Automatic anatomical brain MRI segmentation combining label propagation and decision fusion

Rolf A. Heckemann, a Joseph V. Hajnal, a,* Paul Aljabar,b and Alexander Hammersc

aImaging Sciences Department, MRC Clinical Sciences Centre, Imperial College at Hammersmith Hospital Campus, Du Cane Road, London W12 0HS, UK
bDepartment of Computing, Imperial College London, UK
cDivision of Neuroscience and Mental Health, MRC Clinical Sciences Centre, Imperial College at Hammersmith Hospital Campus, London, UK

Received 22 October 2005; revised 18 May 2006; accepted 23 May 2006
Available online 24 July 2006

Regions in three-dimensional magnetic resonance (MR) brain images can be classified using protocols for manually segmenting and labeling structures. For large cohorts, time and expertise requirements make this approach impractical. To achieve automation, an individual segmentation can be propagated to another individual using an anatomical correspondence estimate relating the atlas image to the target image. The accuracy of the resulting target labeling has been limited but can potentially be improved by combining multiple segmentations using decision fusion. We studied segmentation propagation and decision fusion on 30 normal brain MR images, which had been manually segmented into 67 structures. Correspondence estimates were established by nonrigid registration using free-form deformations. Both direct label propagation and an indirect approach were tested. Individual propagations showed an average similarity index (SI) of 0.754±0.016 against manual segmentations. Decision fusion using 29 input segmentations increased SI to 0.836±0.009. For indirect propagation of a single source via 27 intermediate images, SI was 0.779±0.013. We also studied the effect of the decision fusion procedure using a numerical simulation with synthetic input data. The results helped to formulate a model that predicts the quality improvement of fused segmentation combined. We demonstrate a practicable procedure that exceeds the accuracy of previous automatic methods and can compete with manual delineations.

© 2006 Elsevier Inc. All rights reserved.

Introduction

Functional and structural brain imaging are playing an expanding role in neuroscience and experimental medicine. The amount of data produced by imaging increasingly exceeds the capacity for expert visual analysis, resulting in a growing need for automated image analysis. In particular, accurate and reliable methods for segmentation (classifying image regions) are a key requirement for the extraction of information from images.

Established methods for segmenting brain volumes acquired through magnetic resonance (MR) imaging can be classified into two groups: basic tissue classification and anatomical segmentation procedures. Tissue classification can be automated (e.g., BET and FAST from the FSL library (Smith et al., 2004), Exbrain (Lemieux et al., 2003)) because local MR signal properties hold the information required to distinguish between brain and nonbrain image portions, or between white matter, gray matter and cerebrospinal fluid. Anatomical segmentation is comparatively difficult to automate because structures that are anatomically distinct do not necessarily differ in their signal properties and can be composed of more than one tissue type. Most procedures therefore rely on case-by-case interactive input of human knowledge (e.g., Hammers et al., 2003). The expert time needed to establish such segmentations puts a practical limit on the size of cohorts that can be analyzed with methods of this type.

Label propagation is an approach that provides anatomical segmentations automatically without requiring interactive human input. The to-be-segmented target image is paired with an image with a corresponding pre-prepared segmentation. Anatomical structure labels are “warped” into the space of the target subject, using a transformation derived from an estimate of anatomical correspondences between the atlas and the target images. This estimate can be obtained through image registration of the MR data sets underlying the atlas and the target. Various reports in the literature describe this approach and present measures of agreement (overlap) between automatically and manually delineated structure labels (e.g., Svarer et al., 2005; Iosifescu et al., 1997; Fischl et al., 2002). To assess the applicability of label propagation, it would be desirable to understand when and why the process fails. The present work describes the first systematic investigation of failure modes in the label propagation process.

Drawing on concepts from the field of pattern recognition, Rohlfing et al. (2003) suggest that the results of the process
can be improved upon by treating multiple propagated segmentation
of a single-target image as classifiers and combining them using
decision fusion rules. This has been shown to work for confocal
microscopy images of bee brains (Rohlfing et al., 2004a). A refinement of the method presented
by Warfield et al. (2004) weights the classifiers based on
an expectation–maximization algorithm (“STAPLE”). The assumption
is that decision fusion compensates for errors in the
individual atlases and registrations and therefore arrives at a
“consensus labeling” that is closer to the ground truth than any
of the constituent segmentations. Svärer et al. (2005) employ
this principle for human brain images and fuse multiple
propagated segmentations, but their expert atlases are less
suited for precise anatomical segmentation. In recent work by
Klein and Hirsch (2005), an automatic method is described
which enables identification of large cortical structure volumes
and can also be improved using fusion of multiple classifiers
(Klein et al., 2005). A related set of methods aims to parcellate
and quality assessment of fused label volumes section). We
subject space, new segmentations were obtained (Decision fusion
and application of an atlas is transformed to the subject’s space
using a registration method based on free-form deformation
(Rueckert et al., 1999). The aim of this work was to evaluate the
accuracy of the method by measuring agreement between
automatically and manually determined structure labels, its
precision by measuring agreement between independent fused
segmentations, and its modes of failure by analyzing individual
cases of label disagreement. We describe a model that predicts the
improvement that can be achieved in relation to the number of
input segmentations. We demonstrate a substantially improved
strategy for brain labeling that provides predictable accuracy at a
level sufficient for a number of scientifically and clinically relevant
applications.

Materials and methods

Overview

The following is an overview of the framework we employed
to evaluate atlas-based segmentation. Our study data consisted of
manually segmented MR images (Subjects and MRI data
section). The term “atlas” in this framework refers to the pairing
of an anatomical image and its corresponding label volume
(Expert segmentations section). The geometric transformation
between an atlas and a subject is computed using a nonrigid
registration algorithm (Image registration section). Using this
geometric transformation, the set of labels in the coordinate
system of an atlas is transformed to the subject’s space
(Generation of propagated label volumes section). We assessed
the accuracy and precision of resulting label volumes by
standard metrics, using a leave-one-out cross-validation approach
with manual segmentations as the gold standard reference
(Quality assessment of individual propagated label volumes
section). Using different starting atlases, we obtained a number of
different propagated label volumes for each subject. By
applying vote-rule-based decision fusion at every voxel in
subject space, new segmentations were obtained (Decision fusion
and quality assessment of fused label volumes section). We
compared the direct segmentation–fusion approach with alter-
native, indirect strategies: (1) propagation of a single atlas to a
number of intermediate images, forming a set of atlases, which
are in turn propagated to the subject’s space, and (2) obtaining a
single fused segmentation of an intermediate image, which is
then propagated to the target (Indirect fused label propagation
section). Based on subject data as well as simulated data, we
investigated the influence of the number of atlases on
segmentation accuracy (A model for fused label propagation
section and Model testing section). Visual analysis of individual
segmentations was used to classify discrepancies between
automatically generated and reference segmentations (Visual
comparisons section).

Subjects and MRI data

Data were available from 30 volunteers, age range 20–54 years,
median age 30.5 years, 15 male, 15 female, 25 strongly right-
handed, 5 non-right-handed. All subjects gave written consent to
MR brain scanning and data analysis. The recruitment and
scanning procedures were approved by the local hospital research
ethics committee.

The data consisted of T1-weighted 3D volumes, acquired in the
coronal plane using an inversion recovery prepared fast spoiled
gradient recall sequence (GE), TE/TR 4.2 ms (fat and water in
phase)/15.5 ms, time of inversion (TI) 450 ms, flip angle 20°, to
obtain 124 slices of 1.5-mm thickness with a field of view of
18 × 24 cm with a 192 × 256 matrix, 1 NEX. Scanning took place at the
National Society for Epilepsy MRI Unit (Chalfont St Peter,
Buckinghamshire, UK). Scan data were resliced to create isotropic
voxels of 0.9375 × 0.9375 × 0.9375 mm, using windowed sinc
interpolation.

Expert segmentations

Each data set was accompanied by a set of labels in the
form of an annotated image volume, where every voxel was
coded as one of 67 anatomical structures (or background, code
0). These labels had been prepared using a protocol for
manually outlining anatomical structures on two-dimensional
sections from the image volume. The protocol used was an
augmented version of a published prototype (Hammers et al.,
2003). Most boundary definitions corresponded to signal
intensity differences in the images. Occasionally, boundary
definitions were based on anatomical distinctions that do not
have an intensity correlate in T1-weighted MR images. The
protocol defined such contrived or “knowledge-based” bound-
aries by transverse, coronal or sagittal planes, each intersecting
with a landmark.

Fig. 1 shows a surface rendering of a brain MR data set with
superimposed manual labels distinguished by color. The first row
in Fig. 4 shows sections through the same data set.

A subset of 12 structures that represented a range of tissues,
shapes and locations were examined in detail. Six structures
were selected from the left hemisphere: amygdala, orbitofrontal
cortex, precentral gyrus, lateral ventricle (excluding temporal
horn), pallidum and superior parietal gyrus. From the right
hemisphere, we selected 5 structures: caudate nucleus, inferior
frontal gyrus, superior temporal gyrus, hippocampus and
thalamus. The corpus callosum was selected as an unpaired
structure.
Image registration

Every subject was paired with every other subject for image registration, resulting in 870 (30×29) image pairs. Intracranial structures were extracted from the target MR images using “BET” from the FSL library (Smith et al., 2004). All image pairs were aligned using 3D voxel-based registration, maximizing normalized mutual information (Studholme et al., 1999) in three steps. Rigid and affine registration corrected for global differences. In the third, nonrigid step, alignment of details in the image pair was achieved by manipulating a free-form deformation represented by displacements on a grid of control points blended using cubic B-splines (Rueckert et al., 1999). In this algorithm, the spacing of control points defines the local flexibility of the nonrigid registration. The nonrigid registration was carried out in a multi-resolution fashion using control point spacings of 20 mm, 10 mm, 5 mm and 2.5 mm. Registrations on different atlas-target pairs were carried out in parallel on a cluster of 272 Linux PCs, controlled by Condor software (Version 6.7.3, http://www.cs.wisc.edu/condor/).

The output of the registration process is thus a geometric transformation that, when applied to the image component of an atlas, maximizes the similarity of that image to the target within the constraints of the registration algorithm.

Generation of propagated label volumes

The transformations resulting from the final nonrigid registration step were applied to the source labels using nearest-neighbor interpolation, generating 29 propagated label volumes for each target individual. We refer to these transformed sets as the individual nonrigidly propagated label volumes.

To compare alternative registration strategies, we also generated propagated label volumes based on rigid and affine registration, using the intermediate (rigid, affine and nonrigid with a control point spacing of 20 mm) transformation output to propagate the label volumes. We refer to these transformed sets as the individual affine-, rigid- and coarse nonrigid propagated label volumes.

Quality assessment of individual propagated label volumes

The similarity index (SI) is defined as the ratio between the volume of the intersection (overlap) and the mean volume of a pair of labels in the same coordinate space (Zijdenbos et al., 1994):

\[ \text{SI} = \frac{2n(L_a \cap L_b)}{n(L_a) + n(L_b)} \]

where \( L_a, L_b \) : labels compared; \( n \) : number of voxels.

This measure ranges from 0 (for labels that do not overlap) to 1 (for labels that are identical). To compare the propagated labels of individual structures against the manual segmentation, we determined the similarity index for each structure (SI\(_L\)) in each target. Some authors use a different, but related measure, the overlap ratio (OR), defined as:

\[ \text{OR} = \frac{n(L_p \cap L_m)}{n(L_p) + n(L_m)} \]

It serves the same purpose as SI and can be converted to SI by the following formula:

\[ \text{SI} = \frac{2 \text{OR}}{1 + \text{OR}} \]

To facilitate comparison, we express key results of this study as OR in addition to SI.

To arrive at a summary similarity measure across all labels for a given pair of segmentations, we calculated SI\(_m\) as the mean SI for all labels from 1 to 67, i.e., excluding background. Where appropriate, we use the notation SI\(_L\)(L1, L2), where \( x \) is one of \( s \) or \( m \) and L1 and L2 denote of the label origins.

From the five left/right structure pairs which showed the lowest SI\(_L\) results on average, we selected one of each pair for visual examination. These were the temporal horn of the left ventricle (48), the left fusiform gyrus (16), the right nucleus accumbens (37) and the right cingulate gyrus (25). The fifth structure pair, pallidum (42), already appeared in the list of a priori representative structures described in Expert segmentations section.

We compared all generated segmentations of each target volume using the manual label volume of the target as the reference segmentation.

Decision fusion and quality assessment of fused label volumes

Label volumes represent classifiers that assign a structure label to every voxel in the corresponding MR image volume. To combine the information from multiple individual propagated label volumes into a consensus segmentation, the classifiers were fused: the consensus class of each voxel was defined as the modal value of the distribution of the individual label assignments (vote rule decision fusion as described by Kittler et al., 1998, and used by Hammers et al., 2003; Klein et al., 2005). Fused label volumes were compared to manual segmentations by SI\(_m\) and SI\(_L\). In addition, we compared pairs of fused label volumes derived from combining two independent subsets of individuals, again using SI\(_m\) and SI\(_L\) in order to assess precision. We use the notation F\(_n\)N (fused nonrigidly propagated label volume), where \( n \) indicates the number of classifiers considered.

We chose odd numbers of classifiers for all subsets to increase the chance of unique modal values. If, however, there was more than one modal value in the distribution (a tie between votes), one of the modal values was assigned at random to the voxel in question.
Indirect fused label propagation

We hypothesized that the benefit of fusing propagated labels is in part due to anatomical variability of the atlases used as label sources. To estimate the importance of this factor, we set up an additional label propagation experiment, where a single atlas was used as the label source. The labels were propagated to intermediate subjects, then propagated and fused in the space of the target. The assumed error-compensating effect of label fusion is retained in this approach. For each of the 870 atlas-target combinations, 14 sets of intermediate targets were chosen randomly from the remaining 28 subjects, varying the size of the intermediate set \( n = \{1, 3, \ldots , 27\} \). Each indirectly propagated segmentation was assessed using \( S_{\text{Im}} \) (indirect fused, manual).

In a separate experiment, we determined the accuracy of label propagation when using an atlas which was itself generated by a propagation–fusion process. For this indirect approach, we propagated and fused 29 atlases in the space of a single intermediate subject. The resulting segmentation was then propagated to the target. The procedure is similar to the maximum probability mapping method presented in Hammers et al. (2003).

A model for fused label propagation

Anatomical image segmentation by label propagation is subject to both systematic and random errors.

Systematic errors arise during the stage of defining the reference segmentation as discrepancies between the reference labeling and the unknowable segmentation of the reference object. They can also result from bias in the registration process. Random errors arise from variability in the placement of individual labels and in individual registrations. This suggests that, when multiple labels are fused, the random errors will diminish, leaving systematic differences between the fused label volume and the ground truth segmentation. As the number of input label volumes increases, the SI values are expected to increase from an initially low level resulting from the combined effects of both systematic and random errors, towards an asymptotic value that reflects only the systematic errors. Assuming Gaussian variation of SI, we expect the SI to evolve with the number \( n \) of classifiers fused according to the secular equation:

\[
S_{\text{Im}}(n) = 1 - a - b \frac{n}{\sqrt{n}},
\]

where \( a \) and \( b \) are parameters to be determined. Parameter \( a \) reflects systematic differences between the labels being compared, whereas \( b \) is related to the random variability of the propagated labels. Consistent labels result in a high SI, so that \( a \) becomes small.

Model testing

We tested the above model with data from brain label propagation and with simulated data.

For each of the 30 target brain images, we created consensus (fused) segmentations from subsets of propagated label volumes of varied sizes \( n = \{3, 5, 7, \ldots , 29\} \). Where possible (for \( n \leq 13 \)), we used multiple nonoverlapping subsets. Individual and fused label volumes were then compared with the manual atlas and with other fused label volumes to determine the behavior of \( S_{\text{Im}} \) as a function of \( n \).

Numerical simulations were performed using a two-dimensional model as follows (Fig. 2): A filled circle of radius 10 mm in an image matrix with \( 1 \text{ mm} \times 1 \text{ mm} \) pixels was defined to represent a ground truth label \( I_{\text{orig}} \). A random free-form deformation was applied by overlaying a grid of control points with 8 mm spacing and displacing these control points. The displacements were drawn from a Gaussian distribution with a mean of zero and a standard deviation of \( \sigma_{\text{sys}} \) (step 1 in Fig. 2). The resulting label \( I_{\text{sys}} \) represents an approximation to the ideal circular label containing systematic error. The model label was then deformed multiple times with grid-point displacements drawn from a different Gaussian distribution with mean zero and standard deviation \( \sigma_{\text{rand}} \), producing an ensemble of shapes that contained random errors, representing propagated labels (step 2 in Fig. 2). In step 3, independent ensembles were fused to create labels (\( I_{\text{fused}} \)) that represented estimates of the original, circular label, subject to systematic and random error. The agreement (as measured by SI) of individual and fused labels with the original gold standard label and with other fused labels was then investigated to explore the behavior of SI as a function of the number of classifiers fused.

Both the simulated data and experimental values of SI were fitted to the model of Eq. (4). We explored the variation of the parameters \( a \) and \( b \) as functions of the standard deviations used for the simulated data (\( \sigma_{\text{sys}} \) and \( \sigma_{\text{rand}} \)). Using the brain label data, we determined the variation of \( a \) and \( b \) depending on the registration method employed (unregistered, rigid propagated, affine propagated, coarse nonrigid propagated and nonrigid propagated).

Visual comparisons

For each of the 16 selected structures, we determined a target subject that appeared to be particularly problematic for the label propagation–fusion method. To measure this property, we determined the difference between the values of SI, (F13N, manual) and SI, (F13N, F13N). For the target subject that showed the largest discrepancy value, we generated a structure-specific color overlay on the MR image that identified four types of anatomical structures.
of voxels: background; labeled as part of the structure by manual segmentation only; labeled as part of the structure by F13N only; and labeled by both segmentations as part of the structure (overlap; cf. Fig. 3). Based on appearance, the discrepancies were assigned to one of five types (cf. Table 1). Random error (RND) can result from small discrepancies such as those introduced through interpolation. “Greedy/shy” labeling (GSL) error systematically places the label boundary beyond or short of the reference label but preserves its shape. Label propagation failure (LPF) was noted when a group of connected voxels was assigned to a structure in error. Manual segmentation failure (MSF) was recorded if in retrospect the manual segmentation was questionable in a region showing discrepancy. Planar boundary error (PBE) occurs when a knowledge-based boundary is displaced.

**Results**

**Global segmentation comparisons**

The quality of automatically generated segmentations as measured by the mean similarity index ($S_I$) for comparison with expert manual segmentations is shown in Table 2. Using anatomically detailed correspondence estimates, such as those obtained by nonrigid registration, yields a substantial benefit in agreement over the simple affine method. Applying multiclassifier

---

**Table 1**

Summary of error types

<table>
<thead>
<tr>
<th>Abbreviated error type</th>
<th>Error type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RND</td>
<td>Random error</td>
<td>Individual discrepant voxels resulting from interpolation</td>
</tr>
<tr>
<td>GSL</td>
<td>Greedy/shy labeling</td>
<td>Connected areas of discrepancy, following the structure outline</td>
</tr>
<tr>
<td>LPF</td>
<td>Label propagation failure</td>
<td>Misassignment of a group of connected voxels to a structure</td>
</tr>
<tr>
<td>MSF</td>
<td>Manual segmentation failure</td>
<td>Discrepancies due to mistakes in the expert segmentation</td>
</tr>
<tr>
<td>PBE</td>
<td>Planar boundary error</td>
<td>Misplacement of a knowledge-based boundary</td>
</tr>
</tbody>
</table>

**Table 2**

Mean agreement levels of individually propagated volumes (columns 2, 3) and propagated and fused volumes (columns 4, 5)

<table>
<thead>
<tr>
<th>Registration method</th>
<th>Individually propagated volumes ($n=870$)</th>
<th>Fused propagated volumes ($n=30$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_I_m$, min–max</td>
<td>$OR_m$, min–max</td>
</tr>
<tr>
<td>Affine</td>
<td>0.668±0.028, 0.502±0.032</td>
<td>0.754±0.023, 0.616±0.030,</td>
</tr>
<tr>
<td></td>
<td>0.538–0.731, 0.368–0.576</td>
<td>0.662–0.790, 0.495–0.653</td>
</tr>
<tr>
<td>Nonrigid</td>
<td>0.754±0.016, 0.605±0.020</td>
<td>0.836±0.009, 0.718±0.013,</td>
</tr>
<tr>
<td></td>
<td>0.692–0.799, 0.529–0.665</td>
<td>0.820–0.853, 0.695–0.744</td>
</tr>
</tbody>
</table>
Assessment of individual and fused label propagation for individual structures

Fig. 9 confirms for every single structure the difference in accuracy between several propagation approaches. Table 3 provides a key to the structure codes used in Fig. 9. Agreement with the manual atlases increased in the order of affine-individual<nonrigid-individual<affine-fused<max. probability<nonrigid-fused propagation. Affine-fused labels were numerically at the level of individual nonrigid labels. The sequence of rows in Fig. 4 shows how discrepancies diminish as the automatic labeling approaches improve.

Table 4 shows average SI results for representative as well as problematic structures. Fig. 3 shows sample sections through reviewed structures.

Most structures were segmented with a small amount of boundary error of the RND type. Where the intensity gradient in the MR image that corresponded to the boundary was weak, GSL error was prevalent.

Some structures (anterior cingulate gyri, fusiform gyri, nuclei accumbentes, pallidi, temporal horns of the lateral ventricles) are segmented with comparatively low levels of accuracy and precision. This result was consistent between pairs of these structures.

Due to the small size of the nucleus accumbens, classifying discrepancies between the automatic and manual labels was difficult. The largest discrepancy in the sample reviewed appeared to be of the GSL type.

The temporal horns of the lateral ventricles pose the same challenge, although some of the discrepancies were clearly attributable to LPF.

Label propagation based on nonrigid registration segmented the majority of structures with substantial overlap with the manual label volumes (SI>0.70 in 79% of all measurements). Fusing labels from 29 atlases lifted this percentage to 95%. Visual review of label volumes generated by subtracting a manual from a propagated label volume shows a high level of shape fidelity (Fig. 4).

Discussion

If a ground truth segmentation were available for a given three-dimensional brain image, it could be used to segment another brain image perfectly, assuming that exact voxel-to-voxel anatomical correspondence could be established between the image pair. This idea forms the basis of label propagation.

In reality, however, errors limit the accuracy of segmentations obtained by label propagation. Types of error include inaccuracies of the atlas used as a starting point, errors in the registration process used to estimate correspondence and localized failure of the assumption of one-to-one correspondence between individual brains. Finally, the assessment of a segmentation will contain errors, too, consisting of the discrepancies between the reference segmentation used for the target and its ground truth segmentation.

When multiple segmentations of a single target are available, it becomes possible to distinguish between systematic bias and randomly distributed errors. If segmentations are regarded as classifiers and combined using a suitable decision fusion algorithm, random errors will tend to cancel each other out. The resulting fused segmentation can be more accurate than any of the
Fig. 4. Sectional label representations, superimposed on a gray scale MR image. The top two rows show sections from full label volumes. The remaining rows highlight areas of disagreement between different label pairs.
constituent segmentations but will still be subject to systematic bias arising from the input data and the process itself.

In this study, we established experimentally how label propagation and decision fusion can be combined to automate the segmentation of MR images of the human brain. Although the principle has been described previously (Rohlfing et al., 2004a; Svarer et al., 2005; Klein et al., 2005), we present the first comprehensive assessment using carefully validated input data of 30 subjects. To our knowledge, label propagation–fusion has not been studied on a comparably large data set. We found that for the majority of macroscopic brain structures, a propagation–fusion approach yields consistent segmentations of a quality that compares favorably with expert manual segmentation. Many of these structures are relevant in functional research and investigations of disease progression, e.g., the hippocampus, the thalamus and the orbitofrontal cortex.

To determine whether an automatic segmentation approach is appropriate for a particular application, it is important to understand how and why the process can fail at the level of individual structures. Based on a systematic review of individual cases of discrepancy between fused-propagated and manually prepared reference segmentations, we identified and named distinct patterns which are likely to correspond with different modes of failure. These novel results illustrate the current limitations and point the way towards further improvement of the method.

We investigated the correlation between the number of classifiers used to form a fused segmentation and the accuracy of the result and found a model that describes this correlation. We verified the applicability of this model using simulated data. By performing parameter fitting, it was possible to quantify the impact on agreement levels of the scale of the transformation used to propagate individual brain segmentations (affine, coarse nonrigid, fine nonrigid). Fit parameters obtained in this fashion could be used to evaluate future methods of estimating intersubject correspondence and to assess newly developed atlases.

As a means of combining decisions, vote rule fusion, though basic, is appropriate in this context. Averaging of labels is ill-suited...
because labels assigned to voxels are qualitative variables. Assignment probabilities in our approach are binary. This enabled us to test the method on a large number of structures in a large data set. More complex fusion approaches, e.g., those based on sum or product rules (Kittler et al., 1998), might yield benefit if probabilistic label assignments, for example based on interpolation at the label transformation stage, were made available. Expectation–maximization strategies, such as STAPLE and derived methods, can improve segmentation results by differential weighting of classifiers (Warfield et al., 2004; Rohlfing et al., 2004b). We are planning to test expectation–maximization approaches in a separate study using the results obtained in the present study as a baseline.

Combining even a small number \( n \) of classifiers (propagated label sets) results in a marked improvement of the agreement between the fused label volume and a segmentation prepared by an expert. Increasing \( n \) improves this agreement further, albeit with diminishing benefits.

As Fig. 5 shows, the data from fused label propagation on the human brain was consistent with the model described by Eq. (4). For comparisons between independent fused label volumes, the \( a \) parameter of the model fit was much smaller, proving that the averaging that takes place in the label fusion process leads to highly precise (if not necessarily unbiased) results.

Indirect propagation showed the same functional behavior as direct propagation, but with lower absolute SI values (and an accordingly large \( a \) parameter). The maximum SI, which was achieved with the largest intermediate set, only just exceeds the value for single direct propagation of a source label to a target. This is congruous with the assumption that the fusion process decimates errors introduced in the registration, so that the remaining discrepancies are due to differences in the definitions of the source atlas and the target's reference. The \( b \) parameter was small for indirect propagation, supporting the view that the segmentations of the intermediate subjects were highly consistent with each other, because they were all generated by propagation from a single source atlas and therefore carried the same bias. The results of this experiment are a caveat against “boot-strapping” approaches, where a single manual segmentation is propagated to a multitude of brains in order to then combine them into a “universal” brain map (e.g., Collins et al., 1994). An important factor in the success of our direct propagation–fusion approach is the independence of the individual manual segmentations used as input.

The simulations reproduced all key features of the experimental results: the \( b \) parameter correlates with \( \sigma_{\text{rand}} \) (Fig. 8), showing that it can be used as an indication of the precision of segmentations fused in the experiment. The \( a \) parameter, on the other hand, describes the accumulated systematic error that cannot be eliminated by considering further classifiers in the fusion process. We conclude that the model (Eq. (4)) describes the core behavior well. The model is built on the assumption that SI values produced for the fused labels are normally distributed. Because SI is bounded, this assumption is formally violated, so we used a Kolmogorov–Smirnov test to check for normality in subject data as well as in simulations using randomly varied label shapes. All the subject data was found to be indistinguishable from a normally distributed variable by this metric. In simulations we were only able to cause a detectable violation of the normal distribution for very high SI values (>0.97, i.e., well beyond the range we observed in practice). Thus, Eq. (4) can usefully be applied to both the subject data and in simulations over the range of SI we encountered in practice.

Fused label propagation generally produces reasonable results, even for structures that show large anatomical variability, such as the amygdala (range of manually determined volumes across subjects: 982–2173 mm\(^3\)). There were a small number of exceptions, which we reviewed in order to identify and describe failure modes. The fusiform and cingulate gyri can be particularly problematic if the target subjects show unusual configurations of these structures. The fusiform gyrus can be interrupted, with
At the other end of the scale of intersubject variability are structures such as the thalamus and the caudate nucleus. Average $S_I$ values above 0.90 indicate that automated segmentation works as well as manual segmentation for these structures, considering that interobserver results from the literature indicate comparable or lower interobserver agreement levels for expert segmentations (Fischl et al., 2002; Spinks et al., 2002). Other automatic methods for segmenting these structures have been reported but yield similar or lower levels of agreement and precision. A dynamic contour-model-based approach for segmenting the thalamus by Amini et al. (2004) achieved OR results of 0.80 on average, whereas our method showed an average OR of 0.82 ± 0.02 (F29N sets) for this structure. Using a different overlap measure, Barra and Boire (2001) find spatial accuracy values of 0.90 ± 0.04 and 0.89 ± 0.03 for the left and right thalamus. Applying the same assessment method to our F29N data, we find values of 0.91 ± 0.02 and 0.90 ± 0.03 for the left and right thalamus. The same authors segment the caudate nucleus with a spatial accuracy of 0.85 ± 0.06 (left) and 0.84 ± 0.05 (right), our figures for comparison are 0.89 ± 0.03 and 0.90 ± 0.03. This comparison shows that specialized methods that identify a particular structure are not necessarily superior to our approach, which segments the entire brain into currently 67 structures in a single pass.

A detailed comparative study of various hippocampus segmentation methods has recently been reported by Carmichael et al. (2005). The authors report interrater overlap ratios (OR) of 0.65 ± 0.06 and automatic-to-manual overlaps of 0.61 ± 0.05. Again, our results compare favorably (OR 0.70 ± 0.05 for fused propagated labels).

The work by Svarer et al. (2005) demonstrated that classifier fusion improves MR brain segmentations. Their approach of creating a thresholded probability map per region of interest has been designed for extracting data for samples of anatomical regions from PET images. It does not lend itself to identifying anatomical boundaries or measuring structure volumes because the regions of interest may intersect with each other.

A recent paper by Klein et al. (2005) also showed improvements in label consistency when fusion methods were used on atlases generated by their automated approach. This group reports that maximum label agreement was 82.50% for 10 classifier subjects. Applying the same measure to our data, we find a value of 85.1% for fusion of 9 classifiers. The difference may be attributed to the smaller number of contrived boundaries in our source labels and to the use of a higher dimensional registration procedure. We have also used a larger number of labeled brains and so have been able to move closer to convergence of the labels. The concept of label propagation by these means places a great premium on the fidelity of the source labels, but also justifies the substantial effort required for manual labeling, because the resulting atlases can be used to label new brains as required.

A key question concerning automatic labeling of an unseen brain is how to decide how much confidence to place in the method. Even though results shown in the present study are promising, it is possible that label propagation to a new brain may fail for a given structure. Because by definition there would not be a manual label defined, we need some other measure of

---

**Table 3**

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>$S_I$ (F29N, manual)</th>
<th>$S_I$ (F13N, F13N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Temporal lobe</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1; 2</td>
<td>Hippocampus</td>
<td>0.83; 0.81</td>
<td>0.91; 0.89</td>
</tr>
<tr>
<td>3; 4</td>
<td>Amygdala</td>
<td>0.81; 0.80</td>
<td>0.90; 0.89</td>
</tr>
<tr>
<td>5; 6</td>
<td>Anterior temporal lobe, medial part</td>
<td>0.85; 0.84</td>
<td>0.92; 0.91</td>
</tr>
<tr>
<td>7; 8</td>
<td>Anterior temporal lobe, lateral part</td>
<td>0.85; 0.85</td>
<td>0.91; 0.90</td>
</tr>
<tr>
<td>9; 10</td>
<td>Parahippocampal and ambient gyr</td>
<td>0.82; 0.81</td>
<td>0.90; 0.88</td>
</tr>
<tr>
<td>11; 12</td>
<td>Superior temporal gyr</td>
<td>0.88; 0.88</td>
<td>0.93; 0.93</td>
</tr>
<tr>
<td>13; 14</td>
<td>Middle and inferior temporal gyr</td>
<td>0.87; 0.86</td>
<td>0.93; 0.91</td>
</tr>
<tr>
<td>15; 16</td>
<td>Fusiform gyr</td>
<td>0.74; 0.72</td>
<td>0.84; 0.81</td>
</tr>
<tr>
<td>30; 31</td>
<td>Posterior temporal lobe</td>
<td>0.86; 0.85</td>
<td>0.92; 0.92</td>
</tr>
<tr>
<td><strong>Posterior fossa</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17; 18</td>
<td>Cerebellum</td>
<td>0.95; 0.95</td>
<td>0.97; 0.97</td>
</tr>
<tr>
<td>19</td>
<td>Brainstem</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td><strong>Insula and cingulate gyri</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20; 21</td>
<td>Insula</td>
<td>0.86; 0.86</td>
<td>0.93; 0.93</td>
</tr>
<tr>
<td>24; 25</td>
<td>Gyrus cinguli, anterior part</td>
<td>0.78; 0.83</td>
<td>0.87; 0.89</td>
</tr>
<tr>
<td>26; 27</td>
<td>Gyrus cinguli, posterior part</td>
<td>0.84; 0.82</td>
<td>0.91; 0.89</td>
</tr>
<tr>
<td><strong>Frontal lobe</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28; 29</td>
<td>Middle frontal gyrus</td>
<td>0.85; 0.85</td>
<td>0.91; 0.92</td>
</tr>
<tr>
<td>50; 51</td>
<td>Precentral gyrus</td>
<td>0.84; 0.84</td>
<td>0.91; 0.91</td>
</tr>
<tr>
<td>54; 55</td>
<td>Orbifrontal gyr</td>
<td>0.84; 0.85</td>
<td>0.91; 0.91</td>
</tr>
<tr>
<td>56; 57</td>
<td>Inferior frontal gyrus</td>
<td>0.84; 0.83</td>
<td>0.90; 0.91</td>
</tr>
<tr>
<td>58; 59</td>
<td>Superior frontal gyrus</td>
<td>0.85; 0.86</td>
<td>0.91; 0.92</td>
</tr>
<tr>
<td><strong>Occipital lobe</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64; 65</td>
<td>Lingual gyrus</td>
<td>0.83; 0.83</td>
<td>0.90; 0.90</td>
</tr>
<tr>
<td>66; 67</td>
<td>Cuneus</td>
<td>0.81; 0.82</td>
<td>0.89; 0.89</td>
</tr>
<tr>
<td>22; 23</td>
<td>Lateral remainder of occipital lobe</td>
<td>0.83; 0.83</td>
<td>0.91; 0.91</td>
</tr>
<tr>
<td><strong>Parietal lobe</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52; 53</td>
<td>Gyrus rectus</td>
<td>0.80; 0.81</td>
<td>0.88; 0.89</td>
</tr>
<tr>
<td>60; 61</td>
<td>Postcentral gyrus</td>
<td>0.83; 0.82</td>
<td>0.90; 0.89</td>
</tr>
<tr>
<td>62; 63</td>
<td>Superior parietal gyr</td>
<td>0.86; 0.86</td>
<td>0.92; 0.92</td>
</tr>
<tr>
<td>32; 33</td>
<td>Inferolateral remainder of parietal lobe</td>
<td>0.85; 0.84</td>
<td>0.91; 0.91</td>
</tr>
<tr>
<td><strong>Central structures</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34; 35</td>
<td>Caudate nucleus</td>
<td>0.89; 0.90</td>
<td>0.94; 0.94</td>
</tr>
<tr>
<td>36; 37</td>
<td>Nucleus accumbens</td>
<td>0.72; 0.70</td>
<td>0.85; 0.84</td>
</tr>
<tr>
<td>38; 39</td>
<td>Putamen</td>
<td>0.89; 0.90</td>
<td>0.94; 0.94</td>
</tr>
<tr>
<td>40; 41</td>
<td>Thalamus</td>
<td>0.91; 0.90</td>
<td>0.95; 0.95</td>
</tr>
<tr>
<td>42; 43</td>
<td>Pallidum</td>
<td>0.80; 0.80</td>
<td>0.90; 0.90</td>
</tr>
<tr>
<td>44</td>
<td>Corpus callosum</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Ventricles</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45; 46</td>
<td>Lateral ventricle (excluding temporal horn)</td>
<td>0.89; 0.91</td>
<td>0.94; 0.95</td>
</tr>
<tr>
<td>47; 48</td>
<td>Lateral ventricle, temporal horn</td>
<td>0.66; 0.62</td>
<td>0.78; 0.75</td>
</tr>
<tr>
<td>49</td>
<td>Third ventricle</td>
<td>0.84</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Left-sided structures are referred to by even numbers, right-sided ones by odd numbers.
Table 4
SI results and most prevalent error types for selected structures

<table>
<thead>
<tr>
<th>Structure</th>
<th>Code, SI (individual, manual)</th>
<th>SI (F2N, manual)</th>
<th>Error type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average of 67</td>
<td>−, 0.754±0.016 0.836±0.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Representative structures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L amygdala</td>
<td>4, 0.71±0.064 0.80±0.046</td>
<td>LPF</td>
<td></td>
</tr>
<tr>
<td>R caudate nucleus</td>
<td>35, 0.84±0.030 0.90±0.016</td>
<td>RND</td>
<td></td>
</tr>
<tr>
<td>Corpus callosum</td>
<td>44, 0.81±0.031 0.87±0.023</td>
<td>GSL</td>
<td></td>
</tr>
<tr>
<td>R hippocampus</td>
<td>1, 0.75±0.055 0.83±0.042</td>
<td>LPF</td>
<td></td>
</tr>
<tr>
<td>R inferior frontal gyrus</td>
<td>57, 0.74±0.065 0.83±0.037</td>
<td>LPF</td>
<td></td>
</tr>
<tr>
<td>L lateral ventricle</td>
<td>46, 0.86±0.034 0.91±0.023</td>
<td>GSL</td>
<td></td>
</tr>
<tr>
<td>L orbitofrontal cortex</td>
<td>54, 0.76±0.040 0.84±0.024</td>
<td>LPF,</td>
<td></td>
</tr>
<tr>
<td>R thalamus</td>
<td>41, 0.86±0.021 0.90±0.009</td>
<td>RND,</td>
<td>LPF</td>
</tr>
<tr>
<td>L pallidum</td>
<td>42, 0.72±0.069 0.80±0.047</td>
<td>MSF</td>
<td></td>
</tr>
<tr>
<td>L precentral gyrus</td>
<td>50, 0.75±0.051 0.84±0.024</td>
<td>LPF</td>
<td></td>
</tr>
<tr>
<td>L sup parietal gyrus</td>
<td>62, 0.78±0.049 0.86±0.031</td>
<td>LPF</td>
<td></td>
</tr>
<tr>
<td>R sup temp gyrus</td>
<td>11, 0.82±0.034 0.88±0.022</td>
<td>PBE</td>
<td></td>
</tr>
<tr>
<td>R thalamus</td>
<td>41, 0.86±0.021 0.90±0.009</td>
<td>RND,</td>
<td>LPF</td>
</tr>
</tbody>
</table>

Examples of “poor” structures

R anterior cingulate gyrus | 25, 0.73±0.082 0.83±0.064 | LPF       |
L fusiform gyrus           | 16, 0.57±0.122 0.72±0.090 | LPF       |
R nucleus accumbens        | 37, 0.59±0.076 0.70±0.076 | GSL       |
L temporal horn            | 48, 0.46±0.093 0.62±0.067 | LPF       |

The likelihood of success. Visual inspection provides one such check, but is time consuming and encumbered by intra- and inter-observer variability. The model for improved label consistency obtained in this work provides a possible alternative quality assurance measure. In Fig. 6, all the methods of label propagation show increasing precision with increasing n. This precision is not in itself a guarantee of appropriate labeling (fusion of many labels without any registration will lead to convergence to a mean label, but not a useful one). It is notable, however, that the approach to consistency with the number of fused label was slower for less anatomically detailed registration methods. The b value was largest when no registration was employed and showed the minimum value when the nonrigid registration with the smallest control point spacing was used. This suggests that b may be a useful metric to compare the ability of different image registration algorithms to provide a useful intersubject correspondence estimate. The approach to consistency depends on registration accuracy, because the labeled atlas brains are all different from each other and are all different from the target. Agreement between individually propagated labels (which leads to rapid convergence with increasing n, i.e., a small b value) requires both anatomically consistent labeling and valid anatomical correspondence estimates. When label fusion is employed, it is always possible to calculate intermediate results up to the maximum n set by the number of available atlases, and so a b value can be determined. It remains to be seen if the absolute value of b can be used as a measure of label accuracy to add confidence to the high level of precision that fusion achieves.

Conclusions

Intersubject anatomical correspondence estimates can be used to propagate expert segmentations of three-dimensional magnetic resonance brain images to previously unsegmented image volumes. We have presented an assessment of this method on a set of 30 brain images that have corresponding high-quality manual segmentations.

Decision fusion provides a robust self-consistent way of labeling brains. By introducing a model, we confirmed that the mechanism of convergence of the fused labels is likely to be averaging out of random errors in both the source labels and the registration process. The model also predicts the number of labels required to achieve a required consistency level of the resulting segmentations.

Converged labels are self-consistent, but this does not guarantee an accurate manual segmentation. Anatomically accurate segmentation requires both detailed independent input labels and an anatomically faithful registration method. This implies that indirect methods that fuse many labels that ultimately derive from the same segmentation are intrinsically limited. For independent label sources, the rate of convergence is a likely marker of the veracity of the result, measured by the b parameter introduced here. Although this quality measure remains to be investigated, the results already obtained strongly support the notion that label propagation by fusion is a robust method and is suitable for a wide range of applications where consistent anatomical brain segmentation is required.

The methods described in this paper provide a basis for creating authoritative labels of unseen brains. The approach to convergence of the propagate labels as the number of fused segmentations increases provides a means of assessing the likely suitability of the target labeling. We are now studying clinical exemplars such as temporal lobe epilepsy, where quantifying structural change has hitherto required time-consuming manual segmentation. Other areas of application may include anatomical definition of regions of interest in functional studies, where structural MR imaging is used as an anatomical reference.

Acknowledgments

Rolf Heckemann was supported by a research fellowship from The Dunhill Medical Trust. The authors would like to thank the National Society for Epilepsy, Chalfont St Peter, Buckinghamshire, UK, for making available the original 30 MRI data sets used in the creation of the atlases. This work was funded by the UK Engineering and Physical Sciences Research Council as part of the IXI project.

References

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C.,


