clear distinction between the two, meaning that failures can automatically be detected by evaluating the cost functional. This property is also evident from Fig. 3, which shows that the costs at local minima are much higher than that at the optimum. The actual threshold needed to detect failures will be dependent on the contrast...
properties of the images being registered, so it will need to be computed from a representative set of data. This starting-point test is important because BBR must start within its basin of attraction of the global optimum to be successful. The initialization may be inaccurate due to inaccuracies in the Stage 1 technique (i.e., CR or NMI). For whole-brain data, we do not expect this to be likely from our experience, and (Saad et al., 2009) found that most of the CR and MI errors were no worse than 5 mm.

If a partial FoV registration is being performed using a same-session whole-brain image as an intermediate step, then the subject may have moved between the acquisition of the whole- and partial-brain images. To get an idea of how big this motion might be in a real-world setting, we measured the maximum perturbation between the first and last fMRI runs for each of the 18 subjects (shown as asterisks in Fig. 4A). Seventeen of the subjects were less than 3; the worst was 5, and none were failures. Based on these findings, we expect there to be very few cases where Stage 3 does not start in the capture region. When failures do occur, we expect to be able to automatically detect them.

**BBR sensitivity to parametric variation and B0 masking**

There are four parameters in the BBR cost function: (1) offset ($Q_0$), (2) Slope ($M$), (3) Cortical Projection Distance ($\Delta_c$), and (4) White Matter Projection Distance ($\Delta_w$). The sensitivity to these parameters was tested by varying them individually over a range, re-registering, and seeing how much the registration changed (as measured with AAD). When the offset was varied from $-2$ to $5$ (nominal 0), the worst AAD across all subjects was 0.400 mm. When the slope was varied from 0.1 to 1 (nominal 0.5), the AAD for all subjects was less than 0.225 mm. When the WM Projection Distance was varied from 0.1 to 4 mm (nominal 2), the worst AAD was 0.425 mm. When the Cortical Projection Distance was varied from 0.1 to 1 (nominal 0.5), the AAD was no worse than 0.300 mm. This shows that the BBR solution is exceedingly robust with respect to wide changes in the parameters of the cost function, with the worst results across all subjects and parameter ranges being only 0.425 mm. Including masking of B0 regions only changed the registration by an average of 0.200 mm AAD with a maximum of 0.260 mm AAD. In this case, masking out susceptibility regions had little effect, but this might not be the case in data sets where the B0 distortion is more extreme. To get an idea of the amount of B0 distortion in this data set, we used the FSL PRELUDE and FUGUE programs to compute the amount of voxel shift in various regions. In the medial orbital frontal regions, the shift was about 5 mm (1.5 voxels); in inferior and medial temporal regions, the shift was about 3 mm (.9 voxels).

**Partial FoV consistency test**

Fig. 5 shows how the methods perform when the number of slices is systematically reduced. At full-brain (30 slices) and near full-brain coverage, the methods are similar in their ability to maintain consistency. However, as brain coverage drops, the performance of CR and NMI degrades considerably. At approximately half-brain coverage, the AAD is equal to the average human cortical thickness (2.5 mm), meaning that cortical voxels in the functional images would likely not be mapped to cortex in the anatomical (or would be mapped to the wrong part of cortex). Below 4 slices, CR and NMI fail catastrophically. In contrast, BBR maintains AAD of better than 1 mm down to 2 slices. Even using a single functional slice, the AAD is just above 1 mm. In fact, 17 of the 18 subjects had better than 1 mm AAD for a single slice. This is an important result because it is often the case that researchers need to acquire partial-brain FoV in high-resolution fMRI applications (e.g., Duong et al., 2002; Hyde et al., 2001; Kirwan et al., 2007; Miller et al., 2006).

**Intensity bias consistency test**

Fig. 6 shows how the methods perform as increasing levels of intensity bias are applied. For all methods, the performance drops with more bias, but BBR clearly outperforms both CR and NMI. At $\alpha = 0.5$, CR and NMI have an AAD equal to the average cortical thickness, and beyond that level they begin to fail catastrophically. BBR maintains better than 1 mm consistency all the way out to $\alpha = 0.9$.  

---

**Fig. 5.** The effect of reducing the number of slices on the ability of each registration algorithm to stay at the same position found from the full-brain registration. BBR maintains very good repeatability, even down to one slice. The large error bar for BBR on the single slice is because 1 of the 18 subjects failed. The confidence intervals represent 1 standard error.
Inaccurate surface test

Surfaces generated from the FAST segmentation had some regions of inaccuracy, particularly near the top of the head, the temporal lobes, and any place where a gyrus was very thin; however, the surface was fairly accurate through most of the brain. As mentioned above, this should not be construed as an evaluation or judgment of the FAST program; we simply used FAST to generate surfaces that are less accurate than FreeSurfer. The results show that use of the FAST surfaces had very little effect on the registration with an average AAD of 0.575 mm and the worst case of 1.716 mm.

DOF test

The average difference between 6 DOF and 9 DOF registrations was 1.237 mm AAD. The scale changed by about 2 or 3%, which does not make a big difference at the center of the brain but can be as much as a few millimeters near the edge. It was very difficult to tell from visual inspection whether the scale was helping or hurting the final registration. The average difference between 6 DOF and 12 DOF was 1.315 mm AAD; and the average difference between 9 DOF and 12 DOF was 0.281 mm AAD. From this we conclude that 12 DOFs is probably not necessary, at least for this data set. It is not our purpose here to resolve the question as to what the appropriate number of DOFs is for registering within-subject, cross-modal images, rather we only want to measure the sensitivity of BBR to changes in DOF.

Computational load

Of the 18 subjects in the evaluation data set, none took more than 4 min to complete all stages of the 6 DOF registration on a 64-bit 2 GHz XEON processor. This included the initialization stage using CR or NMI, which took about 30 s by itself. The bulk of the time was spent in Stage 5. Thus BBR only added a few minutes to the total processing time. For 9 DOF, the time increased to only 6 min; for 12 DOF, the time was about 15 min. Note that the computational load only depends on the number of vertices on the surface; it is independent of the size or resolution of the input image. This does not take into account the amount of time needed to generate the surfaces, which can be as much as 30 h if FreeSurfer is used. If one needs to have very accurate registration, then it is likely that one will need a very accurate model of the anatomy as well, and so generating the surfaces might not require any additional computational overhead. We also expect that BBR will perform well on surfaces of much lower quality. For example, the tests above showed very little deviation when less accurate FAST-based surfaces were used, and it only took FAST a few minutes to run.

Other modalities

We have used BBR on several other modalities, including DWI, ex vivo T1, ASL, and B0 maps, which we report on qualitatively here. DWI: we used the low-b volume as the input (the same that is used for motion correction). The low-b images of DWI scans have the same basic contrast and distortions as BOLD scans, and tend to be whole-brain and of higher resolution (ours had 2 mm isotropic voxel size). Visual inspection showed a similar pattern to the BOLD, i.e., CR and NMI performing well but BBR being a little more accurate. Ex vivo T1-weighted: our test images were very high-resolution (80 μm) acquisitions of single hemispheres suspended in fluid and were acquired using a special coil on a 7 T system. This is a real challenge for CR and NMI, and they nearly always failed catastrophically. However, BBR performs very robustly and accurately. ASL: we have run BBR on both the tag and control images as well as on the difference images. Visually, the registrations look to have about the same accuracy as CR and NMI. Magnitude images from B0 maps: the amount of gray–white contrast in these images depends on the echo times at which they were acquired. The ones we evaluated were acquired at 3 ms, which produced very little gray–white contrast. BBR was still able to register them, but the performance as about the same as CR and NMI.

Conclusions

Accurate and robust alignment of brain images is critical for multimodal integration, surface-based intersubject analysis, longitudinal analysis, and pre-surgical planning. Due to the size of the brain structures involved, registration accuracy frequently needs to be better than 1 mm. Our new alignment procedure, called Boundary-Based Registration, or BBR, adjusts alignment by maximizing image contrast across tissue boundaries rather than matching intensities between two images or by matching surface shapes. One of the novelties of BBR is that it treats the two images asymmetrically: one needs a very accurate model of the anatomy as well, and so generating the surfaces might not require any additional computational overhead. We also expect that BBR will perform well on surfaces of much lower quality. For example, the tests above showed very little deviation when less accurate FAST-based surfaces were used, and it only took FAST a few minutes to run.
On whole-brain, artifact-free BOLD data, CR and NMI perform admirably, but BBR still systematically outperforms both. When the entire brain is not imaged or the images have large intensity biases, CR and NMI can fail catastrophically whereas BBR is consistent with the whole-brain, artifact-free result to within 1 mm. BBR can even accurately align a single functional slice to the anatomical, which may facilitate some novel analysis techniques. The BBR cost function capture range is narrow (+/−5 mm and +/−5°), which necessitates having a method to initialize registration in the correct basin of attraction. For this, we use existing implementations of CR or NMI, thus we see BBR as an augmentation of those procedures and not a full replacement. In the case where a partial-brain FoV is needed, we recommend that the researcher also collect a fast whole-brain image in the same session, then use a CR or NMI whole-brain registration to seed the partial-brain BBR registration. BBR is also extremely robust to wide variations in its own parameters, and takes less than 4 min to run.

The key ingredient to BBR's performance is that it is based on a strong anatomical model derived from a high-quality structural image. This removes some of the burden from the (often) low-resolution cross-modal image. For example, it does not need to be skull-stripped and is insensitive to non-rigid effects such as jaw position, tongue position, head-neck orientation and differential image distortions. Using FreeSurfer’s automated cortical anatomical labeling, we can also exclude regions known to have intensity or distortion problems in the cross-modal images (e.g., BO distortion regions in EPI). We have successfully used BBR for several difficult applications in our lab, including high-resolution ex vivo imaging. We freely distribute our implementation of BBR with FreeSurfer as the “bbregister” function. Finally, like Saad et al. (2009), we strongly encourage researchers to visually inspect the quality of their registrations.

Acknowledgments

Support for this research was provided in part by the National Center for Research Resources (P41-RR14075, R01 RR16594-01A1), the NCRR BIRN Morphometric Project BIRN002, and Functional Imaging Biomedical Informatics Research Network (FIBIRN) U24 RR021382, the National Institute for Biomedical Imaging and Bioengineering (R01 EB001550, R01 EB006758), the National Institute for Neurological Disorders and Stroke (R01 NS052585-01) as well as the Mental Illness and Neurosciences Discovery (MIND) Institute, and is part of the National Alliance for Medical Image Computing (NAMIC), funded through the NIH Roadmap for Medical Research, Grant U54 EB005149. Additional support was provided by The Autism and Dyslexia Project funded by the Ellison Medical Foundation. We would also like to thank the FIBIRN for setting up the cognitive fMRI tasks used in this manuscript.

References


dale, a.m., fischl, b., sereno, m.l., 1999. cortical surface-based analysis: i. segmentation and surface reconstruction. Neuroimage 9, 179–194.

desikan, r.s., segonne, f., fischl, b., quinn, b.t., dickerson, b.c., blacker, d., buckner, r.l., dale, a.m., maguire, r.p., hyman, b.t., albert, m.s., killiany, r.j., 2006. an automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31, 968–980.


duong, t.q., yacoub, e., adriany, c., hu, x., ugarbici, k., vaghvan, j.t., merkle, h., kim, s.g., 2002. high-resolution, spin-echo bold, and cbf fMRI at 4 and 7 T. Magn. Reson. Med. 48, 589–593.


fischl, b., sereno, m.l., tootell, r.b.h., dale, a.m., 1999. high-resolution inter-subject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. 8, 272–284.

hampel, f.r., ronchetti, elvezio m., rousseeuw, peter j., stahel, werner a., 1986. robust statistics — the approach based on influence functions. wiley, new york.


Miller, k.l., smith, s.m., jezzard, p., pauly, j.m., 2006. high-resolution fMRI at 1.5T using balanced SSFP. Magn. Reson. Med. 55, 161–170.


