**Automatic Evaluation of Brain MR Segmentation Techniques by Level of Agreement**

Sylvain Bouix\(^1,2\), Marcos Martin-Fernandez\(^2\), Lida Ungar\(^1\), Martha E. Shenton\(^1\), Robert W. McCarley\(^1\)

\(^1\)Dept of Psychiatry, Boston VA Healthcare System, Harvard Medical School.
\(^2\)Surgical Planning Laboratory, Brigham and Women's Hospital, Boston, MA.

**Objective:** We are interested in finding a way of comparing and evaluating different segmentation techniques in a robust and automatic way. In this project, we have compared five different brain tissue classifiers applied on 40 different MR images. The evaluation is based on the notion of agreement, i.e. how consistent is a specific technique with all the others.

**Methods:**

Our data set consists of 40 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 of the brain into four tissue classes: Gray Matter, White Matter, CSF and Background.

The first algorithm (EM) is an implementation of the Expectation Maximization (EM) framework of Wells et al. [1]. The second algorithm (Watershed) is the improved Watershed segmentation [2]. The third algorithm (EMAtlas) is an EM segmenter that uses spatial information provided by a probabilistic atlas [3]. The fourth (FSL) is the multichannel segmentation (MFAST) of the FSL software package. The fifth method (SPM) is the segmentation technique of the SPM2 package.

Two metrics are used to measure the level of agreement between the different techniques. One is the Dice Similarity Coefficient, the other the normalized volume agreement of the segmentation. For each metric, tissue class and classifier a score was obtained by computing the Williams index of the agreements [4]. If the index is above one the classifier under study agrees with the other classifiers as well as they agree with each other. The higher the index the better. Equations of the metrics and Williams index are given in Figure 1.

**Williams Index:**

\[
I_i = \frac{1}{n-1} \sum_{j \neq i} \frac{A_{i,j}}{n(n-2) \sum_{j \neq i} A_{j,j'}}
\]

**Metric 1: Dice Similarity Coefficient (DSC):**

\[
A(i, j) = 2 \frac{|L_i \cap L_j|}{|L_i| + |L_j|}
\]

**Metric 2: Normalized Volume Agreement (VD):**

\[
A(i, j) = 1 - \frac{|L_i| - |L_j|}{|L_i| + |L_j|}
\]

**Fig. 1:** Equations for the Williams Index and the Metrics.
**Results & Discussion:**
The results are presented in Whisker plots in Figures 2 and 3. According to our metrics, one can see FSL is the least consistent of all techniques. Watershed performs best for most tissue classes, EMAtlas and SPM also do well.

**Fig. 2:** Whisker plots of the Williams Index for the Dice Similarity Coefficient metric.

**Fig. 3:** Whisker plots of the Williams Index for the Volume Agreement metric.

An example segmentation is presented in Figure 4. One should also observe different metrics give different rankings.
Conclusions:
We presented a novel approach to automatically evaluate segmentation techniques. The method is flexible and allows for different metrics to be used. One should be cautious when drawing conclusions as the level of agreement does not always reflect the quality of a segmentation. Nevertheless, this method is very good at finding outliers and inconsistent classifiers.

References & Acknowledgements:

We acknowledge the support of NIH (K02 MH01110, R01 MH50747 to MES, R01 MH40799 to RWM), the Dept of Vet. Affairs Merit Awards and REAP award (MES, RWM) and Fulbright (FU2003-0968 to MMF).