Evaluating Automatic Brain Tissue Classifiers

Sylvain Bouix, Lida Ungar, Chandlee C. Dickey, Robert W. McCarley, and Martha E. Shenton

Surgical Planning Laboratory, Harvard Medical School, Boston, MA, USA.
Department of Psychiatry, Boston VA Healthcare System, Boston, MA, USA.

Abstract. We present a quantitative evaluation of MR brain images segmentation. Five classifiers were tested. The task was to classify an MR image into four different classes: background, cortical spinal fluid, gray matter and white matter. The performance was rated by first estimating a ground truth (EGT) using STAPLE and then analyzing the volume differences as well as the Dice similarity measure between each of the 5 classifiers.

Introduction: Classification of brain tissue classes into white matter, gray matter and cortical spinal fluid (CSF) is an essential step in most neuroanatomy studies based on MR images. Several algorithms have been presented over the past decade and their performances are ever improving. Moreover, in recent years, novel evaluation procedures have been developed and it is now possible to rate accurately different methodologies even when a ground truth is not available. In an effort to improve our own segmentation pipeline, we have performed an evaluation of five different brain tissue classifiers. **Methods:** Our data set consisted of 24 pairs of MR volumes acquired on a GE 1.5T scanner. The first volume is a 0.9375x0.9375x1.5mm SPGR coronal scan, the second volume is a 0.9375x0.9375x3mm T2 weighted axial scan. All the classifiers use both volumes for the segmentation. The first algorithm (**1**) is an implementation of the seminal Expectation Maximization (EM) framework of Wells et al. [5]. The second algorithm (**2**) is the output of (**1**) manually edited by an expert to remove non brain tissues. The third algorithm (**3**) is also an EM segmenter, but one that uses spatial information provided by a probabilistic atlas as well as a hierarchical model for the tissue classes [1]. The fourth (**4**) is an implementation of the improved Watershed segmentation by [4]. In the fifth method (**5**) the bias field was corrected with [5] before running [4]. An approximate ground truth classification was estimated using STAPLE [3], and was later used to evaluate volumetric differences between each of the 5 classifiers and the estimated ground truth. The Dice similarity measure between the ground truth and the individual segmentations was also calculated [2].

Table 1. Performance scores of the different classifiers

	Gray Matter					CSF					White Matter				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
`AVG(|X|-|EGT|)`	166	68	**-17**	141	119	-61	-116	**-5**	16	14	140	**6**	-16	-161	-138
`STD(|X|-|EGT|)`	42	28	33	27	30	25	24	22	14	10	30	19	29	25	29
Dice	0.86	**0.93**	0.83	0.85	0.87	0.54	0.57	0.67	0.90	**0.91**	0.90	**1.00**	0.91	0.83	0.87

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3217, pp. 1038–1039, 2004.

Results and Discussion: The results are shown in table 1. Figure 1 presents a box-and-whisker plot of the volume differences. The volumetric analysis ranks method (**3**) as the best method, but method (**5**) has the highest Dice scores. Our experts think method (**3**) is better but this judgment was not quantified. While it is difficult to know which is the best classifier, some conclusions can still be inferred: (i) thankfully, manual brain stripping always improves the results, (ii) bias field correction also always improves the segmentation, (iii) volumetric measurement and overlap measurements such as the Dice measure do not always agree, (iv) if one method is significantly different but better than all others, its score is likely to be low when compared to an *estimated* ground truth. In future work, we propose to investigate other measures, such as the sensitivity and specificity provided by STAPLE. We also plan to compare the different results to small regions of the brain previously manually segmented by experts.

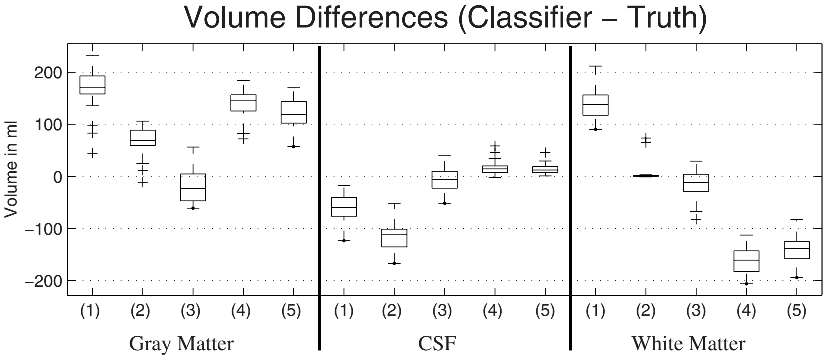

Fig. 1. Box-and-Whisker plot of the volume difference between each classifier and EGT

Acknowledgments. We acknowledge the support of NIH (K02 MH01110, R01 MH50747 to MES, R01 MH40799 to RWM), the Dept of Vet. Affairs Merit Awards (MES, RWM), Career Devel. Award (CCD).

References

1. K. M Pohl, W. M. Wells, A. Guimond, K. Kasai, M. E. Shenton, R. Kikinis, W. E. L. Grimson, S. K. Warfield. Incoperating non-rigid registration into expectation maximization algorithm to segment mr images. In *MICCAI*, pages 564–572, 2002.
2. L.R.Dice. Measure of the amount of ecological association between species. *Ecology*, 26:297–302, 1945.
3. S. K. Warfield, K. H. Zou, W. M. Wells. Validation of image segmentation and expert quality with an expectation-maximazation algorithm. In *MICCAI*, 2002.
4. V. Grau, A.J.U. Mewes,M. Alcaniz, R. Kikinis, S.K. Warfield. Improved watershed transform for medical image segmentation using prior information. *IEEE Trans Med Imag. In Press.*, 2003.
5. W.M. Wells III, W.E.L Grimson, R. Kikinis, F.A Jolesz. Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 15:429–442, 1996.