

Brain Segmentation Performance using T1-weighted Images versus T1 Maps

Xiaoxing Li and Christopher Wyatt

Department of Electrical and Computer Engineering, Virginia Tech,
Blacksburg, VA, 24061

ABSTRACT

The recent driven equilibrium single-pulse observation of T1 (DESPOT1) approach permits real-time clinical acquisition of large-volume and high-isotropic-resolution T1 mapping of MR tissue parameters with improved uniformity. It is assumed that the quantitative nature of maps will facilitate clinical applications such as disease diagnosis and comparison across subjects. However, there is not yet enough quantitative evidence on the actual benefit of adopting T1 maps, especially in computer-aided medical image analysis tasks. In this study, we compare methods with respect to image types, T1-weighted images or T1 maps, in automatic brain MRI segmentation. Our experimental results demonstrate that, using T1 maps, different segmentation algorithms show better agreement with each other, compared to that from using T1-weighted images. Furthermore, through multi-dimensional-scaling projection, we are able to visualize the relative affinity among segmentation results, which reveals that the projections of those segmentations using two different types of input images tend to form two separate clusters. Finally, by comparing to expert segmented reference segmentation of brain sub-regions, our results clearly indicate a better agreement between the manual reference and those automatic ones on T1 maps. In other words, our study provides an evidence for the hypothesis that compared to the conventionally used T1-weighted images, T1 maps lead to improved reliability in automatic brain MRI segmentation task.

Keywords: MRI, T1-weighted, T1 map, segmentation

1. INTRODUCTION

Conventionally, T1-weighted Magnetic Resonance image (MRI) is the most popular image type used in single-channel brain image segmentation, compared with T2-weighted and Proton Density (PD) images, due mainly to its relatively higher contrast between gray matter and white matter.^{1,2} More recently, there has been a rapid development of T1 and T2 mapping techniques,³⁻⁷ in response to the gradually increasing interest on standardized imaging. With these methods, the absolute determination of T1 and T2 ensures intensity consistency within the same type of tissue of an individual image and across multiple images. The high-resolution and rapid computation provided by modern T1 mapping techniques and the intensity consistency provided by T1 map render it a new and straightforward candidate data type in brain image segmentation. New segmentation frameworks either directly take T1 maps as input² or combine them with T1-weighted images in multi-channel image segmentation.⁸ Several potential benefits have been projected for using T1 maps in brain segmentation. First, being parametric images of pure T1, T1 maps are expected to contain reduced bias and other distortions, compared with conventional T1-weighted images, and are believed to better assist tissue discrimination and thus improve the performance of segmentation.⁴ Second, the quantitative tissue characteristic provided in T1 maps can serve as indicators of pathology. For example, studies have uncovered pathological changes in T1 maps within the “normal-appearing white matter” in T1-weighted images,⁹ which is useful in studying the progression of multiple sclerosis (MS). This ability of revealing undergoing pathology in T1 maps is also believed to facilitate segmentation.^{1,2}

We conducted a direct and quantitative comparison of segmentation performance when the same segmentation tools are applied to T1-weighted images and T1 maps. Specifically, in this study, a conventional T1-weighted image (IR-SPRG) and SPRG images at two flip angles are simultaneously collected for a group of patients. The T1 maps are later generated from the SPRG images at the two flip angles using the *driven equilibrium single-pulse observation of T1* (DESPOT1) technique.⁴ These T1 maps and T1-weighted images are then segmented into white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) using a collection of segmentation

Further author information: E-mail: {xxli,clwyatt}@vt.edu , Telephone: 1 (540) 231-6658.

algorithms, giving us a set of 3-class label maps as the segmentation results. We adopt statistical tools to quantitatively analyze the quality of these segmentation results. By comparing these results based on T1 maps and T1-weighted images, we can obtain a systematic evaluation on whether T1 maps facilitate segmentation.

Methods determining the quality of segmentation results have been studied in the past. Ideally, the correctness of a segmentation result can be accurately determined by comparing to the ground truth. Unfortunately, such ground truth is typically not available in patient studies. Instead, some works take phantoms¹⁰ or human expert segmentations¹¹ as the ground truth. When phantoms are used,¹⁰ the ground truth is known exactly, but the synthetic data cannot perfectly mimic the true shape variation, all types of pathology, or the realistic image acquisition. Human expert segmentation results are regarded as the ground truth in many studies,¹¹ although the expert segmentation results are prone to performance variations of the segmentors.¹² More importantly, in most studies, expert segmentation is hard to acquire because of the huge amount of manual work involved in segmenting high-resolution image volumes. As a more general approach, Bouix et. al.¹³ proposed to adopt *common agreement* among results of different segmentation algorithms as an indicator of their relative qualities, where the label map that best agrees with others is expected to be more reliable. In this work, STAPLE¹² algorithm is used to estimate a reference, which is the best agreement among a set of label maps obtained for the same brain. The *Jaccard coefficient* (JC) is taken as the generic similarity measure between the reference and individual segmentation results. Thus, the label map yielding a larger JC is regarded as more accurate. Multi-dimensional scaling (MDS) is another tool used in¹³ to investigate the approximate affinity among different label maps through a 2D visualization and will also be used in our study.

In this paper, we inherit the notion of common agreement to evaluate segmentation qualities using T1-weighted images versus T1 maps. Our work differs from the work by Bouix et. al.¹³ in the sense that, instead of evaluating a set of segmentation algorithms, we evaluate two groups of segmentation results using the same set of segmentation algorithms, but on different input images. As will be shown later, the MDS projection reveals that segmentation results obtained on the same input image tend to form a group. Note that STAPLE assumes uni-model distribution of the inputs, and thus it becomes less reliable to use STAPLE as the reference in evaluating the relative quality of segmentation results obtained on different types of input images. To deal with this, we only use STPALE to estimate the agreement among segmentation results that are obtained on the same input. In our experiments, we find that a smaller cross-segmentation-algorithm variation is observed by segmenting T1 maps, which indicates that there is less tissue characteristic ambiguity carried by T1 maps. Furthermore, we invited a neurologist to manually segment a set of brain sub-regions under the setting that both T1 maps and T1-weighted images are presented. This manual segmentation is then taken as the ground truth, and its agreement with the automatic segmentation algorithms is computed. Our results demonstrate that compared to T1-weighted images, automatic segmentation algorithms have coherently better performances by segmenting T1 maps.

2. SEGMENTATION ON T1-WEIGHTED IMAGES AND T1 MAPS

In our study, brain MR images are collected for a group of $n = 16$ subjects. For each subject, a conventional T1-weighted image (IR-SPRG) and SPRG images at two flip angles are simultaneously collected, while T1 maps (DESPOT1)⁴ are later generated using the SPRG images at the two flip angles. The acquisition time for a T1-weighted image (IR-SPGR) is 7'28" and that for the DESPOT1 data at two flip angles is 6'54". In our experiments, the reconstruction of T1 maps are conducted off-line. However, the technique does permit real-time reconstruction. For each subject, the resulting T1 maps are co-aligned with their corresponding T1-weighted images. An exemplar set of sagittal, coronal and axial view of a pair of T1-weighted image and T1 map is shown in Fig. 1.

2.1 Segmentation algorithms

In this study, we experiment on a collection of popular brain MRI segmentation algorithms. FMRIB's Automated Segmentation Tool (FAST)¹⁴ is distributed as a part of FSL package.¹⁵ This algorithm is based on a hidden Markov random field model, where an expectation-maximization algorithm is used to estimate a bias field, an image classification and a set of class model parameters. EM¹⁶ segmentator is part of the 3D slicer package,¹⁷ which is an expectation-maximization based segmentation algorithm that works with registered tissue probability maps. Statistical Parametric Mapping (SPM)¹⁰ is a MATLAB based software package,¹⁸ which

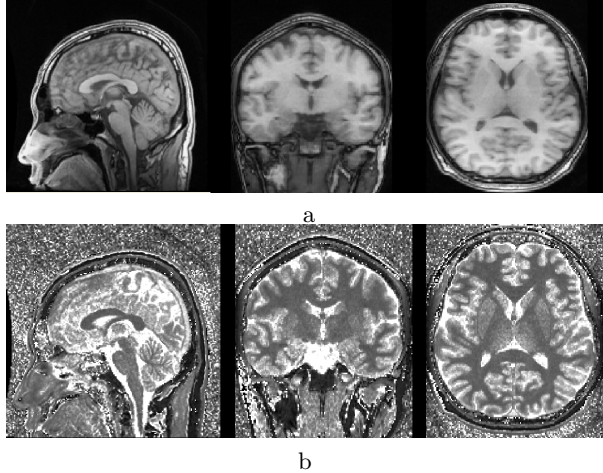


Figure 1. An exemplar set of sagittal, coronal and axial view of a) T1-weighted image and b) T1 map of the same patient.

uses a probabilistic framework that enables image registration, tissue classification and bias correction to be combined within the same generic model. These algorithms use different mathematical techniques with different complexities. In addition to these advanced image segmentation tools, we also include thresholding (THRE) and K-means clustering (KMS) to segment the brains based solely on image intensities. Note that, in general THRE and KMS are not regarded as trustworthy segmentation algorithms due to their model simplicity. The reason of including these algorithms is to highlight the benefit from their quantitative nature in segmenting T1 maps. We emphasize that the purpose of this study is to investigate the segmentation performance variations with respect to different types of input images, i.e., T1-weighted images versus T1 maps. Therefore, the particular choice of segmentation algorithms is not critical, as long as the segmentation is conducted without systematic bias across subjects. The performance evaluation of different segmentation algorithms is not the focus of this work, but can be found in the work by Bouix et. al.¹³

2.2 Segmentation process

After data collection, both T1-weighted images and T1 maps are skull-stripped using FSL¹⁹ tool (BET), followed by minor manual correction. Since we used two flip angles to generate the T1 maps, individually with lower signal-to-noise ratio (SNR) than the T1-weighted images, the SNR of the T1 maps is less than that of the T1-weighted images. We thus used the curvature flow smoothing Insight Toolkit (ITK) filter to pre-process all T1 maps. The segmentation processes are summarized as follows:

- **THRE:** All the T1-weighted images are histogram-matched to the first subject, followed by a bias correction using the MRI Bias-field Correction Filter.²⁰ The bias-corrected T1-weighted images and T1 maps of the first subject are then segmented using manually selected intensity thresholds. These thresholds are then taken as the trained values to segment images of the remaining subjects.
- **KMS:** All the T1-weighted images are histogram-matched and bias-corrected as in THRE. The resulting T1-weighted images and T1 maps are segmented using the standard K-means clustering method.
- **FAST:** In this algorithm, image smoothing and bias correction are included as part of the segmentation pipeline. The skull-stripped images are directly used as inputs.
- **SPM:** The pre-processing in the application includes bias correction, so the T1-weighted images and T1 maps are directly taken as inputs. The probabilistic tissue maps are distributed within the software package.
- **EM:** All the T1-weighted images are histogram-matched and bias-corrected as in THRE. Both the resulting T1-weighted images and T1 maps are segmented using initialization parameters collected from manually-placed samples. For the young subjects in the dataset, we use the aforementioned SPM tissue maps. For the aging subjects, we constructed our own tissue maps by aligning and segmenting 10 aging subjects from a separate aging brain dataset collected in a previous study.

To ensure the consistency of the segmentations, all the free parameters from these algorithms, which are not mentioned above, are set to the default values and used across all images. Note that, there are no available parameters set for T1 maps in the current FAST algorithm and none of the provided parameters can successfully segment T1 maps into 3 meaningful segments. As a result, for subject i , we obtain 9 label maps, 5 of which are the segmentation results of the 5 algorithms using T1-weighted images as inputs, i.e., EM_w^i , $FAST_w^i$, KMS_w^i , $THRE_w^i$ and SPM_w^i . The remaining 4 label maps are the segmentation results on T1 maps for the 4 algorithms other than FAST, i.e., EM_m^i , KMS_m^i , $THRE_m^i$ and SPM_m^i . Fig. 2 shows the segmentation results of one subject, where the segmentation of WM, GM and CSF are marked by blue, green and red, respectively, and is alpha blended into the input images.

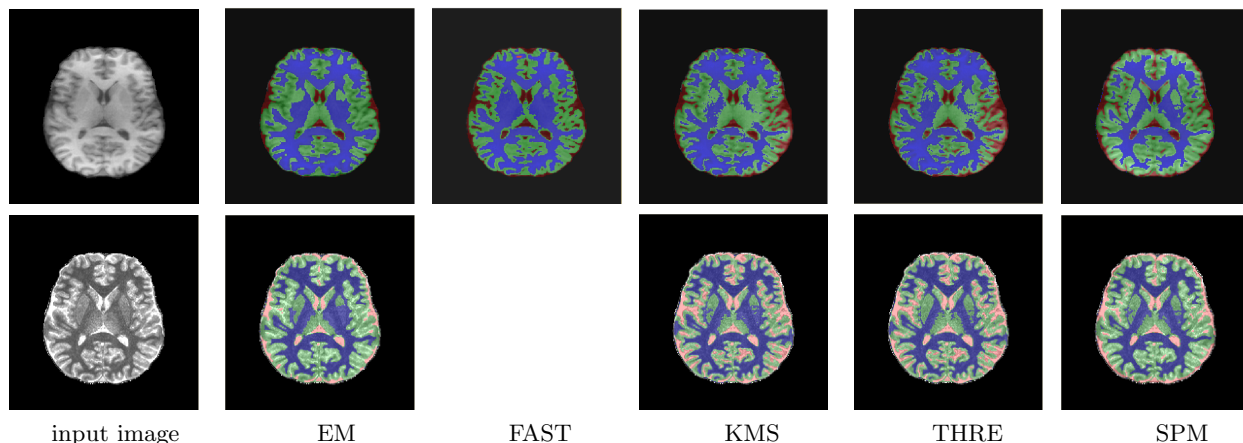


Figure 2. Example of segmentation results. Top row: segmentation obtained using T1-weighted images. Lower row: segmentation obtained using T1 maps.

3. SEGMENTATION QUALITY ANALYSIS

In this section, we compare the segmentation results obtained on T1-weighted images and T1 maps from several different aspects. In order to evaluate the quality of a segmentation result, certain metrics need to be measured between each segmentation result and the reference. In the following, we give a detailed description on how we choose references and metrics for specific tasks. A reproducible research page containing these segmentation results of our dataset can be found at <http://www.bsl.ece.vt.edu/ReproducibleResearch/SPIE10-Maps/maps.php>.

3.1 Agreement among segmentation algorithms.

In this experiment, we compute the agreement among segmentation results on T1-weighted images versus T1 maps. The basic assumption is that if one type of input image leads to a better agreement among the results obtained from different segmentation algorithms, it is believed to help segmentation, i.e., carries less ambiguity in segmentation processes. On each input image, regardless of whether it is a T1-weighted image or a T1 map, we first estimate a reference from the collection of segmentation results we obtained, and then compute the average distance from each individual segmentation result to the estimated reference. By comparing the average distances we obtained on each pair of corresponding T1-weighted images and T1 maps, we can identify which image type better facilitates segmentation.

STAPLE¹² is a tool designed to estimate a common agreement among multi-label segmentation results with a maximum log-likelihood, which assumes that the variations in segmentation results are random samples in a uni-model distribution. It has been demonstrated in¹³ that, when operated on the same input image, the performance difference among different automatic segmentation algorithms can be modeled as random variations. Thus, we can use STAPLE to estimate the reference. Specifically, for the i th subject, we obtain $STAPLE_w^i$, a STAPLE estimation of the 5 segmentations on a T1-weighted image, and $STAPLE_m^i$, that of the 4 segmentations on a T1 map.

Taking $STAPLE_w$ or $STAPLE_m$ as the reference segmentation, we need to compute a metric between each individual segmentation and the corresponding reference. Given two binary label maps X and Y , the Jaccard

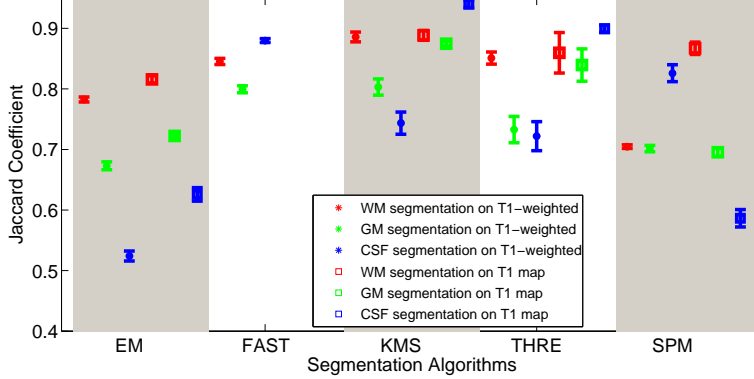


Figure 3. Jaccard coefficients of individual segmentations w.r.t the corresponding STAPLE reference. The red/green/blue symbols show the JC obtained for WM/GM/CSF respectively. For each segmentation algorithm, the first 3 columns plot the JCs on T1-weighted images and the next 3 columns plot those on T1-maps. Error bar demonstrates the variance of the JCs across all subjects.

coefficient (JC) is defined as the cardinality of the intersections of the foregrounds divided by the cardinality of the union of the foregrounds in the two binary label maps:¹³

$$JC(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}. \quad (1)$$

JC is a normalized measure of the relative overlap in $[0, 1]$. A JC close to 1 indicates a high similarity. Since each brain image is segmented into 3 classes, we obtain 3 JCs, one for each class. Except for FAST, where T1 maps are not successfully segmented, we have 6 measurements for each of the four remaining segmentation algorithms. After this computation, the average JCs over all subjects are plotted in Fig. 3. As an example, the first 6 columns in the figures plot the JCs for EM segmentation algorithm, where column 1 and 4, 2 and 5, 3 and 6 show the values obtained on T1-weighted images and T1 maps, for WM, GM and SCF, respectively:

$$JC(\text{EM}, \text{STAPLE})_w = \frac{1}{n} \sum_{i=1}^n JC(\text{EM}_w^i, \text{STAPLE}_w^i),$$

$$JC(\text{EM}, \text{STAPLE})_m = \frac{1}{n} \sum_{i=1}^n JC(\text{EM}_m^i, \text{STAPLE}_m^i).$$

From Fig. 3, we find that except for the CSF segmentation in SPM, in all other cases, segmentations on T1 maps report larger JCs with respect to the corresponding STAPLE reference (the JC of GM segmentation in SPM is about the same in the two cases). This observation indicates that segmentation algorithms tend to better agree with each other on T1 maps, compared to T1-weighted images. We believe this observation is an evidence that the intensity consistency in T1 maps does improve segmentation performance.

3.2 Relative performance of segmentation on T1-weighted images and T1 maps.

In this experiment, we directly compare the relative quality of the segmentation results on T1-weighted images versus T1 maps. To do so, we are first interested in observing the relative affinity of all the segmentation results.

As shown in,¹³ the approximate relative affinity of different segmentations can be visualized by projecting them onto a 2D plane via MDS technique, where in a good MDS projection, the distance between each pair of 2D projections should be close to the actual distance between the pair of samples before projection. Similarly, we use the average $(1 - JC)$ values between each pair of binary segmentation label maps as the distance measure, and show the MDS plots in Fig. 4(a-c). The embedding errors are given in Fig. 4(d-f), where we observe that the majority of the samples are distributed close to the diagonal, i.e., the MDS projections are reasonably accurate.

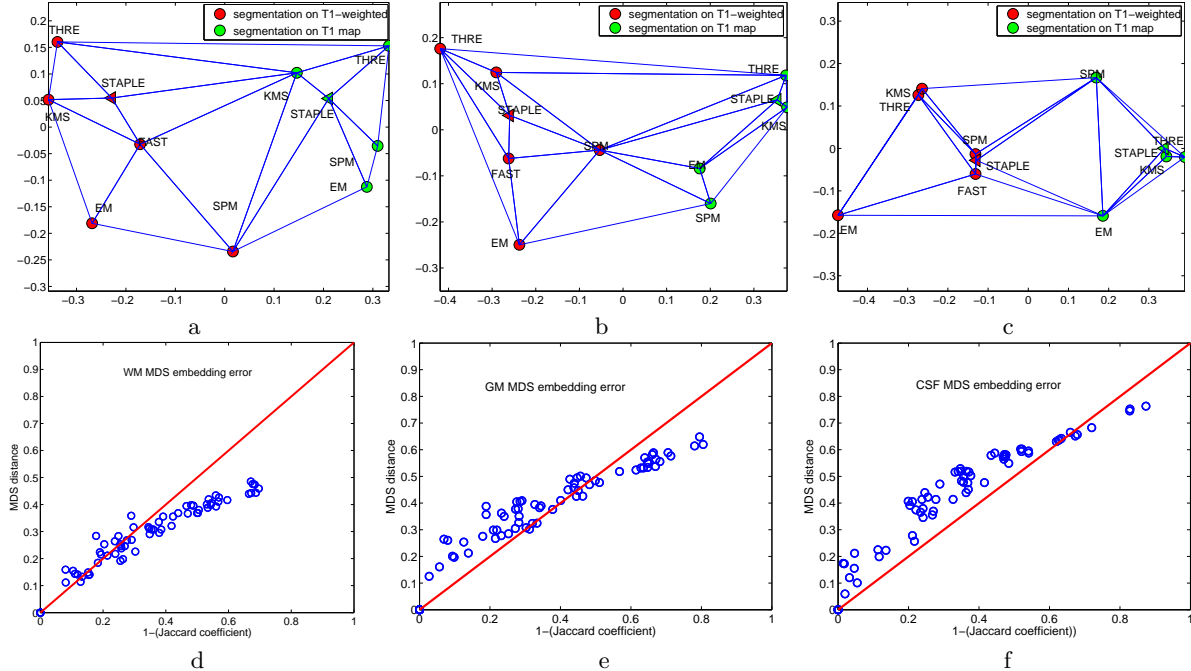


Figure 4. MDS projection. Top row: MDS plot for WM, GM and CSF, respectively. The green nodes represent the segmentation results on T1-weighted images, and the red nodes represent those on T1 maps. Lower row: the plot of the embedding errors of the MDS plot for WM, GM and CSF, respectively.

From Fig. 4(a-c), we clearly notice that for all 3 tissue classes, the nodes of the same color tend to cluster and a clear separation can be seen between nodes of different colors. Based on this observation, we believe the difference of the segmentation results on the two types of input images is not random error (which is shown to be the case when different segmentation algorithms are adopted for the *same* type of input image¹³), but some systematic bias. Consequently, the “common agreement” among all these segmentation results across two types of input images can lie in between the two clusters, and may not agree well with the information in any of the input images. In this case, the “common agreement” estimated by STAPLE is expected to have reduced credibility when serving as the reference.

As a good alternative, if both T1-weighted image and T1 map are presented, the manual work by human expert can be trusted as the reference segmentation. However, due to the huge amount of manual work involved, it is impossible to obtain expert segmentation for the entire dataset. Following the method presented in,¹³ we invited a neurologist to manually segment two representative sub-regions for each subject. The sub-regions selected are from axial slices and of size 35×30 . Examples of the selected sub-regions for a pair of T1-weighted image and T1 map are shown in Fig. 5. The first sub-region is selected around the ventricle, with certain randomness on the exact location across different subjects. We mark this sub-region with red in the Fig. 5, and refer to it as the *red region*. The second sub-region is randomly selected on all other brain regions but does not cover the ventricle area. We mark the second sub-region with blue in Fig. 5, and refer to it as the *blue region*. To obtain the manual segmentation of these two sub-regions, we simultaneously present the entire slice of T1-weighted images and T1 maps with the region to be segmented being marked out, as the ones shown in Fig. 5 (a) and (b). The cut-off sub-regions from T1-weighted images and T1 maps are also presented on the screen, as shown in (c) and (d) for the blue region, or as shown in (f) and (g) for the red region. The expert then cross-references the two cut-off regions, aided by the image context provided by the two whole slices, and chooses to record his/her segmentation on either of the cut-off regions. The manual segmentation results for the samples given in Fig. 5 are shown in (e) and (h) for the blue and the red regions, respectively.

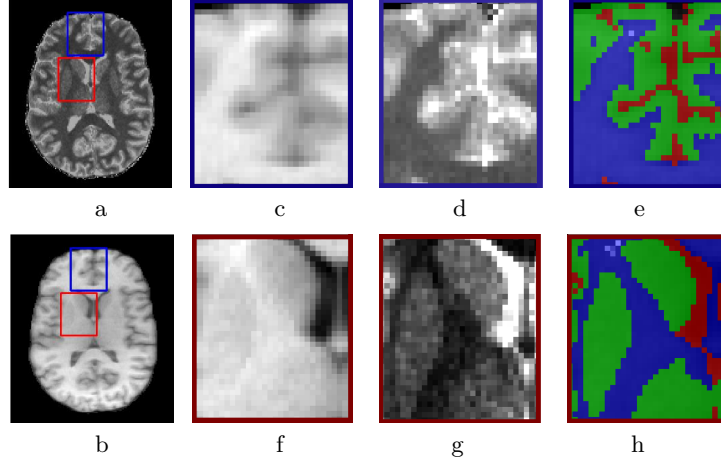


Figure 5. Example of expert segmentation of image sub-regions. (a) and (b): chosen axial slice of the T1-weighted image and the T1 map of one subject, respectively. (c) and (d): the cut-off blue region from (a) and (b), respectively. (f) and (g): the cut-off red region from (a) and (b), respectively. (e) and (h): manual segmentation for the blue and the red regions, respectively.

3.2.1 Segmentation quality with respect to manual segmentation.

Taking the expert manual segmentations as the reference, we can compute JCs between each of the automatic segmentation results and the reference. The JCs computed for the blue regions and the red regions are plotted in Fig. 6 (a) and (b), respectively. Similar as in Fig. 3, the makers indicate the average value of JC over all the subjects and the variance of the values are plotted as error bars. From Fig. 3, we can see that the agreements between the segmentation results on T1 map and the reference are unanimously higher compared to the corresponding ones on T1-weighted images, across all automatic segmentation algorithms, and all tissue classes. We believe this is a strong evidence that utilizing T1 maps does facilitate segmentation. Note that from this analysis, there is no evidence showing that the performance variation of segmentation algorithms is reduced by segmenting T1 maps.

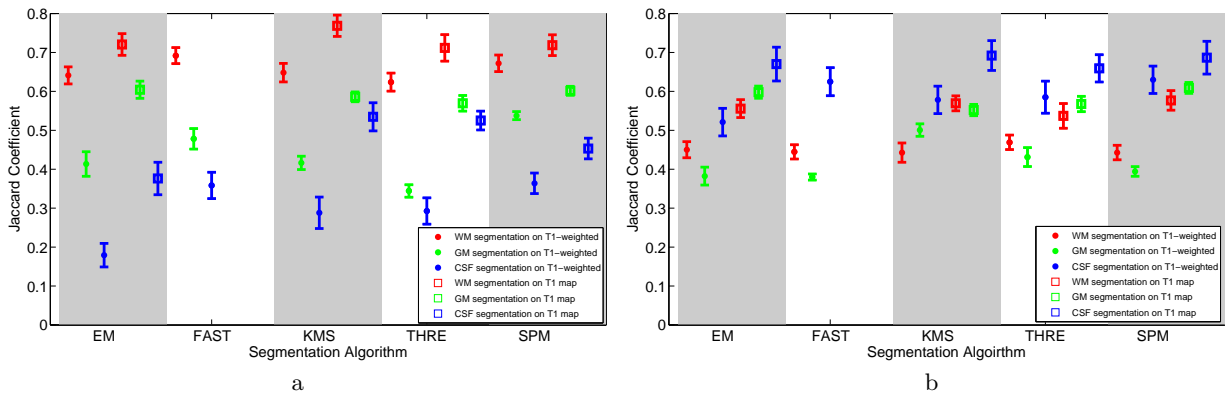


Figure 6. Jaccard coefficients computed between automatic segmentations and expert manual segmented references. The red/green/blue symbols shows the JC obtained for WM/GM/CSF respectively. For each segmentation algorithm, the first 3 columns plot the JC on T1-weighted images and the next 3 columns plot that on T1-maps. Error bar demonstrates the variance of the JCs across all subjects. a) JC values for the blue regions, b) JC values for the red regions.

Quantitatively, averaged over different segmentation algorithms, the improvements of JCs by segmenting T1 maps over T1-weighted images are summarized in Tab. 1. We can see that in both the red and the blue regions, the quality of GM segmentation is improved significantly. Also, for the blue regions, CSF segmentation is largely

improved, where for the red regions more improvement can be seen on the WM segmentation. In general, the GM and WM segmentation around the ventricle area (red region) is very difficult, due mainly to the weak boundaries and uneven intensity of sub-cortical gray structures. Based on our study, we believe the quality of sub-cortical segmentation can be improved by shifting to quantitative images, e.g., T1 maps.

Table 1. Improvement of Jaccard coefficient by segmenting T1 maps over segmenting T1-weighted images. JCs are computed by taking expert segmentation as the reference.

	WM	GM	CSF
Blue regions	0.0836	0.1624	0.1914
Red regions	0.1085	0.1545	0.0985

4. CONCLUSION AND FUTURE WORKS

We provide a systematic study on the segmentation quality when a collection of automatic brain MRI segmentation algorithms are applied on T1-weighted images and T1 maps. In this study, the T1-weighted images and T1 maps are simultaneously captured for a group of 16 subjects, which allow us for a direct comparison between the segmentation results on these two types of images.

The results of several experiments presented in this work supports the assumption that T1 maps, as a quantitative measure of pure T1, facilitates segmentation, compared to widely used T1-weighted images. Using T1 maps, different automatic segmentation algorithms tend to better agree with each other. Also, the segmentations obtained on T1 maps are more similar to human expert manual segmentations.

In the future, we are interested in a similar study on T2 maps, or the combination of T1 and T2 maps. It has been shown that T1- and T2-weighted images can be jointly used in multi-channel segmentation algorithms, for a more robust segmentation result. Thus, we believe T1 and T2 maps potentially can also be combined to further improve segmentation performance. In addition, considering that undergoing pathology can be better identified in T1 maps, lesion segmentation would be an interesting and promising research topic.

Acknowledgment

We would like to thank Dr. Warfield for granting the access to the implementation of STAPLE algorithm. Also, thanks to Merideth Addicott from the School of Medicine at Wake Forest University for the manual segmentation of brain sub-regions.

This research was funded by the National Institutes of Health through the National Institute on Alcohol Abuse and Alcoholism, Grant R01-AA016748; and the NIH Roadmap for Medical Research, Grant U54 RR021813.

REFERENCES

- [1] Helms, G., Kallenberg, K., and Dechent, P., “Contrast-driven approach to intracranial segmentation using a combination of t2- and t1-weighted 3d mri data sets,” *Journal of Magnetic Resonance Imaging* **24**(4), 790–795 (2006).
- [2] Chen, P., Steen, R., Yezzi, A., and Krim, H., “Joint brain parametric t1-map segmentation and rf inhomogeneity calibration,” *International Journal of Biomedical Imaging* (2009).
- [3] Deoni, S., “High-resolution t1 mapping of the brain at 3t with driven equilibrium single pulse observation of t1 with high-speed incorporation of rf field inhomogeneities (despot1-hifi),” *Journal of Magnetic Resonance Imaging* **6**, 1106–1111 (2007).
- [4] Deoni, S., Rutt, B., and Peters, T., “Rapid combined t1 and t2 mapping using gradient recalled acquisition in the steady state,” *Magnetic Resonance in Medicine* **49**(3), 515–526 (2003).
- [5] Bluml, S., Schad, L., Boris, S., and Lorenz, W., “Spin-lattice relaxation time measurement by means of a turboflash technique,” *Magnetic Resonance in Medicine* **30**, 289–295 (1993).
- [6] McKenzie, C., Chen, Z., Drost, D., and Prato, F., “Fast acquisition of quantitative t2 maps,” *Magnetic Resonance in Medicine* **41**, 208–212 (1999).
- [7] Henderson, E., McKinnon, G., Lee, T., and Rutt, B., “A fast 3d look-locker method for volumetric t1 mapping,” *Magnetic Resonance in Medicine* **17**, 11631171 (1999).

- [8] Chen, P., Steen, R., Yezzi, A., and Krim, H., “Brain mri t1-map and t1-weighted image segmentation in a variational framework,” in [*IEEE International Conference on Acoustics, Speech, and Signal Processing*], 417–420 (2009).
- [9] Vrenken, H., Rombouts, S., Pouwels, P., and Barkhof, F., “Voxel-based analysis of quantitative t1 maps demonstrates that multiple sclerosis acts throughout the normal-appearing white matter,” *American Journal of Neuroradiology* **27**(4), 868–874 (2006).
- [10] Ashburner, J. and Friston, K., [*Spatial normalization using basis functions*], Academic Press (2003). In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner and W. Penny, Editors, *Human Brain Function* (2nd edition).
- [11] Grau, V., Mewes, A., Alcaiz, M., Kikinis, R., and Warfield, S., “Improved watershed transform for medical image segmentation using prior information,” *IEEE Transaction on Medical Imaging* **23**(4), 447458 (2004).
- [12] Warfield, S., Zou, K., and Wells, W., “Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation,” *Journal of Magnetic Resonance Imaging* **23**, 903–921 (2004).
- [13] Bouix, S., Martin-Fernandez, M., Ungar, L., Nakamura, M., Koo, M., McCarley, R., and Shenton, M., “On evaluating brain tissue classifiers without a ground truth,” *Neuroimage* **07**, 447458 (2007).
- [14] Zhang, Y., Brady, M., and Smith, S., “Segmentation of brain mr images through a hidden markov random field model and the expectation maximization algorithm,” *Journal of Magnetic Resonance Imaging* **20**(1), 45–57 (2001).
- [15] FSL. <http://www.fmrib.ox.ac.uk/fsl/>.
- [16] Pohl, K., Bouix, S., Kikinis, R., and Grimson, W., “Anatomical guided segmentation with non-stationary tissue class distributions in an expectationmaximization framework,” in [*IEEE International Symposium on Biomedical Imaging*], 81–84 (April 2004).
- [17] Slicer3. http://www.slicer.org/slicerWiki/index.php/Main_Page.
- [18] SPM. <http://www.fil.ion.ucl.ac.uk/spm/>.
- [19] Smith, S., “Fast robust automated brain extraction,” *Human Brain Mapping* **17**(3), 143–155 (2002).
- [20] M. Styner, G. Gerig, C. B. and Szekeley, G., “Parametric estimate of intensity inhomogeneities applied to mri,” *IEEE Trans. on Medical Imaging* **19**(3), 153–165 (2000).