Original Research

Comparison of 10 Brain Tissue Segmentation Methods Using Revisited IBSR Annotations

Sergi Valverde, MS,* Arnau Oliver, PhD, Mariano Cabezas, PhD, Eloy Roura, MS, and Xavier Lladó, PhD

Purpose: Ground-truth annotations from the well-known *Internet Brain Segmentation Repository* (IBSR) datasets consider Sulcal cerebrospinal fluid (SCSF) voxels as gray matter. This can lead to bias when evaluating the performance of tissue segmentation methods. In this work we compare the accuracy of 10 brain tissue segmentation methods analyzing the effects of SCSF ground-truth voxels on accuracy estimations.

Materials and Methods: The set of methods is composed by FAST, SPM5, SPM8, GAMIXTURE, ANN, FCM, KNN, SVPASEG, FANTASM, and PVC. Methods are evaluated using original IBSR ground-truth and ranked by means of their performance on pairwise comparisons using permutation tests. Afterward, the evaluation is repeated using IBSR ground-truth without considering SCSF.

Results: The Dice coefficient of all methods is affected by changes in SCSF annotations, especially on SPM5, SPM8 and FAST. When not considering SCSF voxels, SVPASEG (0.90 ± 0.01) and SPM8 (0.91 ± 0.01) are the methods from our study that appear more suitable for gray matter tissue segmentation, while FAST (0.89 ± 0.02) is the best tool for segmenting white matter tissue.

Conclusion: The performance and the accuracy of methods on IBSR images vary notably when not considering SCSF voxels. The fact that three of the most common methods (FAST, SPM5, and SPM8) report an important change in their accuracy suggest to consider these differences in labeling for new comparative studies.

Key Words: brain MRI; tissue segmentation; permutation tests; IBSR
J. Magn. Reson. Imaging 2014;00:000–000.

MOST AUTOMATIC BRAIN tissue segmentation methods are evaluated on common data which permit to reproduce and compare the results between different studies. In the past 15 years, some public datasets have been proposed based on simulated data such as the Brainweb dataset (1) or real MRI data acquired from different sources such as the Internet Brain Segmentation Repository (IBSR)[1]. IBSR contains two sets of T1-w images acquired from healthy subjects and composed by 20 (IBSR20) and 18 (IBSR18) images, respectively. Both sets provide images and ground-truth segmentations labeled by experts which permit to quantify the accuracy of methods with respect to the images of the dataset (2–5).

However, on both datasets all original ground-truth annotations do not contain sulcal parts of cerebrospinal fluid (CSF) tissue and include it inside the gray matter (GM) segmentation (see Fig. 1). Although this fact is known by authors, most of the studies which make use of the IBSR datasets tend to publish their findings assuming the deviation on CSF tissue.

On all these studies, CSF is not considered (2,6–8), or a weak performance of CSF is obtained compared with GM and white matter (WM) (3,5,9). In contrast, several studies also using IBSR datasets have relabeled sulcal CSF (SCSF) voxels as GM tissue before evaluating the accuracy of methods (10–12) or combined all CSF with GM as a single tissue (13) to minimize the differences between segmentation masks and ground-truth labels.

Changes in CSF tissue labels can not only affect the accuracy of CSF tissue but also the GM estimation. For instance, with original IBSR data, the accuracy of a method segmenting GM tissue can be benefited by misclassified SCSF voxels, as long as they are labeled as GM on ground-truth images. Misclassified SCSF voxels can bias the real accuracy of methods segmenting GM tissue because these voxels are labeled as GM and can compensate the performance on genuine GM tissue. More importantly, on comparative studies using IBSR datasets this fact can induce to inaccurate measurements, and real differences between methods can be hidden behind the bias introduced by SCSF voxels.

In this study, we evaluate the effects of the IBSR annotations on the accuracy of ten commonly used

Department of Computer Architecture and Technology, University of Girona, Girona, (Spain).

Contract grant sponsor: Generalitat de Catalunya; Contract grant number: FI-DGR2013.
Eloy Roura also holds a research grant from Universtitat de Girona (ref UdG BR-GR13).

*Address reprint requests to: S.V., Campus Montilivi, University of Girona, 17071 Girona (Spain). E-mail: svalverde@eia.udg.edu

Received July 23, 2013; Accepted October 22, 2013.
DOI 10.1002/jmri.24517
View this article online at wileyonlinelibrary.com.

[1]Available from the Center for Morphometric Analysis at Massachusetts General Hospital http://www.cma.mgh.harvard.edu/ibsr

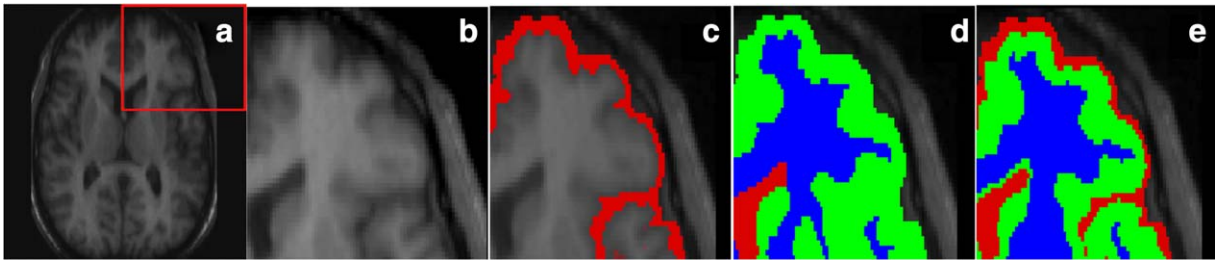

Figure 1. Graphical description of IBSR ground-truth annotations. **A:** Image IBSR01 from IBSR18 dataset. **B:** Zoomed view of part of the image. **C:** Highlighted sulcal CSF voxels. **D:** IBSR ground-truth for the current image. E: Example of tissue segmentation. SVPASEG correctly classifies the majority of sulcal CSF voxels as CSF. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

automatic brain tissue segmentation methods with the aim to establish a reference mark that can be used in future comparisons between these and new proposed methods.

MATERIALS AND METHODS

MRI Image Data

The MRI image data is composed by 20 and 18 T1-w scans of normal subjects from the *Internet Brain Segmentation Repository* (IBSR), available from the Center for Morphometric Analysis at Massachusetts General Hospital. The first set of scans is commonly known in the literature as IBSR20 while the second is known as IBSR18. Although the Center for Morphometric Analysis at Massachusetts General Hospital provide both datasets, the characteristics for each set of images is different.

IBSR20 Dataset

The IBSR20 image set is composed by 20 T1-w scans with 3.1 mm slice thickness (256 $\times$ 63 $\times$ 256). The authors also provide labeled volumes with main tissue annotations for evaluation (GM, WM, and CSF), based on trained experts using a semi-automated intensity contour mapping algorithm (14), and signal intensity histograms. These images are sorted by level of difficulty. The most challenging scans contain important acquisition artifacts and irregularities.

IBSR18 Dataset

The set is composed by 18 T1-w scans with 1.5 mm slice thickness (256 $\times$ 128 $\times$ 256). IBSR18 scans present higher resolution and image quality than IBSR20, with no apparent acquisition artifacts that can bias the accuracy of some scans. Scans are provided after processing them with the Autoseg bias field correction routines from the Center for Morphometric Analysis. The dataset is also supplied with manually labeled volumes into 84 structures obtained using the NVM program[2] and three-class labeled volumes by assigning each of the 84 structures into GM, WM, and CSF tissues.

Preprocessing

First, we remove from the images all nonbrain tissue parts such as eyes, fat, spinal cord or brain skull using a binary mask created from the provided ground-truth. Afterward, all images from IBSR20 dataset are corrected from possible intensity nonuniformities and acquisition artifacts using the N3 package[3] (15). To guarantee a common preprocessing framework between segmentation techniques, we disabled all possible preprocessing options on methods for all experiments.

Segmentation Methods

In this study, we evaluate 10 brain tissue segmentation algorithms on T1-w MRI data with the aim to include a wide set of different segmentation techniques and available tools. We include in our study simpler techniques such as ANN, FCM, KNN, and public available toolboxes such as FAST, SPM5, SPM8, PVC, GAMIXTURE, SVPASEG, and FANTASM. Characteristics and requirements for each method are summarized in Table 1. After testing different parameter configurations for each method, we run all methods with default options because they provide the best overall results for both datasets.

FAST (16) and SVPASEG (8) methods are both based on Markov Random Fields models. In the case of FAST, Markov Random Field parameters are estimated using the iterative Expectation Maximization algorithm with K-means initialization. SVPASEG is based on an Iterative Conditional Modes algorithm and initialization parameters are estimated by a real-coded Genetic Algorithm. PVC (17) is based on a Maximum-a-Posteriori classifier and also optimized by the Iterative Conditional Modes algorithm. Tissue distributions are estimated by combining the tissue measurement model with a spatial prior model of the brain.

SPM5 and SPM8 are two of the available versions of the SPM toolbox. SPM5 *segment* and SPM8 *new_segment* methods are both based on the same algorithm, which comprises the parameter estimations of a Gaussian Mixture Model, atlas registration and bias-field correction at the same time iteratively (18). SPM8 is the current version of the toolbox and introduces

[2]http://neuromorphometrics.org:8080/

[3]http://www.bic.mni.mcgill.ca/ServicesSoftware/MINC

Table 1
Brain Tissue Segmentation Methods Evaluated In This Study *

Method reference Name	Algorithm characteristics				Version	
	Algorithm	Image Type	IC	SS	Implementation	Source
FAST [16]	HRMF+EM	All	Yes	No	FSL v.5.0 (Sept 2012)	http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FAST
SPM5 [18]	GMM	All	Yes	Yes	SPM8 v.4667 (Feb 2012)	www.fil.ion.ucl.ac.uk/spm/
SPM8 [19]	GMM	All	Yes	Yes	SPM8 v.4667 (Feb 2012)	www.fil.ion.ucl.ac.uk/spm/
GAMIXTURE [20]	GA-GMM	T1	No	No	GAMIXTURE v1.1 (Feb 2007)	http://www.loni.ucla.edu/Software/GAMixture
ANN [23]	SOM	T1	No	No	MATLAB 7.12	–
FCM [21]	FCM	T1	No	No	MATLAB 7.12	–
KNN [24]	KNN	T1,PD	No	No	MATLAB 7.12	–
SVPASEG [8]	GA-MRF	T1,T2	No	No	v.2.1 (Oct 2010)	http://www.cs.tut.fi/jupeto/svpaseg/
FANTASM [22]	RFCM	T1	Yes	Yes	MIPAV v.R3c (Mar 2012)	http://mipav.cit.nih.gov/
PVC [17]	MAP	T1	Yes	Yes	Brainsuite 11.a (May 2011)	http://neuroimage.usc.edu/neuro/BrainSuite

*The acronyms for the algorithms stand for: Expectation Maximization (EM), Fuzzy C-Means (FCM), Gaussian Mixture Model (GMM), Genetic Algorithm (GA), Hidden Random Markov Fields (HRMF), Maximum-a-Posteriori (MAP), K-nearest Neighbor (KNN), Markov Random Fields (MRF), Robust Fuzzy C-Means (RFCM), Self Organized Map (SOM). The acronyms for the column definition stand for Intensity Correction (IC) and Skull-stripping (SS).

several differences with respect to SPM5 such as a different treatment of mixing proportions, an improved registration model, an extended set of probabilistic atlases, and a more robust initial registration (19). Similar to SPM approaches, tissue distributions on GAMIXTURE (20) are also obtained by a Gaussian Mixture Model. However, this program estimates both the initial and the successive parameters of the Gaussian Mixture Model by a real-coded Genetic Algorithm.

FCM (21) and FANTASM (22) are based on Fuzzy Clustering techniques. FCM refers to the classical fuzzy-c-means clustering algorithm, which does not take into account spatial information. In contrast, FANTASM extends the FCM approach by modifying the objective function with a penalty term based on the membership of neighbors to other classes that makes the method more robust to noise and acquisition artifacts.

ANN (23) implements a Self Organizing Map or Kohonen network, which clusters image data based on an iterative process of comparison of related changes within voxels organizing unknown data into groups of similar patterns, according to a similarity criterion (e.g., Euclidean distance). Finally, KNN (24) implements a self-trained K-Nearest Neighbor algorithm based on automatic registration of prior probability atlases into the input image, which are used to label training voxels with the class with maximum probability obtained from the tissue atlases.

Data Analysis

First, we evaluate the accuracy of methods segmenting GM, WM, and CSF tissues on both IBSR20 and IBSR18 datasets by computing the overlap between the segmentation masks and the ground-truth annotations using the Dice similarity coefficient (25):

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN}$$

where *TP* is the number of true positives between the segmentation result and the ground-truth, *FP* is the number of false positives and *FN* is the number of false negatives. Hence, the DSC coefficient ranges from 0 to 1, and a higher Dice coefficient represents more accurate segmentation. Compared with other measures such as the Sensitivity, Specificity, False Negative, and False Positive Rates (13,26,27), which only state on positive or negative outcomes and are only based on size, the Dice coefficient is a compromise between the positive and negative outcomes, allowing a better understanding of the overall similarity and dissemination of the location of the differences between the segmentation masks and the ground-truth (28).

Second, we rank methods on both IBSR20 and IBSR18 datasets using significant pairwise method permutation tests of the obtained Dice values on GM and WM tissues (29,30). Permutation tests select small sets of independent values obtained by the different segmentation methods, choose all possible method pairs, and for each pair permute an arbitrary number of times a random number of values. Permutation tests permit to compute the exact *P*-value, and are not limited by any statistical distribution or minimum number of subjects. The *P*-value for each method pair is computed by counting the number of times that the mean difference of the permuted pair is higher than the mean difference without permuting. Finally, the significance of the results between the method pair is stored as the percentage of times where the *P*-value ≤ 0.05.

For our experiments, we have adapted the implementation provided by Klein et al. (30)[4]. The test returns the mean μ and standard deviation σ of the fraction of times when each method provides a higher Dice value than the rest of the methods with significant *P*-value ≤ 0.05. Consequently, methods with higher means have passed a higher number of pairwise comparisons with other methods using randomly chosen subsets of values. Methods are presented in ranks determined by the mean and standard

[4]Available for download in http://www.mindboggle.info/papers/evaluation_NeuroImage2009/

Table 2
Mean Dice Coefficient for Each Method and Tissue Computed From IBSR20 Scans

a. IBSR20 with original ground-truth

Method	GM	WM	CSF
FAST	0.68±0.06	0.79±0.10	0.13±0.04
SPM5	0.76±0.06	0.80±0.04	0.17±0.07
SPM8	0.78±0.06	0.81±0.08	0.21±0.07
GAMIXTURE	0.77±0.09	0.74±0.16	**0.25±0.12**
ANN	0.69±0.09	0.77±0.14	0.15±0.06
FCM	0.69±0.09	0.77±0.14	0.14±0.05
KNN	0.64±0.09	0.80±0.06	0.13±0.04
SVPASEG	**0.82±0.04**	**0.81±0.07**	0.21±0.06
FANTASM	0.70±0.10	0.77±0.14	0.15±0.06
PVC	0.66±0.11	0.63±0.23	0.13±0.05

b. IBSR20 with ground-truth not considering SCSF

Method	GM	WM	CSF
FAST	0.82±0.06	0.78±0.12	0.76±0.12
SPM5	0.86±0.03	**0.82±0.02**	0.82±0.08
SPM8	0.86±0.06	0.81±0.07	**0.83±0.08**
GAMIXTURE	0.83±0.06	0.74±0.16	0.75±0.10
ANN	0.81±0.08	0.77±0.14	0.76±0.12
FCM	0.81±0.08	0.77±0.14	0.76±0.12
KNN	0.78±0.08	0.80±0.06	0.75±0.12
SVPASEG	**0.88±0.04**	0.81±0.07	0.76±0.09
FANTASM	0.81±0.08	0.77±0.14	0.77±0.12
PVC	0.79±0.09	0.62±0.24	0.77±0.12

Table (a) shows the results on IBSR20 scans evaluated with original ground-truths. Table (b) shows the results on the same scans when not considering SCSF on the evaluation. Reported values are mean ± standard deviation. The highest mean Dice value for each tissue is shown in bold text.

deviation of the method with highest mean (μ_o, σ_o). Ranks are decided in terms of the distance of the mean of each method to the best mean observed. Specifically, Rank 1 methods are those in (μ_o–σ_o, μ_o), Rank 2 methods fall in (μ_o–$2\sigma_o$, μ_o–σ_o) and Rank 3 methods are those in the interval (μ_o–$3\sigma_o$, μ_o–$2\sigma_o$). In all our experiments we have repeated the permutation tests within N = 1000.

RESULTS

IBSR20 Dataset

Table 2a shows the Dice overlap coefficients obtained for each method and tissue. SVPASEG provides the highest Dice values on GM, followed by SPM8, GAMIXTURE and SPM5. SVPASEG and both SPM approaches take advantage of the prior atlas information to improve their performance on the most difficult images, which also explains their low standard deviation in comparison with the rest of the methods. GAMIXTURE compensates a lower performance on difficult scans with a similar performance to SVPASEG on easier images. In contrast, FAST, PVC, and KNN provide the lowest Dice values, mostly because their poor performance on CSF. On WM, again SVPASEG but also SPM5 and SPM8 provide the highest overlap coefficients while PVC is the method of the study with the lowest Dice coefficient.

All methods provide very low values for CSF due to divergences on the classification of SCSF tissue with respect to IBSR20 ground-truth. Methods tend to classify SCSF voxels mostly as CSF but those are labeled as GM in the ground-truth. This fact has also a direct effect on GM tissue because voxels segmented as SCSF are considered as false positives by the ground-truth, decreasing the mean Dice GM values on all methods. In contrast, the number of false positives caused by differences between the segmentation masks and the ground-truth appears to be lower on WM tissue.

To avoid possible influences of SCSF tissue on the overall accuracy of the methods, we recomputed again the Dice overlap coefficient on the same 20 images of the IBSR20 dataset but without taking into account SCSF voxels. Table 2b shows the new values for all methods and tissues. All methods improve their performance after not considering SCSF, especially those with lower Dice values on original images. On WM, SPM5, SVPASEG, SPM8, and KNN are the methods with highest Dice values. Contrary to GM, all methods except PVC provide very similar values on easy scans, and differences on the mean WM Dice values are mostly due to the performance of each method on the most challenging scans. GAMIXTURE and PVC provide the lowest Dice values on WM. The voxels affected by image artifacts appear brighter than normal voxels. The inclusion of these voxels into the tissue distributions of methods that incorporate the voxel intensity into the segmentation process such as FCM, FANTASM, ANN, or GAMIXTURE can increase the threshold between GM and WM tissue intensities, and voxels that would have been classified as WM will be misclassified as GM. The results without considering SCSF show that these differences between methods cannot be attributed to CSF because 8 of 10 methods maintain their performance after the modifications introduced on the ground-truth.

Table 3 shows the ranking of methods returned by the permutation tests for the IBSR20 dataset using the original ground-truth and without considering SCSF voxels. With original ground-truth annotations, SVPASEG outperforms significantly the rest of the methods in 90% of the times on GM. Although SVPASEG performs similarly to SPM5 and SPM8 on difficult scans, the fact that SVPASEG provides the highest Dice values on 15 out of 20 scans makes the method invariant to the different permutations. Permuting the Dice values of several scans with other methods will maintain significant real differences between both methods if one of them has most of the highest values. On WM, SVPASEG, SPM8, SPM5, KNN, and FAST are ranked as Rank 1. The results of the permutation test without considering SCSF reveal that the methods more penalized by original images ground-truth are SPM5, SPM8 and FAST. In the case of SPM5 and SPM8, after not considering SCSF the permutation test classifies both methods in Rank 1 along with SVPASEG while in the case of FAST is

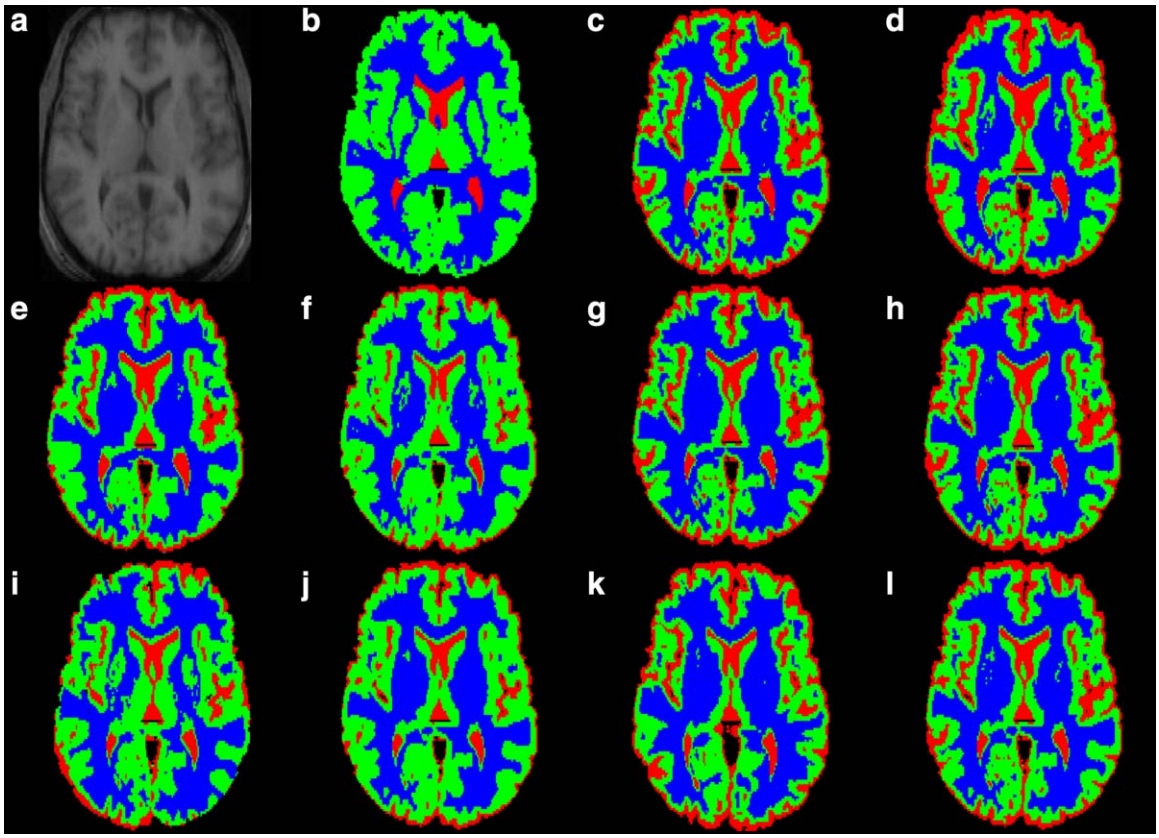

Figure 2. Tissue segmentation masks for IBSR_05 image. Original image (**A**), ground-truth (**B**), and segmentation results from FAST (**C**), SPM5 (**D**), SPM8 (**E**), GAMIXTURE (**F**), ANN (**G**), FCM (**H**), KNN (**I**), SVPASEG (**J**), FANTASM (**K**), and PVC (**L**). In segmentation mask images, CSF tissue is labeled in red, GM in green, and WM in blue.

promoted from Rank 3 to Rank 2. On the contrary, GAMIXTURE and PVC appear to be benefited by the original ground-truth because both methods are moved to Rank 3 after not considering SCSF.

IBSR18 Dataset

Table 4a shows the mean Dice values obtained for all methods and tissues with original ground-truth annotations (see also Fig. 2). In general, better results are

Table 3
Permutation Tests for Obtained Dice Overlap Coefficients on the IBSR20 Dataset With Original Ground-Truth and Not Considering SCSF Voxels *

	IBSR20 (original ground-truth)				IBSR20 (not considering SCSF)			
	GM		WM		GM		WM	
	Method	$\mu\pm\sigma$	Method	$\mu\pm\sigma$	Method	$\mu\pm\sigma$	Method	$\mu\pm\sigma$
Rank 1	SVPASEG	0.90±0.32	SVPASEG	0.70±0.48	SVPASEG	0.80±0.42	SVPASEG	0.64±0.48
			SPM8	0.60±0.52	SPM5	0.70±0.48	SPM8	0.60±0.52
			SPM5	0.50±0.53	SPM8	0.60±0.70	SPM5	0.50±0.53
			KNN	0.40±0.70			KNN	0.46±0.58
			FAST	0.30±0.82			FAST	0.30±0.82
Rank 2	SPM5	0.50±0.71			GAMIXTURE	0.30±0.95	ANN	−0.30±0.82
	SPM8	0.50±0.71			FAST	0.10±0.99	FANTASM	−0.30±0.82
	GAMIXTURE	0.50±0.71					FCM	−0.30±0.82
Rank 3	FANTASM	0.03±0.95	ANN	−0.30±0.82	FANTASM	−0.19±0.92	GAMIXTURE	−0.70±0.67
	ANN	−0.03±0.94	FANTASM	−0.30±0.82	ANN	−0.50±0.71	PVC	−0.90±0.32
	FAST	−0.50±0.70	FCM	−0.30±0.82	FCM	−0.50±0.71		
	FCM	−0.50±0.71	GAMIXTURE	−0.70±0.67	PVC	−0.51±0.52		
	PVC	−0.60±0.52	PVC	−0.91±0.32	KNN	−0.80±0.42		
	KNN	−0.80±0.42						

*Reported values are mean and standard deviation (μ, σ) of the fraction of times when each method produced significant *P* values. Positive values indicate that on average, the method overperformed the other methods in pair-wise significant tests. Negative values indicate the contrary. Rank 1: $(\mu_o - \sigma_o, \mu_o]$, Rank 2: $(\mu_o - 2\sigma_o, \mu_o - \sigma]$, Rank 3 $(\mu_o - 3\sigma_o, \mu_o - 2\sigma_o]$.

Table 4
Mean Dice Coefficient for Each Method and Tissue Computed From IBSR18 Scans

a. IBSR18 with original ground-truth

Method	GM	WM	CSF
FAST	0.74±0.04	**0.89±0.02**	0.12±0.05
SPM5	0.68±0.07	0.86±0.02	0.10±0.05
SPM8	0.81±0.02	0.88±0.01	**0.17±0.08**
GAMIXTURE	0.78±0.08	0.87±0.02	0.15±0.09
ANN	0.70±0.07	0.87±0.03	0.11±0.06
FCM	0.70±0.06	0.88±0.03	0.11±0.06
KNN	0.79±0.03	0.86±0.03	0.16±0.07
SVPASEG	**0.81±0.03**	0.86±0.02	0.16±0.07
FANTASM	0.71±0.06	0.88±0.03	0.11±0.06
PVC	0.70±0.08	0.83±0.07	0.13±0.06

b. IBSR18 with ground-truth not considering SCSF

Method	GM	WM	CSF
FAST	0.88±0.01	**0.89±0.02**	0.47±0.18
SPM5	0.89±0.02	0.87±0.02	**0.79±0.08**
SPM8	**0.91±0.01**	0.88±0.01	0.77±0.08
GAMIXTURE	0.89±0.03	0.87±0.02	0.52±0.15
ANN	0.87±0.03	0.88±0.03	0.52±0.15
FCM	0.88±0.02	0.88±0.03	0.52±0.15
KNN	0.87±0.03	0.86±0.03	0.46±0.16
SVPASEG	0.90±0.01	0.87±0.02	0.57±0.13
FANTASM	0.88±0.02	0.88±0.03	0.53±0.15
PVC	0.83±0.08	0.84±0.07	0.52±0.15

Table (a) shows the results on IBSR18 scans evaluated with original ground-truths. Table (b) shows the results on the same scans when not considering SCSF on the evaluation. Reported values are mean ± standard deviation. The highest mean Dice value for each tissue is shown in bold text.

obtained compared with IBSR20 images, due to the higher spatial resolution and the better image quality of IBSR18 images. SVPASEG and SPM8 provide the highest Dice values on GM, followed by KNN and GAMIXTURE. Again, it appears that SVPASEG and SPM8 take advantage of the prior atlas information to outperform the other methods. Surprisingly, SPM5 provides the lowest Dice value on original IBSR18 images. Unexpected low values for SPM5 on GM are caused by two factors: first, the provided high standard deviation suggests that the mean Dice value of the method has been affected by a low performance on punctual images. Second, although both SPM5 and SPM8 use a probabilistic atlas to guide the segmentation, the atlas is different among the two versions. The low Dice values yield by SPM5 appear to be caused by the own SPM5 atlas. The probability of SCSF voxels to pertain to CSF determines the amount of SCSF voxels that will be classified as CSF. Compared with the SPM8, this probability appears higher in the SPM5 atlas and most of the SCSF voxels are classified as CSF by the method.

On experiments without considering SCSF (see Table 4b), SPM5 provides values similar to the best methods on GM. FAST is also affected by this aspect, and Dice values on images without considering SCSF are notably higher than those evaluated with original ground-truth. On WM, FAST is the method that provides the highest accuracy, followed by SPM8, FANTASM and FCM. In contrast, we observe that the performance of SVPASEG on WM with this second dataset is lower than previous results with IBSR20 images.

The ranking of methods returned by the permutation tests for the IBSR18 dataset with original and modified ground-truths is shown in Table 5. On images with original ground-truth, SVPASEG is the best ranked method on GM tissue, followed by SPM8, GAMIXTURE and KNN. The rest of methods are assigned to Rank 3. On WM, FAST is the only method that is ranked in the first group, while FANTASM is ranked in the second group and the rest of methods in Rank 3. As we have seen on IBSR20 images, the permutation tests on both tissues are again influenced by not considering SCSF. Thus, on GM FAST is now classified in Rank 1, while FANTASM and FCM are assigned to Rank 2. On WM, SPM8 is now assigned to Rank 2 in detriment of FANTASM, which is assigned to Rank 3.

DISCUSSION

In the literature, there are numerous studies comparing the accuracy of their proposed methods with some of the methods of our study (31–33). In fact, we have reviewed other studies which also used IBSR20 and IBSR18 datasets to compare their results with our findings. These previous studies have used original IBSR ground-truths without extracting SCSF voxels.

On IBSR20 data, our results are similar to those reported in other studies for FAST (6,9), SPM5 (6), SPM8 (9), and SVPASEG (31). Our results for FCM and FAST are slightly higher than those reported also by Shahvaran et al (31) while lower for PVC than those published by Shattuck et al (17). The fact that we are observing differences in the accuracy for the same method between studies (FAST [6,9,31]) can be caused by changes in the preprocessing pipeline, initialization parameters or changes in the skull-stripping masks. On IBSR18 data, our findings are also similar to those published in other studies for FAST (32), SPM8 (32), FANTASM (33), GAMIXTURE (32), and SVPASEG (8).

On images evaluated with original ground-truths, the accuracy of methods is in general lower on IBSR20 than IBSR18 images, especially on WM. This fact is explained by the acquisition artifacts found in several images from IBSR20 that have a direct effect on WM tissue distributions, reducing the mean value of the methods, and increasing their variability. Furthermore, the lack of SCSF labeling on ground-truth masks appears to have a weak impact on WM tissue because the improvement on the performance of methods after not considering SCSF is inappreciable. In contrast, we have found that differences in SCSF have a direct impact on GM tissue accuracy. All methods tend to segment SCSF tissue as CSF which decreases the Dice values for both GM and CSF tissues.

Table 5
Permutation Tests for Obtained Dice Overlap Coefficients on the IBSR18 Dataset With Original Ground-Truth and Not Considering SCSF Voxels*

	IBSR18 (Original ground-truth)				IBSR18 (not considering SCSF)			
	GM		WM		GM		WM	
	Method	$\mu\pm\sigma$	*Method*	$\mu\pm\sigma$	*Method*	$\mu\pm\sigma$	*Method*	$\mu\pm\sigma$
Rank 1	SVPASEG	0.80±0.42	FAST	0.90±0.32	SPM8	0.70±0.48	FAST	0.90±0.32
	SPM8	0.70±0.48			SVPASEG	0.60±0.52		
	GAMIXTURE	0.50±0.71			FAST	0.40±0.70		
	KNN	0.40±0.84			GAMIXTURE	0.32±0.47		
Rank 2			FANTASM	0.32±0.65	FANTASM	0.10±0.74	SPM8	0.38±0.68
					FCM	−0.22±0.79		
Rank 3	FAST	−0.10±0.88	SPM8	0.25±0.63	KNN	−0.29±0.49	FCM	0.25±0.63
	FANTASM	−0.11±0.86	FCM	0.19±0.57	SPM5	−0.32±0.64	FANTASM	0.24±0.63
	PVC	−0.40±0.52	ANN	0.11±0.56	ANN	−0.50±0.71	ANN	0.24±0.63
	SPM5	−0.59±0.51	GAMIXTURE	−0.13±0.56	PVC	−0.79±0.41	SPM5	−0.20±0.42
	FCM	−0.60±0.52	SVPASEG	−0.27±0.40			KNN	−0.33±0.40
	ANN	−0.60±0.52	KNN	−0.28±0.40			GAMIXTURE	−0.38±0.68
			SPM5	−0.49±0.52			SVPASEG	−0.50±0.53
			PVC	−0.60±0.52			PVC	−0.60±0.52

*Reported values are mean and standard deviation (μ, σ) of the fraction of times when each method produced significant *P* values. Positive values indicate that on average, the method over-performed the other methods in pair-wise significant tests. Negative values indicate the contrary. Rank 1: ($\mu_o - \sigma_o$, μ_o], Rank 2: ($\mu_o - 2\sigma_o$, $\mu_o - \sigma_o$], Rank 3 ($\mu_o - 3\sigma_o$, $\mu_o - 2\sigma_o$]. (as table 4).

The majority of the reviewed studies use the Dice coefficient to evaluate the accuracy of methods segmenting GM, WM, and CSF tissues. However, other measures such as the Sensitivity and Specificity, False Positive Fraction, and False Negative Fraction (13,26,27) can be used. To compare our results with these coefficients, we re-computed the accuracy of all ten methods using the Sensitivity and Specificity measures. These coefficients are inversely related with the FNR and FPR, respectively, and therefore the obtained results can be directly extrapolated to these measure rates. These new measurements showed that, in general, all methods tended to penalize one of these two coefficients in detriment of the other. Therefore, the rank of methods was clearly distinct, because the measures were only focused on positive or negative outcomes. For instance, PVC was the method from our study that clearly most overestimated WM tissue. Based on the Sensitivity coefficient, we found that PVC was the best ranked method on WM due to the low number of False Negatives. Conversely, it was the worst ranked method when the Specificity was evaluated. In contrast, as stated by the majority of the reviewed studies and our experiments, the Dice coefficient allows a better understanding of the overlap between the segmentation and the ground-truth masks because takes into account both positive and negative outcomes.

The permutation tests return the fraction of times that the mean Dice value of a current method is higher than the rest of the methods in pair-wise significant tests. We have observed that even the ranking of methods returned by each permutation test follows in general the same order produced by simply sorting the mean Dice values, the permutation test allows to disseminate better the differences among methods. Permutation test breaks the linearity of a ranking based on sorting the mean Dice values by differentiating methods by the relevance of their results (30). Furthermore, a low performance of a method on one or several images of the dataset can decrease notably the mean Dice value and increase the standard deviation, while the permutation test tends to minimize the effect on the images with low performance.

Our results suggest that evaluating the accuracy of the 10 methods on original images of both IBSR datasets introduce an artificial bias, because most of the methods are penalized by the lack of SCSF on IBSR ground-truth. Thus, if we analyze the results on images where SCSF is not considered on the evaluation, SVPASEG is the best ranked method on both GM and WM tissues of IBSR20 images. SVPASEG, SPM5 and SPM8 are ranked in the first group on both tissues with significant mean values. On IBSR18 images, the results are not so clear and change between tissues. On GM, SPM8 and SVPASEG are the best ranked methods while FAST is the only method that is assigned to Rank 1 on WM. Both SPM methods on IBSR20, and FAST on IBSR18 are assigned to Rank 1 after removing SCSF from the evaluation. This fact is especially interesting, because these three methods are the most common used as baseline on comparative studies.

Furthermore, comparing the accuracy of the methods on both datasets can give us a better idea of the robustness of methods to different acquisition artifacts or their independence to changes produced by intensity correction. We observe that on GM, the performance of FAST, SPM5, GAMIXTURE, FCM, and FANTASM is dependent of the characteristics of the dataset. In contrast, ANN, PVC and KNN obtain the lowest performance on both datasets. Interestingly, SPM8 and SVPASEG are the only methods that are grouped to Rank 1 on both datasets. On

WM, FAST is the only method that is assigned to Rank 1 on both datasets, while the rest of methods present a different performance on each dataset. In our opinion, the fact that SVPASEG, SPM8 and FAST have been assigned to Rank 1 in three out of four permutation tests performed without considering SCSF, make these methods suitable for accurate brain tissue measurements.

Most of the methods were sensible to changes in acquisition sequences, intensity inhomogeneities or special attributes of the different datasets. In our opinion, the results of this paper highlight the fact that the brain tissue segmentation problem is still open, because there is not a single method that achieves a very high accuracy on all brain tissues. Although the design of more accurate methods should be the most common choice to follow on future research, it would be also interesting to analyze other alternative frameworks such as fusion processes based on the best segmentation results on each tissue to reduce the inner limitations of each individual brain tissue segmentation method.

The most important limitation of our study is the lack of new IBSR ground-truths that incorporate SCSF voxels as CSF tissue. However, given the limitation on this aspect, we propose to overpass the artificial bias introduced by the IBSR ground-truth SCSF voxel labels, by not considering these voxels into the accuracy measurements.

In conclusion, changes on original ground-truth annotations of IBSR images should be taken into account, especially in comparative studies that include several automatic segmentation methods. On these images, SCSF voxels are labeled as GM, and the inclusion of these voxels in the accuracy measurements can bias the results, due to differences in the amount of SCSF tissue classified as CSF by each method.

ACKNOWLEDGMENT

S.V. was funded by a FI-DGR2013 grant from the Generalitat de Catalunya. E.R was funded by a grant UdG BR-GR13 from the Universitat de Girona.

REFERENCES

1. Cocosco CA, Kollokian V, Kwan RKS, Pike GB, Evans AC. Brainweb: online interface to a 3D MRI simulated brain database. Neuroimage 1997;5:425.
2. Awate SP, Tasdizen T, Foster N, Whitaker RT. Adaptive markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification. Med Image Anal 2006;10:726–739.
3. Akselrod-Ballin A, Galun M, Gomori M, Basri R, Brandt A. Atlas guided identification of brain structures by combining 3D segmentation and SVM classification. In: Proceedings of the 9th International Conference on Medical Image Computing and Computer-Assisted Intervention 2006;4191:209–216.
4. Caldairou B, Rousseau F, Passat N, Habas P, Studholme C, Heinrich C. A non-local fuzzy segmentation method: Application to brain MRI. Comput Anal Images Patterns 2009;5702:606–613.
5. Wels M, Zheng Y, Huber M, Hornegger J, Comaniciu D. A discriminative model-constrained EM approach to 3D MRI brain tissue classification and intensity non-uniformity correction. Phys Med Biol 2011;56:3269.
6. Tsang O, Gholipou A, Kehtarnavaz N, Gopinath K, Briggs R, Panahi I. Comparison of tissue segmentation algorithms in neuroimage analysis software tools. In: Proceedings of the 30th International Conference of the IEEE in Engineering in Medicine and Biology Society 2008;3924–3928.
7. Kasiri K, Kazemi K, Dehghani M, Helfroush M. Atlas-based segmentation of brain MR images using least square support vector machines. In: Proceedings of the 2nd International Conference on Image Processing Theory Tools and Applications 2010;306–310.
8. Tohka J, Dinov ID, Shattuck DW, Toga AW. Brain MRI tissue classification based on local markov random fields. Magn Reson Imaging 2010;28:557–573.
9. Tian G, Xia Y, Zhang Y, Feng D. Hybrid genetic and variational expectation-maximization algorithm for gaussian-mixture-model-based brain MR image segmentation. IEEE Transactions on Information Technol Biomed 2011;15:373–380.
10. Bazin PL, Pham D. Topology-preserving tissue classification of magnetic resonance brain images. IEEE Trans Med Imaging 2007;26:487–496.
11. Shiee N, Bazin PL, Ozturk A, Reich DS, Calabresi PA, Pham DL. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. Neuroimage 2010;49:1524–1535.
12. Roy S, Carass A, Bazin PL, Resnick S, Prince JL. Consistent segmentation using a rician classifier. Med Image Anal 2012;16:524–535.
13. Ortiz AR, Gorriz JM, Ramirez J, Salas-Gonzalez D, Llamas-Elvira JM. Two fully-unsupervised methods for MR brain image segmentation using som-based strategies. Appl Soft Comput 2013;13:2668–2682.
14. Filipek PA, Richelme C, Kennedy DN, Caviness VS. The young adult human brain: an MRI-based morphometric analysis. Cereb Cortex 1994;4:344–360.
15. Sled JG, Zijdenbos AP, Evans CP. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 1998;17:87–97.
16. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 2001;20:45–57.
17. Shattuck DW, Sandor-Leahy DR, Schaper K, Rottenberg DA, Leahy RM. Magnetic resonance image tissue classification using a partial volume model. Neuroimage 2001;13:856–876.
18. Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005;26:839–851.
19. Ashburner J, Barnes G, Chen CC, et al. SPM8 Manual. Wellcome Trust Centre for Neuroimaging Institute of Neurology, UCL 2011.
20. Tohka J, Krestyannikov E, Dinov ID, et al. Genetic algorithms for finite mixture model based voxel classification in neuroimaging. IEEE Trans Med Imaging 2007;26:696–711.
21. Pham DL. Spatial models for fuzzy clustering. Computer Vision and Image Understanding 2001;84:285–297.
22. Pham DL. Robust fuzzy segmentation of magnetic resonance images. In: Proceedings of the 14th IEEE Symposium on Computer-Based Medical Systems, Bethesda, Maryland, 2001. p 127–131.
23. Tian D, Fan L. A brain MR images segmentation method based on SOM neural network. Bioinform Biomed Eng 2007;2:686–689.
24. De Boer R, Vrooman HA, Van Der Lijn F, et al. White matter lesion extension to automatic brain tissue segmentation on MRI. Neuroimage 2009;45:1151–1161.
25. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.
26. Anbeek P, Vincken KL, Van Osch M, Bisschops R, Van der Grond J. Probabilistic segmentation of white matter lesions in MR imaging. Neuroimage 2004;21:1037–1044.
27. Vrooman HA, Cocosco CA, van der Lijn F, et al. Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification. Neuroimage 2007;37:71–81.
28. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging 1994;13:716–724.
29. Menke J, Martinez TR. Using permutations instead of student's t-distribution for p-values in paired-difference algorithm comparisons. Proceedings of the IEEE Joint Conference on Neural Networks, Budapest, 2004. p 1331–1335.

30. Klein A, Andersson A, Ardekani BA, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 2009;46:786–802.
31. Shahvaran Z, Kazemi K, Helfroush HS, Jafarian N, Noorizadeh N. Variational level set combined with markov random field modeling for simultaneous intensity non-uniformity correction and segmentation of MR images. J Neurosci Methods 2012; 209:280–289.
32. Zhang T, Xia Y, Feng D. Clonal selection algorithm for gaussian mixture model based segmentation of 3D brain MR images. In: Proceedings of the Second Sino-foreign-interchange Conference on Intelligent Science and Intelligent Data Engineering, Xi'an, China, 2011. p 295–302.
33. Nguyen TM, Wu QM. Robust student's-t mixture model with spatial constraints and its application in medical image segmentation. IEEE Trans Med Imaging 2012;31:103–116.