Reliability of tissue volumes and their spatial distribution for segmented magnetic resonance images

Valerie A. Cardenas\textsuperscript{a,}\textsuperscript{*}, Frank Ezekiel\textsuperscript{b}, Victoria Di Sclafani\textsuperscript{d}, Bryon Gomberg\textsuperscript{c}, George Fein\textsuperscript{d}

\textsuperscript{a}Department of Radiology, University of California, San Francisco, and San Francisco Veterans Affairs Medical Center, 4150 Clement St., San Francisco, CA 94121, USA
\textsuperscript{b}Department of Psychiatry, University of California, San Francisco, and San Francisco Veterans Affairs Medical Center, 4150 Clement St., San Francisco, CA 94121, USA
\textsuperscript{c}Department of Bioengineering, University of Pennsylvania Medical Center, 3400 Spruce St., Philadelphia, PA 19104, USA
\textsuperscript{d}Neurobehavioral Research, Inc., Corte Madera, CA 94925, USA

Received 5 September 2000; received in revised form 31 January 2001; accepted 25 February 2001

Abstract

Before using MRI tissue segmentation in clinical studies as a dependent variable or as a means to correct functional data for differential tissue contribution, we must first establish the volume reliability and spatial distribution reproducibility of the segmentation method. Although several reports of volume reliability can be found in the literature, there are no articles assessing the reproducibility of the spatial distribution of tissue. In this report, we examine the validity, volume reliability, and spatial distribution reproducibility for our K-means cluster segmentation. Validation was examined by classifying gray matter, white matter, and CSF on images constructed using an MRI simulator and digital brain phantom, with percentage volume differences of less than 5\% and spatial distribution overlaps greater than 0.94 (1.0 is perfect). We also segmented repeat scan MRIs from 10 healthy subjects, with intraclass correlation coefficients greater than 0.92 for cortical gray matter, white matter, sulcal CSF, and ventricular CSF. The original scans were also coregistered to the repeat scan of the same subject, and the spatial overlap for each tissue was then computed. Our overlaps ranged from 0.75 to 0.86 for these tissues. Our results support the use of K-means cluster segmentation, and the use of segmented structural MRIs to guide the analysis of functional and other images. © 2001 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: K-means clustering; Magnetic resonance image; Brain; Image processing; Tissue classification

* Corresponding author. Tel.: +1-415-221-4810, ext. 4086; fax: +1-415-502-2110.
E-mail address: valerie@itsa.ucsf.edu (V.A. Cardenas).
1. Introduction

Use of structural magnetic resonance images (MRIs) to guide analysis of functional images (e.g. PET, SPECT, fMRI) or data from other imaging modalities (e.g. magnetization transfer imaging, MR spectroscopy, diffusion weighted imaging) is becoming increasingly popular. Anatomic regions of interest (ROIs) can be better delineated on high resolution structural MRIs than on lower resolution functional and other images, and these ROIs can then be used to guide quantitative analysis of the lower resolution images (e.g. Migneco et al., 1994; Mountz et al., 1994; Weiner et al., 1998). Tissue segmentation from the structural MRIs can be used to correct the lower resolution images for atrophic and differential tissue compartment contribution effects (Meltzer et al., 1990, 1996a,b; Müller-Gärtner et al., 1992; Weiner et al., 1998).

MRI segmentation is most often used to assess tissue specific volumes as measures of atrophy or as differences in brain organization between diagnostic groups. The validity of the segmentation tissue classification can only be assessed against a true gold standard. Investigators have assessed the reliability of MRI segmentation by evaluating the repeatability of the resultant tissue volume measurements (Bonar et al., 1993; Byrum et al., 1996; Cohen et al., 1992; Fisher et al., 1997; Harris et al., 1999; Kikinis et al., 1992; Reiss et al., 1998). Sources of unreliability of MRI segmentation volumetric measures identified by these investigators include intra- and inter-operator variability, imperfections in data acquisition (RF inhomogeneity, motion and flow artifacts), drift in imager function over time, and partial volume effects (PVE). When segmented structural MRIs are used to guide the analysis of functional and other images, their utility depends on the validity and reliability of the segmentation on a pixel-by-pixel basis. In these cases, it is not just the volume of cortical gray matter that is important, but it is the spatial location of the gray matter voxels in the brain that is used to guide the analysis of the other imaging modality. We have found no publications assessing the reliability of segmentation algorithms on a pixel-by-pixel basis. It is difficult to validate tissue segmentation algorithms using data acquired in vivo, since there is no way of determining the true tissue classifications of each MRI voxel. Manual segmentation by an expert is often used as a gold standard for validating segmentation algorithms (Harris et al., 1999), but such efforts are hindered by intra- and inter-rater variability. Investigators at the Montreal Neurologic Institute (MNI) (Collins et al., 1998; Kwan et al., 1996) have developed a realistic digital brain phantom and MRI simulator, which can be used to evaluate image-processing methods. Using a web interface (http://www.bic.mni.mcgill.ca/brainweb), differently weighted simulated MRIs can be downloaded and used to test segmentation algorithms. The output of segmentation can be compared to the digital brain phantom to compute an objective measure of performance.

This report examines the validity of our K-means clustering segmentation approach by applying it to MRI phantom data, and then focuses on the reproducibility or reliability of the volumetric measures and of the spatial distribution of tissue categories in serially collected, segmented, anatomic images. Only by comparing the reproducibility of the spatial distribution of tissue categories across repeat imaging studies can the utility of segmentation for voxel-by-voxel co-analysis of functional images be assessed.

2. Materials and methods

2.1. McGill University brainweb images

We obtained simulated MRIs from http://www.bic.mni.mcgill.ca/brainweb, using their normal brain model. All images had 1-mm slice thickness with $1 \times 1$ mm$^2$ in-plane resolution. We obtained T1, PD, and T2 weighted images, with the following parameters: T1 (simulated 3D spoiled FLASH TR/TE, FA — 18/10, 30°), 3% noise level, 0% inhomogeneity; PD (simulated early echo from 2D multislice dual spin echo TR/TE — 3300/35), 3% noise level, 20% inhomogeneity, T2 (simulated late echo from 2D multislice dual spin echo TR/TE —
We also obtained their discrete normal brain model, which consists of a class label (integer) at each voxel, representing the tissue which contributes the most to that voxel. We then constructed a brain mask from the anatomical model (method described below), such that the skull and meninges were effectively stripped from the simulated MRIs before segmentation.

2.2. Subjects

We obtained MRIs from 10 volunteers who gave informed consent. All had clinically normal MRIs. Their age (3 males and 7 females) was 22–53 years of age; the average age was 32.4 ± 10.1 years.

2.3. MRI data acquisition

The MRI sequence consisted of a double spin-echo (DSE) and a magnetization prepared rapid acquisition gradient echo (MPRAGE) sequence performed on a Siemens 1.5 Tesla MR system. The coronal MPRAGE (1.5-mm slice thickness, 1 x 1 mm² in-plane resolution) was acquired orthogonal to the optic nerve, and the axial DSE was angulated −10° from the planum sphenoidale (3-mm slice thickness, 1 x 1 mm² in-plane resolution). The DSE sequence (TR/TE1/TE2/excitations 5000/20/80/1, 6/8 FOV, 192 x 256 acquisition matrix zero padded to 256 x 256, alpha flip angle = 90°) yielded 48 contiguous slices, each with proton-density-weighted contrast (the 20-ms echo) and T2-weighted contrast (the 80-ms echo) images. The MPRAGE (TR/TE/excitations 9.7/4/1, 6/8 FOV, 192 x 256 acquisition matrix zero padded to 256 x 256, alpha flip angle = 15°) yielded 164 contiguous slices with T1-weighted contrast.

Each subject exited the imager after undergoing the 40-min MRI imaging protocol (timepoint 1). Approximately 5 min later, the subject returned to the magnet for repositioning and the identical protocol was repeated (timepoint 2). All of the subjects completed their studies within the same 2-week period.

2.4. Image segmentation pre-processing

We removed the skull and meninges from the acquired MRIs, using the algorithm developed by Itti et al. (1997). An operator checked the brain mask and edited as needed, a 5–10-min process. The final mask was applied to the DSE data, and the voxels within the brain mask were scaled from 12 to 8 bits. We next co-registered the 1.5-mm slice thickness coronal MPRAGE data to the masked, scaled 3-mm axial DSE data using the AIR 1.0 algorithm for intermodality registration (Woods et al., 1993), resulting in a final data set of 3-mm axial images with T1, T2, and PD weighting. The coregistrations were visually checked, and misregistered DSE and MPRAGE images were rerun, using different initial parameters. In our experience, misregistration rarely occurs and did not happen with these 10 healthy subjects.

Before K-means clustering, the T2- and PD-weighted data were corrected to remove RF field inhomogeneity using the method of Jernigan et al. (1990). The MPRAGE images had greater tissue contrast and less inhomogeneity and were not corrected.

2.5. K-means cluster segmentation

We created a three-dimensional (3D) feature space using the T1-, T2-, and PD-weighted axial brain images, and used K-means cluster analysis to classify the brain tissue into gray matter, white matter, and CSF compartments. Three images were used to take advantage of the strengths of each sequence, and K-means clustering incorporated all that when classifying voxels to different tissue types. The T1 images have good gray/white/CSF contrast but do not show white matter signal hyperintensities (WMSH). The PD images are useful for detecting WMSH and delimiting subcortical structures. The T2 images are excellent for discriminating CSF from gray/white matter. We experimented with choosing additional tissue classes in our feature space, but could neither differentiate WMSHs from white and gray matter and CSF, nor separate subcortical gray matter from surrounding white matter. We therefore designed the segmentation (i.e. we chose the
centroids for the clusters, see below) so that WMSHs would end up being classified as gray matter or CSF. We then used anatomic information (i.e. gray matter or CSF appearing in an anatomic location that must be white matter) to reclassify such voxels as WMSHs. For delineation of subcortical gray matter structures, more operator-intensive methods were used.

The segmentation into gray matter, white matter and CSF was accomplished using the Statistical Analysis System (SAS) FASTCLUS Procedure clustering algorithm (SAS Institute, 1990). The centroids of three tissue clusters were selected by an automated histogram analysis procedure, were verified by the operator and were fixed throughout the clustering procedure. The FASTCLUS procedure then grew connected clusters around the cluster centroids in the three-dimensional feature space, assigning observations to the three clusters to minimize the within-cluster sum of squared Euclidean distances from the cluster centroid. The centroids in feature space (i.e. PD-weighted, T2-weighted, and T1-weighted image intensity space) were chosen to reflect the mean intensities for conservatively chosen samples of white matter, gray matter, and CSF. The automated algorithm used the voxel intensity histogram of the entire co-registered (axial, 3-mm slice thickness) T1-weighted MPRAGE data set (Fig. 1) to choose the centroids. This voxel intensity histogram is trimodal, with gray matter as the middle intensity peak and white matter as the brightest peak (consistent with T1 weighting). We defined a conservative white matter tissue sample as the 10-voxel intensities bounding the white matter mode (Fig. 1, left). We defined a similar gray matter tissue sample, except that the boundary on the high intensity side of the peak was moved into the minimum between the gray and white matter peaks. As a result of moving the gray matter boundary in this manner, WMSH fell predominantly into the gray matter rather than the white matter category on the segmented image (Fig. 1, middle). In the event of overlap between

![Histograms and images of tissue samples](image)

Fig. 1. The standardized procedure for choosing tissue samples for MRI segmentation is illustrated. The top row shows the voxel intensity histogram or its 1st derivative of the entire coregistered MPRAGE data set, and the range of tissue intensities that defined the white matter (left), gray matter (middle), and CSF (right) tissue samples. The bottom row shows a ‘middle’ slice from the MPRAGE study, with the tissue samples highlighted in black.
the intensity boundary of the white and gray matter samples, the algorithm incremented the lower boundary of the white matter sample until there was no overlap. The CSF sample was chosen by bounding the first peak of the first differential of the MPRAGE intensity histogram by 10-voxel intensities (Fig. 1, right). The algorithm for defining these tissue samples was arrived at on a trial-and-error basis after examination of different criteria on images from a very large number of subjects. We displayed these conservative tissue samples on a ‘middle’ slice of the MPRAGE study (on the same screen as the histogram) and an operator verified that the samples were anatomically correct. In cases in which the image quality was compromised (e.g. by excessive movement), the tissue category samples were easily identified as being anatomically incorrect and the study was deemed as being of insufficient quality for automated segmentation. Review of the pre-chosen tissue categories took 1–2 min of operator time per study. Fig. 2 shows the cluster centers for the 20 scans analyzed.

The voxel intensities from the T1-, T2-, and PD-weighted images for all voxels in the brain, together with the centroids for the conservative tissue samples, were then written to an ascii file. In addition, an SAS program file is created within our windows program that calls the FASTCLUS procedure with appropriate parameters (including the input data). The windows STARTUPINFO structure is used to allow calling of a child process, and that process is started with our temporary .sas file as the input. The program calculates the Euclidean distance between each voxel and each cluster center in the intensity feature space, and each voxel is assigned to its nearest cluster center. The SAS output is then read back into our program and stored as an image. Fig. 3 shows an example of the classification that resulted from K-means clustering; gray matter is displayed in dark gray, white matter in white, and CSF in light

Fig. 2. The cluster centers for 20 MRIs (10 subjects, 2 MRIs/subject) are plotted in intensity feature space. The circles show the CSF clusters, the squares show the gray matter clusters, and the triangles show the white matter clusters. This illustrates good separation of the CSF cluster on the T2 axis and good separation of all clusters on the T1 axis. Though not obvious because of subject overlap, the PD axis typically showed good separation of the gray and white clusters, with the CSF cluster falling very close to the white cluster.
gray. Segmentation of the phantom MRIs was completed at this point.

For the MRIs from our volunteers, we further divided brain tissue into more specific compartments based on anatomy. Different operators segmented the scans from timepoints 1 and 2. We separated CSF into ventricular vs. sulcal CSF and gray matter into cortical and subcortical gray matter. We defined WMSH as voxels that the K-means procedure classified as either gray matter or CSF but that were white matter by anatomic location (these voxels were also added to the white matter category). We edited the most inferior slice of the subcortical region (where the thalamus is clearly identifiable) through the top of the brain. The caudate, lenticular nucleus (globus pallidus and putamen), and thalamus were delineated from the gray scale images. The protocol used to mark these structures is in Appendix A. Completion of editing ranged from 45 min to 1 h 45 min, depending on the amount of WMSH present. Fig. 3 shows an edited segmentation image.

Voxel counts were generated from the edited segmented image and transformed into tissue volumes. The intraclass correlation coefficient (ICC) (Shrout and Fleiss, 1979) between timepoints 1 and 2 tissue volumes was computed to assess the reliability of segmentation volumes. The ICC is a true reliability coefficient in that it estimates the percentage of each measurement that reflects the underlying entity. Thus, an ICC of 0.95 means that 95% of the variance of the measurement reflects the true variability of the underlying entity and the remaining 5% reflects measurement error. We also computed the coefficient of variation (COV), defined as:

\[
\text{COV}_i = \frac{1}{N} \sum_j \frac{\sigma_{ij}}{\bar{x}_{ij}} \times 100\%
\]

where \(i\) indexes tissue type and \(j\) indexes subjects, \(\bar{x}_{ij}\) is the mean volume of tissue \(i\) for subject \(j\), and \(\sigma_{ij}\) is the standard deviation of tissue \(i\) for subject \(j\). The COV was computed to compare our interscan variability with that from other laboratories.

2.6. Coregistration

In order to quantify the repeatability of tissue spatial distributions between timepoints 1 and 2, we needed to compute the transformation necessary to coregister the segmented images. We used AIR 1.0 (Woods et al., 1992, 1993) to perform coregistrations between timepoint 1 DSE images and...
and timepoint 2 MPRAGE images. Once these transformations had been computed, the segmented images were ‘split’ into separate tissue masks. For example, we defined a white matter mask by setting all voxels classified as white in the edited segmentation image to 255 and all other voxels to 0. We repeated this for all tissue categories. We applied the timepoint 1 DSE to timepoint 2 MPRAGE transformation to each tissue mask, resulting in timepoint 1 tissue masks in the same orientation as the timepoint 2 MPRAGE images. We then applied the inverse of our existing timepoint 2 MPRAGE to DSE transformation (computed prior to segmentation) to each tissue mask, resulting in timepoint 2 tissue masks coregistered to the timepoint 1 tissue masks. Although the timepoint 1 and 2 DSE images could have been registered directly, the resolution was too poor to ensure good results. The high resolution MPRAGE images from each timepoint could also have been registered and combined with both existing MPRAGE to DSE transformations, but that would have combined the errors from three registrations, instead of two. Trilinear interpolation was used for all reslicing.

Due to the interpolation that occurred during transformation, the tissue masks were no longer true binary masks; i.e. voxel intensities other than 0 and 255 were present. Voxels with intermediate intensities in the tissue images represent partially volumed voxels. In our analysis, a mask intensity of 255 represented a 100% volumed voxel, 0 represented a 0% volumed voxel, and all other intensities represented an interpolated scale between these values. Hereafter, we refer to these mask data in terms of percentage volume for each voxel. We computed the tissue overlaps between timepoints 1 and 2 as described below.

2.7. Computation of tissue overlap

The overlap between the timepoint 1 and 2 images was 100% for spatially corresponding voxels that were both fully volumed, and 0% when both voxels had no tissue contribution. The overlap was not well defined when voxels on either image were partially volumed. In such cases, the ‘true’ overlap is dependent on the location of the tissue within each voxel, which is unknown. We estimated overlap using two methods that make different assumptions about the location of tissue in the voxels from the two images. One method underestimates the degree of overlap for partially volumed voxels, and the other overestimates the overlap.

The first method underestimates the amount of true overlap by assuming independence of the distribution of tissues in voxels from the two images. In this case, the proportion of overlap is the product of the tissue proportions in each voxel. In this study, we expected tissues to be located in the same general area of a partially volumed voxel, and believed the assumption of independence resulted in an underestimate of the overlap. The second method assumes total agreement in the spatial distribution of tissues within the corresponding voxels and most likely overestimates the overlap. In this method, the proportion of overlap is defined as the minimum of the tissue proportions in the voxels from the two segmented images.

We then computed the volumes for timepoint 1, timepoint 2, and both overlap tissue images. The overlap was computed as,

$$\text{overlap} = \frac{V_{\text{overlap}}}{0.5(V_{\text{Timepoint 1}} + V_{\text{Timepoint 2}})},$$

where $V$ is the volume of the tissue. Our computed overlap quantifies the volume of voxels that were classified as the same tissue category across subsequent scans, and normalizes by the average volume for that tissue category. If the overlap for gray matter was 0.75, this means that 75% of the voxels were classified as gray matter in both the timepoint 1 and timepoint 2 images.

Our overlap is the same as the similarity measure described by Zijdenbos et al. (1994). We also computed the overlap between the brain phantom discrete anatomical model and the output of our K-means cluster segmentation when applied to the phantom T1-, PD-, and T2-weighted images. In this case, both segmentation images were fully
Table 1
Tissue volumes of gray, white, and CSF computed from the discrete anatomical brain phantom and from the segmentation image computed from the T1-, PD-, and T2-weighted phantom images.

<table>
<thead>
<tr>
<th>Tissue type</th>
<th>Tissue volume from discrete phantom (ml)</th>
<th>Segmentation tissue volumes from phantom MRIs (ml)</th>
<th>% Difference</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gray</td>
<td>902.91</td>
<td>869.22</td>
<td>3.80</td>
<td>0.95</td>
</tr>
<tr>
<td>White</td>
<td>674.78</td>
<td>707.40</td>
<td>-4.72</td>
<td>0.96</td>
</tr>
<tr>
<td>CSF</td>
<td>371.95</td>
<td>363.09</td>
<td>2.41</td>
<td>0.94</td>
</tr>
</tbody>
</table>

*The percentage difference in tissue volumes and the tissue overlaps between the two methods are also shown. Volumes were computed using all 181 slices of the phantom.

Fig. 4. Representative slices from the discrete anatomical brain phantom are shown on the left, and the corresponding slices from the segmentation image generated by the K-means clustering segmentation of the T1-, PD-, and T2-weighted simulated images is shown on the right. CSF is light gray, gray matter is dark gray, and white matter is white. The discrete model also contains a separate class (black) for the glial tissue that lines the ventricles.
3. Results

Table 1 shows the volumes obtained from the discrete anatomical brain phantom for gray matter, white matter, and CSF compared to the volumes output by the segmentation of the T1-, PD-, and T2-weighted images generated by the MRI simulator based on the discrete anatomical brain phantom. The percentage difference between these volumes were all less than 5%, and the overlaps were 0.94 or greater. Fig. 4 shows slices from the discrete anatomical brain phantom and the corresponding slices from the segmentation image. As the figure shows, in addition to gray, white, and CSF tissue classes, the discrete model contains a class for the glial tissue that lines the ventricles. Our three-category K-means cluster segmentation misclassifies the glial tissue, usually as gray matter.

The tissue volumes, ICCs, and COVs from the segmentation of the timepoint 1 and 2 MRIs are shown in Table 2. The reliabilities were greater than 0.92 for the tissue categories defined primarily by the K-means clustering output (white, cortical gray, sulcal and ventricular CSF) and greater than 0.80 for the remaining tissue categories, which depend to a much greater extent on operator editing of the segmented images. The COV values were approximately 6% or less for all tissues except WMSH. Because we examined young, healthy subjects, there were very few WMSHs noted on any subject. One subject had a WMSH tissue volume of 0.8 ml, and the remaining subjects had WMSH volumes of less than 0.2 ml.

Fig. 5 presents tissue overlaps computed assuming spatial independence vs. spatial correspondence of partially volumed voxels. The 'true' overlaps for each tissue are between these two estimates, and probably lie closer to the high estimate because we expect the tissue boundaries in partially volumed voxels to be closely aligned. Computed over the eight tissue categories, overlaps computed assuming spatial independence of partially volumed voxels were smaller by 0.076 ± 0.032 (mean ± S.D.) than overlaps computed assuming spatial correspondence of partially volumed voxels. In the latter case, overlaps ranged from 0.75 to 0.86 for all tissue types except the WMSH, where the overlap was only 0.16. Table 2 shows the overlaps for all eight tissue types when spatial correspondence is assumed. The tissues are listed in order of decreasing volume.

We performed additional experiments where we translated our timepoint 1 tissue masks to

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Volume (mean ± S.D. in ml)</th>
<th>ICC</th>
<th>COV (%)</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cortical gray</td>
<td>388.05 ± 38.85</td>
<td>0.95</td>
<td>1.73</td>
<td>0.83</td>
</tr>
<tr>
<td>White</td>
<td>365.93 ± 43.48</td>
<td>0.92</td>
<td>2.90</td>
<td>0.86</td>
</tr>
<tr>
<td>Sulcal CSF</td>
<td>118.13 ± 31.15</td>
<td>0.96</td>
<td>5.11</td>
<td>0.75</td>
</tr>
<tr>
<td>Ventricular CSF</td>
<td>22.28 ± 9.29</td>
<td>0.99</td>
<td>5.07</td>
<td>0.83</td>
</tr>
<tr>
<td>Lenticular nucleus</td>
<td>13.07 ± 1.90</td>
<td>0.86</td>
<td>5.12</td>
<td>0.86</td>
</tr>
<tr>
<td>Thalamus</td>
<td>12.70 ± 1.97</td>
<td>0.81</td>
<td>6.17</td>
<td>0.85</td>
</tr>
<tr>
<td>Caudate</td>
<td>7.10 ± 0.88</td>
<td>0.80</td>
<td>3.39</td>
<td>0.80</td>
</tr>
<tr>
<td>WMSH</td>
<td>0.16 ± 0.25</td>
<td>0.80</td>
<td>58.94</td>
<td>0.16</td>
</tr>
</tbody>
</table>

*Tissues are ordered from largest to smallest average volume. Volumes were computed using a subset of slices, from the most inferior slice of the subcortical region (where the thalamus is clearly identifiable) through the top of the brain.*
simulate timepoint 2 data, and computed the overlaps. This simulated the condition of perfectly reproducible segmentation, but misregistration of the images. We experimented with different degrees of translation, until we achieved the same overlap as we had with our true timepoint 1 and timepoint 2 data. When we translated the timepoint 1 tissue masks by 0.70 voxels in each direction (misregistration of 1.44 mm), we achieved an average overlap of 0.74, the same as that obtained when we coregistered the timepoint 1 and 2 images. These results illustrate that our overlap results are comparable to perfectly reliable segmentation with coregistration errors less than 2 mm.

4. Discussion

Our K-means clustering segmentation method, which utilizes intensity information from T1-, PD-, and T2-weighted images, performed extremely well on the data generated by MNI using their realistic digital brain phantom and MRI simulator. Our method produces a classification into gray matter, white matter, and CSF, and the tissue volume differences between the segmentation image and ‘truth’ were less than 5% for these three tissues. These tissue volume differences are similar to those for other segmentation algorithms that were validated using the MNI simulator (Reiss et al., 1998; Suckling et al., 1999). More importantly, we achieved overlaps of 0.94 and greater for these three tissues. The overlap reflects both volume and localization agreement between the tissue classifications, and our results compare favorably to those reported by Kollokian (1996). Kollokian used simulated MRI data from MNI to compare seven MRI tissue classification techniques. He compared the algorithms under different levels of noise, varying slice thicknesses, and several levels of RF inhomogeneity. At the 3% noise level, 1-mm slice thickness, or the 20% level of inhomogeneity, the overlaps from the various segmentation algorithms ranged from approximately 0.7 to 0.9. Our overlaps for comparable simulated data were higher than those reported by Kollokian.
When we used the K-means clustering segmentation on serially collected MRIs from volunteers, our volume reliabilities were 0.80 and above for all tissues except the WMSH. Our tissue volumes were more reliable than those reported by Cohen et al. (1992) for gray, white, and CSF, and more reliable than those reported by Byrum et al. (1996) for gray, white, and ventricular CSF. Our gray and white matter reliabilities were slightly lower than the scan–rescan reliabilities reported by Harris et al. (1999), although our CSF reliabilities were comparable. Reiss and colleagues examined the individual percent scan difference in order to assess the stability of their segmentation algorithm on four subjects studied twice. They reported mean percent scan differences for gray matter, white matter, and CSF volumes of 101, 99 and 105% (Reiss et al., 1998). On our 10 subjects, comparable mean percent scan differences for gray matter, white matter, ventricular CSF and sulcal CSF were 100, 100, 101, and 101%. Our COV values were 6% and less for all tissues except WMSH, and are comparable to those reported in the literature (Byrum et al., 1996; Kikinis et al., 1992).

Using K-means clustering segmentation and AIR 1.0 coregistration between the timepoint 1 and timepoint 2 scans, we achieved overlaps of approximately 0.75 or better for all tissue categories except WMSH. The overlaps for the white and cortical gray matter categories were approximately 0.83 or better. Müller-Gärtner and colleagues tested their MRI-based correction of PET for partial volume effects when coregistration errors were introduced (Müller-Gärtner et al., 1992). When they simulated perfect segmentation and coregistration errors of 1–3 MRI voxels (0.94–2.82 mm), they observed only a very small mean error in gray matter activity (0.4 ± 3.0%). When we simulated perfect segmentation but inaccurate coregistration of 0.70 voxels in each direction (1.44 mm), the overlaps were comparable to the scan–rescan overlaps. Based on these results, we conclude that the repeatability of our segmented MRIs is adequate for use in correction of partial volume or tissue contribution effects on functional images. Our results also imply that for all tissue categories except WMSH more than 70% of the voxels included in any tissue category truly belong to that category.

Zijdenbos et al. (1994) used a semiautomatic technique to quantify white matter lesions on the human brain, and achieved overlaps greater than 0.75 for WMSH. However, they achieved their results by repeated segmentation of the same datasets, so they were only evaluating overlap errors due to operator variability. In this study, decreases in overlaps from a perfect 1.0 reflect operator-induced variability, RF inhomogeneity, motion and flow artifacts, drift in imager function over time, PVEs, and errors in coregistration. In addition, the Zijdenbos study chose MRI slices from patients with a clinical diagnosis of Alzheimer’s disease and/or multi-infarct dementia. They required the WMSHs to have a minimum size of approximately 10 voxels (≈ 8 mm³), and the average total area of WMSH per slice was approximately 600 voxels (≈ 500 mm²). In contrast, our subjects were healthy young adults, and the average total area of WMSH per brain (approx. 25 slices) was approximately 100 voxels (≈ 100 mm²). We suspect that many of the WMSHs in our subjects were so small that PVEs dominated the segmentation. In other words, small differences in head positioning during the repeat scan could result in the partial volume of WMSH in a given voxel to decrease to the degree that the voxel was no longer identified as WMSH.

In general, overlaps were better for tissue categories with larger volumes and/or tissue categories that are contiguous with many interior voxels as compared to boundary voxels. We suspect this morphology contributes to high overlaps because these structures will have fewer partially volumed voxels in the transformed tissue masks than a category such as sulcal CSF.

We achieved overlaps of greater than 0.94 when comparing segmented simulated MRIs with the digital phantom, but were only able to achieve overlaps of 0.75–0.86 when comparing the serially collected scans. We believe that our inability to do better is due to coregistration errors between our timepoint 1 and timepoint 2 images. Our simulations showed that small registration errors led to large decreases in overlap. We were registering fairly high resolution images, with resolu-
tions of $1 \times 1 \times 1.5 \text{ mm}$ and $1 \times 1 \times 3 \text{ mm}$. We expect even larger errors to occur when coregistering high resolution MRIs to lower resolution images such as PET, SPECT, or MR spectroscopy, where in-plane resolution can be on the order of $2 \text{ mm}$, and slice thicknesses are typically greater than $6 \text{ mm}$. Therefore, if using a segmented structural MRI to define the caudate on a coregistered PET image, we expect the overlap between the MRI-defined caudate and the ‘true’ caudate on the PET image to have a lower overlap than that reported between the coregistered repeat MRIs. Conversely, we expect coregistration errors to decrease with improved image resolution, with a resultant increase in overlaps.

The results of this article support the use of K-means cluster analysis for segmentation, and the use of segmented structural MRIs to guide the analysis of functional and other images. However, we did not directly compare our segmentation method with any others, such as methods based on prior knowledge, neural networks, or deformable geometric models. It may be that some of these techniques will perform as well or better than our current method, but a similar approach to that used in this article must be taken to evaluate the spatial reproducibility of the methods. Only then can the anatomic segmentation information obtained with these techniques be confidently used in the analysis of functional images.

Acknowledgements

The authors would like to thank Diana Truran, Alanna McAlorum, Rosanna Jeremias, and Dawn Hardin for their assistance in recruiting subjects, running the magnet, and processing the MRIs. This work was supported by NIA grant AG12435, NIAAA grant P01AA11493, NIDA grant R01DA08365 and a DVA Research Career Scientist Award (George Fein).

Appendix A

All structures are marked on images in the axial orientation. Editing begins at the most inferior slice of the subcortical region through the top of the brain.

A.1. Thalamus

Begin marking when the thalamus is fully volumed and all borders are discernible (as with all tissues). This typically occurs 2–3 slices after the disappearance of the red nucleus. The inferior and medial borders are the ambient cistern (sulcal CSF) and the 3rd ventricle, respectively. The superior and lateral border is the posterior limb of the internal capsule.

A.2. Caudate

Begin marking when the head of the caudate nucleus is distinguishable from the putamen. The medial border is at first the cortical gray from the interhemispheric fissure, then progresses to be the frontal horns of the lateral ventricles in the middle slices to the body of the lateral ventricles at the top slices. The lateral border is the anterior limb of the internal capsule.

A.3. Lenticular nuclei

The globus pallidus typically volumes before the putamen fully does and, therefore, begin marking the globus pallidus first. Include the putamen when it, too, becomes fully volumed and a clear separation from the head of the caudate nucleus is discernible. The lateral border must distinguish the putamen from the claustrum. The superior border is the anterior limb of the internal capsule. The medial border is the posterior limb of the internal capsule.

References


