Two Methods for Validating Brain Tissue Classifiers

Marcos Martin-Fernandez[1,2], Sylvain Bouix[1,3], Lida Ungar[3],
Robert W. McCarley[3], and Martha E. Shenton[3]

[1] Laboratory of Mathematics in Imaging,
Brigham and Women's Hospital, Boston, MA, USA
[2] Laboratorio de Procesado de Imagen, Universidad de Valladolid, Spain
[3] Department of Psychiatry, Harvard Medical School,
VA Boston Healthcare System, Brockton, MA, USA

Abstract. In this paper, we present an evaluation of seven automatic brain tissue classifiers based on level of agreements. A number of agreement measures are explained, and we show how they can be used to compare different segmentation techniques. We use the Simultaneous Truth and Performance Level Estimation (STAPLE) of Warfield et al. but also introduce a novel evaluation technique based on the Williams' index. The methods are evaluated using these two techniques on a population of forty subjects, each having an SPGR scan and a co-registered T2 weighted scan. We provide an interpretation of the results and show how similar the output of the STAPLE analysis and Williams' index are. When no ground truth is required, we recommend the use of Williams' index as it is easy and fast to compute.

1 Introduction

In today's medical imaging field when one introduces a new segmentation technique, one has to thoroughly validate it and compare it to previously published, well accepted techniques. If a true segmentation exists this task is relatively easy as one only needs to choose a metric measuring differences between the ground truth and the output of the segmentation algorithm. Common metrics are volume differences or measures of overlap [1]. Unfortunately, ground truths or even human expert segmentations are rarely available especially for brain tissue classification where labeling a single brain into gray matter (GM), white matter (WM), cortical spinal fluid (CSF) and background (BG) would take days. Nevertheless, in recent years, novel evaluation procedures have been developed to overcome this problem, and it is now possible, to a certain degree, to rate different methodologies even when a ground truth is not available [2, 3]. In this work, we introduce a novel technique to evaluate brain segmenters based on agreement level and compare it to evaluating each segmenter with STAPLE's estimated ground truth. Seven different classifiers are tested over a data set of 40 different subjects. Our findings show few differences between the results of our technique and those of STAPLE, except for the fact that our evaluation is much faster.

J. Duncan and G. Gerig (Eds.): MICCAI 2005, LNCS 3749, pp. 515–522, 2005.

2 Measuring Segmentation Quality

2.1 Williams Index

Consider a set of r raters labeling a set of n voxels with labels $\{1, \cdot, l\}$. Let $\mathbf{D}$ denote the set of all labeled voxels (the label map) of all raters. D_{ij} represents the label of rater j for voxel i; $\mathbf{D}_j$ denotes the label map of rater j; and $a(\mathbf{D}_j, \mathbf{D}_{j'})$ is the agreement between rater j and j' over all n voxels. Several agreement measures can be used and a few will be defined in section 2.3. Williams' index I_j for rater j is defined as [4]:

$$I_j = \frac{(r-2)\sum_{j' \neq j}^{r} a(\mathbf{D}_j, \mathbf{D}_{j'})}{2\sum_{j' \neq j}^{r} \sum_{j'' \neq j}^{j'} a(\mathbf{D}_{j'}, \mathbf{D}_{j''})} \tag{1}$$

If the upper limit of the confidence interval of this index is greater than one, it can be concluded that rater j agrees with the other raters at least as well as they agree with each other [4]. Using the agreements defined in section 2.3 we can study the statistics of Williams' index for each algorithm, for each label and for each subject.

2.2 Multi-label STAPLE Algorithm

In this section, we describe the multi-label version of the Simultaneous Truth and Performance Level Estimation (STAPLE) Algorithm [3]. This algorithm calculates an estimated ground truth label map out from a set of r given segmentations (raters). Consider a label map with n voxels taking one of l possible labels. Let $\boldsymbol{\theta}_j$ be an $l \times l$ matrix. Each element $\boldsymbol{\theta}_j(s', s)$ describes the probability that rater j labels a voxel with s' when the true label is s. The perfect rater will have $\boldsymbol{\theta}_j$ equal to the identity matrix. Let $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r]$ be the unknown set of all probability matrices characterizing all r raters. Let $\mathbf{T} = (T_1, \ldots, T_n)^T$ be a vector representation of the unknown ground truth segmentation and $\mathbf{D}$ an $n \times r$ matrix whose columns are the r known segmentations. $\mathbf{D}$ is the incomplete data and $(\mathbf{D}, \mathbf{T})$ the complete data. STAPLE is an estimation process based on the EM algorithm which can estimate the ground truth $\mathbf{T}$ and the parameter matrix $\boldsymbol{\theta}$ at the same time by maximizing the complete data log likelihood $f(\mathbf{D}, \mathbf{T}|\boldsymbol{\theta})$. We refer the reader to [3] for the technical details of the optimization process. Once the ground truth is known, we can use any of the normalized metrics defined in section 2.3 for each algorithm, for each label and for each subject and study the resulting statistics.

2.3 Similarity Measurements

Consider two binary images I_1 and I_2 defined over a finite grid (lattice) L of N spatial sites x. Let $X = \{x \in L, I_1(x) = 1\}$ and $Y = \{x \in L, I_2(x) = 1\}$, and let us define four scalar measurements, $a_1 = |X \cap Y|, a_2 = |X| - a_1, a_3 = |Y| - a_1$ and $a_4 = N - |X \cup Y|$ as shown schematically in figure 1. We can then express, using these four values, the following similarity measurements, all of them taking values between 0 and 1.

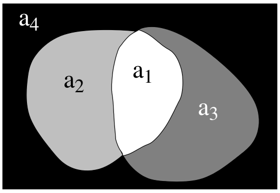

Fig. 1. Schematic diagram for sets X and Y and scalar values a_1 (white), a_2 (light gray), a_3 (dark gray) and a_4 (black)

- **Jaccard (JC)** [5]: $\frac{a_1}{a_1+a_2+a_3} = \frac{|X\cap Y|}{|X\cup Y|}$. It is zero when X and Y are disjoint and one when the sets are equal, i.e. $a_2 = a_3 = 0$.
- **Tanimoto (TN)** [6]: $\frac{a_1+a_4}{a_1+2a_2+2a_3+a_4} = \frac{|X\cap Y|+|\overline{X\cup Y}|}{|X\cup Y|+|\overline{X\cap Y}|}$ where $\overline{X}$ is the set $L - X$. It is zero when X and Y are disjoint and $X \cup Y = L$ and it is one when the sets are equal.
- **Volume similarity (VS)**: $1 - \frac{|a_2-a_3|}{2a_1+a_2+a_3} = 1 - \frac{||X|-|Y||}{|X|+|Y|}$. It is one when the number of elements in both sets are equal, and zero when one of the sets is empty. The positions of the points is irrelevant, only their number counts.

3 Experiments

3.1 Segmentation Pipelines

Data Set: Our data set consists of forty female subjects. The acquisition protocol involves two MR pulse sequences acquired on a 1.5-T GE scanner. First, a SPoiled Gradient-Recalled (SPGR) sequence yielded a coronal MR volume of size 256 × 256 × 124 and voxel dimensions 0.9375 × 0.9375 × 1.5mm. Second, a double-echo spin-echo sequence gave two axial MR volumes (proton density and T2 weighted) of size 256 × 256 × 54 and voxel dimensions 0.9375 × 0.9375 × 3mm. For each subject, both axial volumes were co-registered and resampled to the SPGR volume coordinate space using a Mutual Information rigid registration algorithm [7]. Due to limitations on the number of inputs of some of the classification algorithms, only the resampled T2 weighted and the original SPGR were used for segmentation.

Segmentation Techniques: Seven different automatic classifiers were evaluated. The task given was to segment the brain into four classes: BG, CSF, GM and WM. The algorithms were used "as is" with no special tuning of the parameters. A description of each segmenter follows:

- **kNN:** A statistical classification, whose core is a k Nearest Neighbor classifier algorithm trained automatically by non linear atlas registration [8].
- **MINC:** A back-propagation Artificial Neural Network classifier, trained automatically by affine atlas registration [9], the pipeline also includes its own bias field correction tool [10].

Table 1. Segmentation Pipeline Features. The "O" marks a missing feature in the pipeline. In such cases, standard tools were used (see text).

	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
Filtering	O	X	X	X	X	X	O
Bias Correction	O	X	X	X	X	X	O
Brain stripping	O	O	X	O	O	X	X

– **FSL:** A classification algorithm which makes use of a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm [11].
– **SPM:** A mixture model clustering algorithm, which has been extended to include spatial priors and to correct image intensity non-uniformities [12].
– **EM:** The original implementation of the Expectation Maximization algorithm designed by Wells et al. [13].
– **EMAtlas:** An EM based segmentation incorporating a Markov Random Field Model, and spatial prior information aligned to subject's space by non linear registration [14].
– **Watershed:** A watershed based segmentation which also incorporates spatial prior information in the form of a non linearly aligned atlas [15].

For atlas based classifiers, all techniques use differently defined spatial priors, except for EMAtlas and Watershed which share the same atlas.

Pre- and Post-Processing: Most segmentation techniques are usually a full pipeline involving (i) filtering, (ii) bias field correction, (iii) tissue classification and (iv) brain stripping. As shown in table 1, some methods did not have the all these tools embedded in their framework. We thus used the following three techniques when necessary: **filtering,** the data was smoothed using a diffusion based anisotropic filter [16]; **bias field correction,** was done using the technique of Wells et al. [13]; **brain stripping,** the brain was extracted using the Brain Extraction Tool [17].

3.2 Statistical Analyses

Let $X_{il} = \{x, \mathbf{D}_i(x) = l\}$ be the set of voxels labeled l by rater i. Let T be the estimated ground truth computed by STAPLE from all seven label maps, with T_l the set of voxels labeled l in T. For each subject, for each label, four different types of analysis were done. First, Williams' index was computed for each label using the seven X_{il} as input and all three agreement measures (Williams 1). The mean and standard deviation of the index over all subjects for each label and segmentation algorithm are shown in table 3. Second, for each rater, the three agreement measures between X_{il} and the estimated ground truth for that label T_l were computed (Staple 1). The mean and standard deviation of the index over all subjects are shown in table 2. Third, $C_l = \cap_{1 \leq i \leq r} X_{il}$, the set of all points with the same label l in all r label maps was computed. Williams' index was then calculated for each label l using all seven $X_{il} - C_l$ as input (Williams 2).

Table 2. Staple 1

BG	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	1.00(0.00)	0.99(0.00)	0.99(0.00)	0.99(0.00)	0.99(0.00)	0.99(0.00)	0.99(0.00)
TN	0.99(0.00)	0.99(0.00)	0.99(0.00)	0.98(0.01)	0.98(0.01)	0.98(0.01)	0.98(0.01)
VS	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.99(0.00)	1.00(0.00)
GM	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	*0.85(0.09)*	0.77(0.03)	0.73(0.23)	0.82(0.03)	0.72(0.03)	0.77(0.03)	**0.87(0.03)**
TN	*0.98(0.01)*	0.97(0.01)	0.96(0.04)	0.97(0.01)	0.96(0.00)	0.97(0.01)	**0.98(0.00)**
VS	**0.98(0.01)**	0.96(0.02)	0.96(0.06)	0.97(0.01)	*0.98(0.02)*	*0.98(0.02)*	*0.98(0.02)*
CSF	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	**0.73(0.11)**	0.46(0.14)	0.63(0.24)	0.60(0.12)	0.36(0.11)	0.42(0.12)	*0.62(0.10)*
TN	**0.98(0.01)**	0.97(0.01)	*0.98(0.02)*	**0.98(0.01)**	0.97(0.01)	0.96(0.01)	**0.98(0.01)**
VS	**0.96(0.04)**	0.76(0.11)	0.89(0.12)	*0.91(0.07)*	0.52(0.12)	0.81(0.07)	0.86(0.06)
WM	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	*0.87(0.09)*	0.80(0.04)	0.80(0.13)	0.82(0.03)	0.80(0.03)	0.83(0.03)	**0.90(0.02)**
TN	*0.99(0.01)*	0.98(0.00)	0.97(0.02)	0.98(0.00)	0.98(0.00)	0.98(0.00)	**0.99(0.00)**
VS	**0.98(0.01)**	*0.97(0.02)*	0.92(0.10)	0.91(0.02)	0.89(0.02)	0.96(0.03)	*0.97(0.02)*

The rationale for this process is to try to get a more focused analysis on the differences between raters. The results are show in table 5. Finally, the three agreement measures between $X_{il} - C_l$ and $T_l - C_l$ was calculated (Staple 2). The results are shown in table 4. In all the tables the best score is in **bold** and the second best score in *italic*.

4 Results

Table 2 shows results for Staple 1. We can see that for BG, none of the three agreement measurements give meaningful results. For GM and WM, Watershed is best one followed very closely by kNN when looking at JC and TN. Using VS, kNN performs slightly better than Watershed, which is similar to EM and EMAtlas for GM and similar to MINC for WM. For CSF, kNN is best, and Watershed and SPM rank second.

Table 3 shows the results for Williams 1. Similarly to Staple 1, none of the measurements are meaningful for BG. For GM, using JC and TN, Watershed is best followed closely by kNN. Using VS, kNN and EMAtlas are slightly better than Watershed. For WM, using JC and TN, Watershed is also best, followed by kNN for JC and SPM and EMAtlas for TN. Using VS, kNN is best, MINC and Watershed are second. All of these results mostly agree with the ones using Staple 1. Finally, for CSF, SPM is best followed by Watershed using JC. Watershed is best using TN and VS. In this case, Staple 1 and Williams 1 only agree for TN, and there is disagreement for JC and VS.

Table 4 shows the results for Staple 2. In comparison to Staple 1 (table 2), similar rankings are obtained for GM, CSF and WM for all the agreement measures. The main difference are in evaluating BG. More significance is

Table 3. Williams 1

BG	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
TN	1.00(0.00)	1.00(0.00)	0.99(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
VS	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
GM	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	*1.07(0.10)*	1.00(0.07)	0.90(0.25)	1.03(0.05)	0.94(0.07)	1.00(0.07)	**1.09(0.06)**
TN	**1.01(0.01)**	*1.00(0.01)*	0.99(0.03)	*1.00(0.01)*	0.99(0.01)	*1.00(0.01)*	**1.01(0.01)**
VS	**1.01(0.01)**	0.99(0.03)	0.97(0.06)	*1.01(0.02)*	1.00(0.02)	**1.01(0.01)**	*1.01(0.02)*
CSF	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	1.04(0.12)	1.08(0.14)	0.79(0.36)	**1.23(0.08)**	0.81(0.17)	0.99(0.13)	*1.18(0.09)*
TN	0.99(0.01)	*1.00(0.00)*	0.99(0.01)	*1.00(0.00)*	*1.00(0.00)*	*1.00(0.00)*	**1.01(0.00)**
VS	1.00(0.05)	1.04(0.08)	0.99(0.05)	1.08(0.06)	0.72(0.13)	*1.09(0.04)*	**1.11(0.04)**
WM	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	*1.04(0.09)*	0.97(0.04)	0.96(0.13)	0.99(0.04)	0.98(0.05)	1.01(0.05)	**1.07(0.03)**
TN	1.00(0.01)	1.00(0.01)	0.99(0.02)	*1.00(0.00)*	1.00(0.01)	*1.00(0.00)*	**1.01(0.00)**
VS	**1.04(0.02)**	*1.03(0.02)*	0.97(0.09)	0.96(0.03)	0.94(0.04)	1.02(0.03)	*1.03(0.02)*

Table 4. Staple 2

BG	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	0.37(0.16)	**0.50(0.22)**	0.28(0.11)	0.40(0.22)	0.40(0.23)	0.34(0.19)	*0.41(0.22)*
TN	0.99(0.00)	0.99(0.00)	0.99(0.00)	0.98(0.01)	0.98(0.01)	0.98(0.01)	0.98(0.01)
VS	*0.68(0.23)*	**0.70(0.25)**	0.65(0.26)	0.54(0.22)	0.54(0.24)	0.48(0.21)	0.56(0.24)
GM	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	*0.71(0.11)*	0.57(0.10)	0.55(0.24)	0.65(0.08)	0.47(0.10)	0.56(0.10)	**0.74(0.07)**
TN	*0.98(0.01)*	0.97(0.01)	0.96(0.04)	0.97(0.01)	0.96(0.00)	0.97(0.01)	**0.98(0.00)**
VS	**0.96(0.03)**	0.91(0.05)	0.93(0.08)	*0.95(0.03)*	0.94(0.05)	**0.96(0.03)**	0.95(0.04)
CSF	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	**0.69(0.12)**	0.36(0.17)	*0.59(0.23)*	0.53(0.14)	0.22(0.16)	0.32(0.15)	0.54(0.12)
TN	**0.98(0.01)**	0.97(0.01)	*0.98(0.02)*	**0.98(0.01)**	0.97(0.01)	0.96(0.01)	**0.98(0.01)**
VS	**0.95(0.05)**	0.68(0.16)	0.88(0.12)	*0.90(0.09)*	0.33(0.20)	0.76(0.10)	0.83(0.08)
WM	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	*0.63(0.11)*	0.46(0.08)	0.53(0.16)	0.38(0.12)	0.52(0.05)	0.48(0.11)	**0.67(0.06)**
TN	*0.99(0.01)*	0.98(0.00)	0.97(0.02)	0.98(0.00)	0.98(0.00)	0.98(0.00)	**0.99(0.00)**
VS	**0.94(0.04)**	*0.90(0.06)*	0.80(0.20)	0.58(0.13)	0.69(0.05)	0.84(0.12)	*0.90(0.06)*

achieved for JC and VS, but not for TN which is still not very meaningful. MINC is performing best for BG using Staple 2, followed by Watershed using JC and by kNN using VS.

Table 5 shows the results for Williams 2. Again, the results for GM, CSF and WM are similar to Williams 1 (table 3) overall. The only difference is for JC and for WM, for which Watershed is best for Williams 1, while EM is best, followed by Watershed for Williams 2. With respect to BG, TN is still not meaningful. Using JC and VS, there is significance for Williams 2. For JC, EM performs best followed by SPM and for VS, Watershed is best followed by SPM and EM.

Table 5. Williams 2

BG	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	0.36(0.23)	1.09(0.24)	0.26(0.09)	*1.50(0.10)*	**1.54(0.09)**	1.46(0.10)	1.42(0.12)
TN	1.00(0.00)	1.00(0.00)	0.99(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
VS	0.64(0.23)	1.14(0.09)	0.54(0.14)	*1.25(0.05)*	*1.25(0.06)*	1.14(0.08)	**1.26(0.06)**
GM	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	*1.17(0.15)*	1.00(0.13)	0.89(0.38)	1.10(0.09)	0.80(0.18)	0.95(0.14)	**1.21(0.10)**
TN	**1.01(0.01)**	*1.00(0.01)*	0.99(0.03)	*1.00(0.01)*	0.99(0.01)	*1.00(0.01)*	**1.01(0.01)**
VS	**1.03(0.02)**	0.97(0.06)	0.96(0.09)	1.01(0.03)	0.99(0.05)	*1.02(0.03)*	*1.02(0.03)*
CSF	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	1.20(0.17)	1.02(0.22)	0.93(0.46)	**1.42(0.09)**	0.52(0.33)	0.93(0.18)	*1.29(0.13)*
TN	0.99(0.01)	*1.00(0.00)*	0.99(0.01)	*1.00(0.00)*	*1.00(0.00)*	*1.00(0.00)*	**1.01(0.00)**
VS	1.05(0.09)	1.05(0.13)	1.03(0.08)	1.15(0.10)	0.52(0.24)	*1.15(0.07)*	**1.19(0.08)**
WM	kNN	MINC	FSL	SPM	EM	EMAtlas	Watershed
JC	1.19(0.22)	0.89(0.15)	1.10(0.22)	0.60(0.20)	**1.27(0.12)**	0.91(0.23)	*1.21(0.13)*
TN	1.00(0.01)	1.00(0.01)	0.99(0.02)	*1.00(0.00)*	1.00(0.01)	*1.00(0.00)*	**1.01(0.00)**
VS	**1.17(0.06)**	*1.14(0.07)*	0.99(0.22)	0.71(0.18)	0.87(0.11)	1.07(0.15)	1.13(0.07)

5 Discussion

We have investigated different approaches to evaluate the quality of a segmentation only based on agreement measures. A great number of similarities have been found between Staple 1 and Williams 1, suggesting kNN is the most consistent segmentation method. In general for GM, WM and CSF, all ranking techniques give similar results. For BG, better significance is achieved for JC and VS after discarding the common agreement among algorithms and focusing only on differences (Williams 2 and Staple 2). Staple 2 presents MINC as the best method for BG and Williams 2 has EM as the better technique. In general, FSL, SPM and EMAtlas are less well ranked.

Overall, Williams' index gives similar results to STAPLE, and unless one absolutely needs the *estimated* ground truth of STAPLE for further processing, using Williams' index is sufficient. The biggest advantage is that of speed, as STAPLE in our experimental setup can take as much as 20mn to process one case, whereas Williams index takes a few seconds. We are now interested in studying the probability matrices provided by STAPLE and also finding more intuitive and compact ways to present the results of the evaluation.

Acknowledgments

We acknowledge the support of NIH (K02 MH01110, R01 MH50747, U54 EB005149 to MES, R01 MH40799 to RWM), the Dept of Vet. Affairs Merit Awards and REAP award (MES, RWM) and Fulbright (FU2003-0968 to MMF).

References

1. Zou, K.H., Wells, W.M., Kikinis, R., Warfield, S.K.: Three validation metrics for automated probabilistic image segmentation in brain tumors. Statistics in Medicine **23** (2004) 1280–1291
2. Bello, F., Colchester, A.C.F.: Measuring global and local spatial correspondence using information theory. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. (1998) 964–973
3. S. K. Warfield, K. H. Zou, W.M.W.: Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. IEEE Trans. Medical Imaging **23** (2004) 903–921
4. Williams, G.W.: Comparing the joint agreement of several raters with another rater. Biometrics **32** (1976) 619–627
5. Jaccard, P.: Étude comparative de la distribuition florale dans une portion des alpes et de jura. Bulletin de la Societé Voudoise des Sciences Naturelles **37** (1901) 547–579
6. Rogers, J.S., Tanimoto, T.T.: A computer program for classying plants. Science **132** (1960) 1115–1118
7. Wells, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R.: Multi-modal volume registration by maximization of mutual information. Medical Image Analysis **1** (1996) 35–52
8. Warfield, S.: Fast knn classification for multichannel image data. Pattern Recognition Letters **17** (1996) 713–721
9. Zijdenbos, A.P., Forghani, R., Evans, A.C.: Automatic "pipeline" analysis of 3-d mri data for clinical trials: application to multiple sclerosis. IEEE Transactions on Medical Imaging **21** (2002) 1280–1291
10. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A non-parametric method for automatic correction of intensity non-uniformity in mri data. IEEE Transactions On Medical Imaging **17** (1998) 87–97
11. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden markov random field model and the expectation maximization algorithm. IEEE Transactions on Medical Imaging **20** (2001) 45–57
12. Ashburner, J., Friston, K.: Spatial normalization using basis functions. In Frackowiak, R.S.J., Friston, K.J., Frith, C., Dolan, R., Friston, K.J., Price, C.J., Zeki, S., Ashburner, J., Penny, W., eds.: Human Brain Function. 2nd edn. Academic Press (2003)
13. Wells, W.M., Kikinis, R., Grimson, W.E.L., Jolesz, F.: Adaptive segmentation of mri data. IEEE Transactions on Medical Imaging **15** (1996) 429–442
14. Pohl, K.M., Bouix, S., Kikinis, R., Grimson, W.E.L.: Anatomical guided segmentation with non-stationary tissue class distributions in an expectation-maximization framework. In: IEEE International Symposium on Biomedical Imaging, Arlington, VA (2004) 81–84
15. Grau, V., Mewes, A.U.J., Alcaiz, M., Kikinis, R., Warfield, S.K.: Improved watershed transform for medical image segmentation using prior information. IEEE Transactions on Medical Imaging **23** (2004) 447–458
16. Krissian, K.: Flux-based anisotropic diffusion applied to enhancement of 3-d angiogram. IEEE Transactions on Medical Imaging **21** (2002) 1440–1442
17. Smith, S.: Fast robust automated brain extraction. Human Brain Mapping **17** (2002) 143–155