Discriminative, Semantic Segmentation of Brain Tissue in MR Images

Zhao Yi[1], Antonio Criminisi[2], Jamie Shotton[2], and Andrew Blake[2]

[1] University of California, Los Angeles, USA. zyi@ucla.edu.
[2] Microsoft Research Cambridge, UK.
{antcrim, jamie.shotton, ablake}@microsoft.com.

Abstract. A new algorithm is presented for the automatic segmentation and classification of brain tissue from 3D MR scans. It uses discriminative Random Decision Forest classification and takes into account partial volume effects. This is combined with correction of intensities for the MR bias field, in conjunction with a learned model of spatial context, to achieve accurate voxel-wise classification. Our quantitative validation, carried out on existing labelled datasets, demonstrates improved results over the state of the art, especially for the cerebro-spinal fluid class which is the most difficult to label accurately.

1 Introduction

This paper introduces a new, supervised technique for the classification of 3D MR scans of the brain. The ultimate goal is to assign a class label to each brain voxel from the following set: white matter, grey matter and cerebro-spinal fluid. Such automatic analysis is of practical interest to many clinical applications related to early detection and treatment of schizophrenia [1], epilepsy [2] and Alzheimer's [3]. Automatic segmentation of brain tissue is a challenging problem, owing to acquisition noise, non-uniformities in the MR magnetic field, the complex anatomy of the brain, limited resolution and partial volume effects.

In order to address these problems we propose an algorithm in three steps: 1) bias field correction using polynomials of optimal degree; 2) learned models for automatic tissue classification/segmentation; and 3) partial volume estimation. Model training accounts for much of the accuracy of our technique, and is utilized as much as possible, not only in the segmentation process, but also during bias field correction and partial volume estimation. The tissue classification step is achieved via randomized decision trees [4, 5], an efficient, state-of-the-art discriminative classification technique.

Previous Work. The substantial existing literature on this topic may be roughly grouped into the following four different sets:

Clustering algorithms. Representative work in this area includes the use of K-means [6], mean-shift [7], and expectation-maximization (EM) [8–10]. Their limitation is that the cluster geometry and the number of clusters have to be known, where parametric forms such as Gaussian or Gaussian mixtures are commonly assumed but without taking into consideration existing domain knowledge.

Atlas-based approaches. Segmentation is reduced to a template matching problem, where labels are transferred from a prelabeled atlas to the subject volume via registration techniques [2, 3, 11]. However, registration itself is challenging, especially for the human cortex due to the high variability of the cortical shape and the location of sulci and gyri across individuals.
Deformable models. Relying on curve propagation, deformable models minimize a certain energy associated with the curve to partition the image domain, like active contours [12] and level sets [13, 14]. Those techniques typically suffer from problems with initialization and local minima.
Supervised learning. Surprisingly, supervised learning has received relatively little attention in brain tissue segmentation. In [2] the intensity distribution of each class at every location is modeled as a Gaussian, with spatial information encoded globally via a probabilistic atlas and locally via an anisotropic non-stationary Markov random field. This Gaussian assumption, however, is restrictive to inter-subject variability and image distortions. The work in [15] learns a multi-class discriminative appearance model by a probabilistic boosting tree together with a generative active shape model for each subcortical structure. It works well for regular subcortical structures, but is not suitable for the brain tissue segmentation task which involves highly convoluted cortical surfaces.

2 Discriminative Brain Tissue Segmentation

Given the observed MR brain volume $I : \Omega \subset \mathbb{R}^3 \mapsto \mathbb{R}^+$ our goal is to assign to each voxel a class label from the following set: white matter (`WM`), gray matter (`GM`), and cerebro-spinal fluid (`CSF`). This task is formulated as a maximum-a-posteriori (MAP) classification problem, whose output is the label map $L^\star : \Omega \mapsto \{\texttt{CSF, GM, WM}\}$ such that

$$L^\star = \arg\max_L \ \log P(L|I) = \arg\max_L \ \log P(I|L) + \log P(L). \tag{1}$$

Under the simplistic but common assumption that voxel intensities are mutually independent given their labels, the data likelihood in (1) can be rewritten as

$$\log P(I|L) = \sum_{\mathbf{x}\in\Omega} \log P(I(\mathbf{x})|L(\mathbf{x})). \tag{2}$$

The label prior in (1) can be decomposed into two terms in the Markov Random Field framework, i.e.

$$\log P(L) = \sum_{\mathbf{x}\in\Omega} \log U(L(\mathbf{x})) + \sum_{\mathbf{x},\mathbf{y}} V(L(\mathbf{x}), L(\mathbf{y})), \tag{3}$$

where U is the unary location prior, and V imposes spatial smoothness between neighboring labels (not considered yet). The following sections describe details of how to model $P(I(\mathbf{x})|L(\mathbf{x}))$ as well as $U(L(\mathbf{x}))$. We start by looking at the likelihood $P(I(\mathbf{x})|L(\mathbf{x}))$ and how it is affected by the magnetic bias field.

2.1 Bias Field Correction

Owing to the bias field induced by the MR scanner, the observed intensity of voxels is a corrupted version of the true intensity of the underlying tissue. In order to model the likelihood $P(I(\mathbf{x})|L(\mathbf{x}))$, we need to recover the true intensity of each voxel by estimating the bias field and correcting for it.

Let $\bar{I}$ denote the true intensity, b the bias field, and n the random noise. Here a multiplicative bias with i.i.d. Gaussian noise is assumed, i.e., $I(\mathbf{x}) = b(\mathbf{x}) \cdot \bar{I}(\mathbf{x}) + n(\mathbf{x})$. This MR image formation model has been used frequently [16, 17] as it is simple and known to be consistent with the inhomogeneous sensitivity of the reception coil. Since the bias field is smoothly varying in space, we adopt a low-order polynomial model: $b(\mathbf{x}) = \boldsymbol{\lambda} \cdot \boldsymbol{\Gamma}^n(\mathbf{x})$, where $\boldsymbol{\lambda}$ is the coefficient vector, $n \in \{0, 1, 2, \ldots\}$ is the order of polynomial, and $\boldsymbol{\Gamma}^n$ is the base polynomial vector. For example, $\boldsymbol{\Gamma}^1(\mathbf{x}) = (x, y, 1)^T, \boldsymbol{\Gamma}^2(\mathbf{x}) = (x^2, xy, x, y^2, y, 1)^T$. As MR acquisition is done sequentially, it is reasonable to assume that $\boldsymbol{\lambda}$ is different slice by slice. Thus, our bias model holds for every individual slice and $\boldsymbol{\Gamma}^n$ is applied to (x, y) only (not to the third dimension).

On the other hand, we can assume that the true intensity of each voxel depends only on the underlying tissue label $\bar{I}(\mathbf{x}) = \mu_{L(\mathbf{x})}$, where $\mu \in \{\mu_{\texttt{CSF}}, \mu_{\texttt{GM}}, \mu_{\texttt{WM}}\}$ is the tissue intensity for label $L(\mathbf{x})$, and have uniform values throughout the volume. Given the values of n, I and $\boldsymbol{\Gamma}^n$, if L were known then iterative least squares fitting could be applied to determine the optimal solution of $\boldsymbol{\lambda}$ for each slice and $\mu_{\texttt{CSF}}, \mu_{\texttt{GM}}, \mu_{\texttt{WM}}$ for every volume. In practice, however, a ground-truth labeling for L is not available but probabilistic tissue labeling may be used to tackle the problem. Let $q_{\texttt{CSF}}(\mathbf{x}), q_{\texttt{GM}}(\mathbf{x}), q_{\texttt{WM}}(\mathbf{x})$ denote the probabilities of the voxel $\mathbf{x}$ belonging to each tissue. The expected value of the true intensity in this case is a weighted sum of all tissue intensities and our intensity model changes to

$$\bar{I}(\mathbf{x}) = \sum_{L \in \{\texttt{CSF},\texttt{GM},\texttt{WM}\}} q_L(\mathbf{x}) \mu_L. \tag{4}$$

The same iterative fitting procedure can be applied as before. The optimal degree of the polynomial is obtained on the validation set by performing model selection using T-tests for successive degrees ($n = 0, \cdots, 4$). The Jaccard index $\mathrm{JAC}(L, S) = |L \cap S| \ / \ |L \cup S|$ is used to measure the accuracy of the output label map L given the manual segmentation S. We obtain P values on `WM` and `GM` less than 5% between n and $n-1$ when $n \leq 3$, and greater than 5% when $n > 3$. P values on `CSF` are always greater than 5% indicating no statistically significant difference between degrees. This is because dark `CSF` regions are insensitive to multiplicative bias. Thus we choose $n = 3$ for accuracy.

2.2 Maximum a Posteriori (MAP) Classification

In brain MR images, the (bias-corrected) intensity of a given tissue is approximately uniform, and the spatial assignment of different tissues is constrained by the underlying anatomy. Thus, it makes sense to use both intensity and location

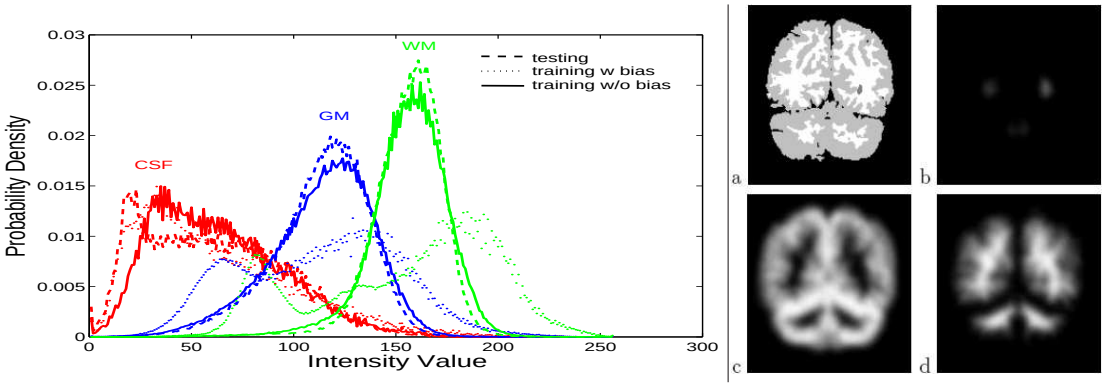


Fig. 1. MAP Classification Model: (left) Tissue intensity likelihood. The improved alignment of the training/testing distributions is an indication of the benefit of bias field correction (best viewed in color). **(right)** Probabilistic atlas example. **(a)** reference brain segmentation; **(b-d)** `CSF/GM/WM` probability maps.

features as the basis of our tissue models. In the Bayesian formalism (1) we use intensity as a likelihood and location as a prior (see Fig. 1).

Our intensity models are multi-modal and non-parametric since they take the form of simple histograms. This overcomes the unimodal limitations of single Gaussian [11, 17] and the inefficiencies of EM-based Gaussian mixtures [8–10], without loss of accuracy. Location information is also exploited by constructing a probabilistic atlas from our own training set. We randomly select a reference volume from the training set, and then affinely register all other volumes to the chosen one. The atlas is obtained by averaging and Gaussian smoothing the label maps of the registered brain volumes. Our model so far has incorporated intensity information and location prior. Next we show how to incorporate further features such as gradient, texture, and context in our discriminative framework.

2.3 Tissue Classification via Random Decision Forests

A random decision forest [4] is a collection of T deterministic decision trees which differ from each other due to random repartitions of training data. This is known to aid generalization accuracy — intuitively, where one tree fails the others do well. Furthermore, a decision forest provides posterior probabilities for labels, as opposed to hard labellings, by pooling votes across the population of trees.

Training. During training, each point $\mathbf{x}$ is associated with a known class label $L(\mathbf{x}) = \{$ `GM, WM, CSF` $\}$, and is pushed through each of the trees starting at the root. Each tree node applies a binary test of the form: $f(\mathbf{x}; \boldsymbol{\theta}) > \tau$ and sends the data to one of its two child nodes accordingly. $f(\cdot)$ is a function characterized by its parameters $\boldsymbol{\theta}$ and applied to the voxel $\mathbf{x}$. τ is a threshold. For now it suffices to say that f computes certain visual features on the point at hand. At training time the parameters $\boldsymbol{\theta}, \tau$ of each node and the tree structure are all optimized by minimizing the data information gain. Randomness in the trees arises from

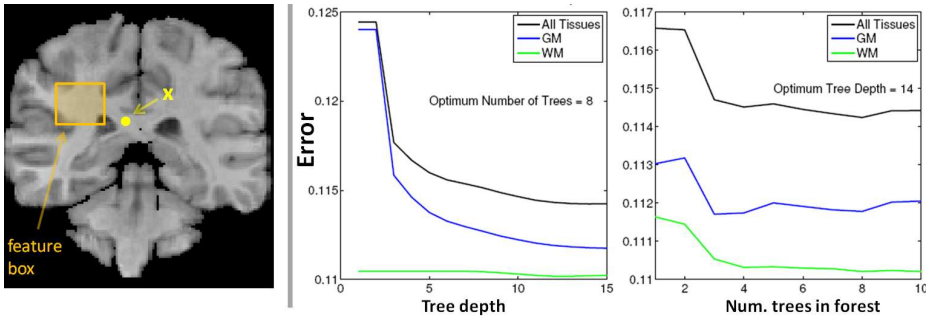


Fig. 2. Decision forests filters and results: (left) A shape filter used to provide context for the point $\mathbf{x}$ is computed from the feature box shown. **(middle-right)** Classification error — total (black), GM (blue), and WM (green) — as a function of the tree depth and the number of trees in the forest (best viewed in color).

noise being injected in the selection of the optimal node parameters. During training the leaf nodes update and store the empirical distributions over classes $P_{l_t(\mathbf{x})}(L(\mathbf{x}) = c)$, where l_t indexes the leaf node in the tth tree.
Testing. During testing each point $\mathbf{x}$ is pushed through each tree until it reaches a leaf node. The same input point $\mathbf{x}$ will end up in different leaf nodes, with different posterior probabilities. The output of the forest, for the point $\mathbf{x}$ is defined simply as the mean of all such posteriors: $P(L(\mathbf{x}) = c) = \sum_{t=1}^{T} P_{l_t(\mathbf{x})}(L(\mathbf{x}) = c) \ / \ T$. Now, a Maximum Likelihood classification for each voxel is obtained as: $c^\star = \arg\max_c P(L(\mathbf{x}) = c)$. Spatial prior could now be incorporated as before (1) but a more effective approach is described below.
Context-rich visual features. Here we use "shape filters" similar to the ones used in [5]; but applied to the 3D volume and without the need for "textonization". Fig. 2(left) illustrates these concepts on a 2D slice. For each voxel $\mathbf{x}$ a feature box F of random size and shape is selected at a random displacement from $\mathbf{x}$. The size of the feature box is selected between 1 and 30 voxels. The feature response is then defined as $f(\mathbf{x}; F) = \sum_{\mathbf{q} \in F} C_i(\mathbf{q})$ where C_i indicates different "channels". In particular, here we make use of the following five image channels: the raw intensities $C_1(\mathbf{x}) = I(\mathbf{x})$, the image gradient $C_2(\mathbf{x}) = |\nabla I(\mathbf{x})|$, the atlas-based probabilities $C_3(\mathbf{x}) = P(L(\mathbf{x}))$, the intensity likelihood $C_4(\mathbf{x}) = P(I(\mathbf{x})|L(\mathbf{x}))$ from the trained histograms, and the label posterior $C_5(\mathbf{x}) = P(L(\mathbf{x})|I(\mathbf{x}))$ output of the MAP classifier from Sect. 2.2. The ability of our features to look at a large distance from the center pixel $\mathbf{x}$ yields context-rich information. As illustrated in Fig. 2(middle-right), increasing the tree depth or the forest size tends to decrease the decision forest classification error.

2.4 Modeling Partial Volume Effects

The limited image resolution causes many voxels to contain material from multiple tissues, which is the main reason of misclassification. The goal of this section

is to locate such partial voxels $\Omega_m \subset \Omega$ in the image, as well as estimating the mixing fraction $\alpha : \Omega_m \mapsto [0, 1]$.

Modeling mixed tissue classes. Here we assume that partial voxels contain at most two different tissue types and we adopt a mixture model to capture the mixing effect as follows: $I(\mathbf{x}) = \alpha(\mathbf{x})I_1 + (1 - \alpha(\mathbf{x}))I_2$, where I_1, I_2 are the underlying tissue intensities and α the mixing factor. When considering partial volume effects, the tissue classification problem is modified by extending the set of class labels to the following: { `CSF, GM, WM, CSF/GM, GM/WM` } (the transition `CSF/WM` is ignored here as it occurs rarely in practice [17]). Since partial voxels usually occur at tissue boundaries, we identify them by labelling the voxels at each side of the boundaries as partial. Then we learn the models for the `CSF/GM` and `GM/WM` mixtures directly, via the same method described before for pure tissue modeling. This technique proves to work better than modeling the mixed tissues by mixing the models of the pure tissues (see Fig. 4d,e for comparison).

Mixing fraction estimation. Using the models described above we can now assign one of the five class labels to each voxel. Then, we estimate the mixing fraction α by maximum-likelihood: $\alpha^\star(\mathbf{x}) = \arg\max_{\alpha \in [0,1]} \log P(I(\mathbf{x})|\alpha)$. Since we conservatively consider both sides of the tissue boundaries to be partial voxels, the built partial volume classifier tends to underestimate pure voxels. Thresholding the mixing fraction, so that partial voxels with $\alpha(\boldsymbol{x}) \leq \delta$ or $\alpha(\boldsymbol{x}) \geq 1 - \delta$ are relabeled as pure, marginally improves labelling accuracy. This threshold is also learned from the validation set, and in practice we found $\delta = 0.1$ to work well.

3 Results and Validation

Our approach is validated on the Internet Brain Segmentation Repository [3], where 20 normal subjects of T1-weighted brain MR images with expert segmentation are available. The volume size is around $256 \times 256 \times 60$, with voxel resolution $1mm \times 1mm \times 3mm$. We compute voxel-wise classification accuracy and the associated standard error by running our measurements on different random training-validation-testing splits. In each run the forest classifier is employed for discriminative optimization. Our results are compared to the state-of-the-art in Fig. 3. We achieve nearly 40% improvement on `CSF`, 5% on `GM`, and parity on `WM`, compared with the best methods. Note that our results are close to the "ideal" score obtained by human experts (last row).

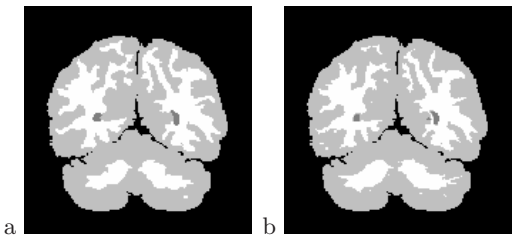
Next, we demonstrate how the test paradigm may further be improved by taking into account partial volume effects. In Fig. 4d we show the confusion matrix results of partial volume classification, and show that, relative to our own results of pure tissue classification, error can be reduced if partial voxels are labelled in datasets. We propose that this is the way brain tissue labelling algorithms should be evaluated in the future.

[3] http://www.cma.mgh.harvard.edu/ibsr/

[4] Dice index reported by [19], different from Jaccard index we use in Fig. 3. By definition, Dice > Jaccard. Thus for a fair comparison we also present here the mean Dice indices of our approach: `CSF` 0.699, `GM` 0.900, `WM` 0.831.

Method	CSF	GM	WM
Adaptive MAP	0.069	0.564	0.567
Biased MAP	0.071	0.558	0.562
Fuzzy c-means	0.048	0.473	0.567
Maximum-a-posteriori (MAP)	0.071	0.550	0.554
Maximum-likelihood	0.062	0.535	0.551
Tree-Structure k-means	0.049	0.477	0.571
MPM-MAP [11]	0.227	0.662	0.683
BSE/BFC/PVC [17]	—	0.595	0.664
Constrained GMM [8]	—	0.680	0.660
Spatial-varying GMM [9]	—	0.768	0.734
Coupled surface [14]	—	0.701	—
FSL [10]	—	0.756^4	—
SPM [18]	—	0.790^4	—
MAP with histograms	$\mathbf{0.549 \pm 0.017}$	$\mathbf{0.814 \pm 0.004}$	0.710 ± 0.005
Decision Forest Classifier	$\mathbf{0.614 \pm 0.015}$	$\mathbf{0.838 \pm 0.006}$	$\mathbf{0.731 \pm 0.007}$
Inter-rater consistency	—	0.876	0.882

Fig. 3. Comparison of our approaches with the state of the art. Mean and std. error of Jaccard indices are obtained from repeated random runs. We are targeting CSF/GM/WM segmentation only, but note that [17] also classifies the background.



a b

c

78	19	3
0	94	6
0	17	83

d

90	2	0	8	0
0	85	1	0	14
0	1	86	0	13
16	26	1	45	12
0	19	13	0	68

e

86	2	0	12	0
0	84	1	0	15
0	1	86	0	13
13	32	0	41	14
0	19	13	0	68

Fig. 4. Segmentation results: (a) Ground-truth with black-gray-white corresponding to CSF-GM-WM; **(b)** Label map obtained by our approach. **(c-e)** Confusion matrices (in %) for tissue classification. **(c)** without partial volume classification; **(d)** modeling partial voxels by direct histogram learning; **(e)** modeling partial voxels by uniform mixture of pure voxels. Matrix rows (top-bottom) correspond to ground-truth, while columns (left-right) are our labelling, both in CSF, GM, WM, CSF/GM, GM/WM order. (d) yields best results, i.e., maximal overlap averaged on 3 pure tissue classes.

4 Conclusions

We have proposed a learning-based method combining bias field correction, histogram tissue likelihood, and atlas based prior, with a decision forest classifier that uses context, to achieve substantial improvements on tissue labelling of brain MR images. Performance obtained is now very close to that of expert practitioners. We also showed that further improvements could be obtained in classification error performance by taking account of partial volume effect, and this suggests a modified test paradigm for future studies.

References

1. Styner, M., et al.: Morphometric analysis of lateral ventricles in schizophrenia and healthy controls regarding genetic and disease-specific factors. In: Proc. National Academy of Science. (2005) 4872–4877
2. Fischl, B., et al.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. NeuroImage **33** (2002) 341–355
3. Maintz, J., Viergever, M.: A survey of medical image registration. Medical Image Analysis **2** (1996) 1–37
4. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural Computation **9** (1997) 1545–1588
5. Shotton, J., Winn, J., Rother, J., Criminisi, A.: Textonboost: joint appearance, shape, and context modeling for multi-class object recognition and segmentation. In: Eur. Conf. Computer Vision. (2006) 1–15
6. Pham, D., Prince, J.: Adaptive fuzzy segmentation of magnetic resonance images. IEEE Trans. Medical Imaging **18** (1999) 737–752
7. Ramon, J., Veronica, M., Oscar, Y.: Data-driven brain MRI segmentation supported on edge confidence and a priori tissue information. IEEE Trans. Medical Imaging **25** (2006) 74–83
8. Greenspan, H., Rurf, A., Goldberger, J.: Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. IEEE Trans. Medical Imaging **25** (2006) 1233–1245
9. Peng, Z., Wee, W., Lee, J.: Automatic segmentation of MR brain images using spatial-varying Gaussian mixture and Markov random field approach. In: Proc. Conf. Computer Vision and Pattern Recognition Workshop. (2006) 80–87
10. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov model and the expectation-maximization algorithm. IEEE Trans. Medical Imaging **20** (2001) 45–57
11. Marroquin, J., Vemuri, B., Botello, S., Calderon, F., Fernandez-Bouzas, A.: An accurate and efficient Bayesian method for automatic segmentation of brain MRI. IEEE Trans. Medical Imaging **21** (2002) 934–945
12. McInerney, T., Terzopoulos, D.: Deformable models in medical image analysis. Medical Image Analysis **1** (1996) 91–108
13. Yang, J., Staib, L., Duncan, J.: Neighbor-constrained segmentation with level set based 3D deformable models. IEEE Trans. Medical Imaging **23** (2004) 940–948
14. Zeng, X., Staib, L., Schultz, R., Duncan, J.: Segmentation and measurement of the cortex from 3D MR images using coupled-surface propagation. IEEE Trans. Medical Imaging **17** (1998) 74–86
15. Tu, Z., et al.: Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Trans. Medical Imaging **27** (2008) 495–508
16. Vovk, U., Pernus, F., Likar, B.: A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans. Medical Imaging **3** (2007) 405–421
17. Shattuck, D., Sandor-Leahy, S., Schaper, K., Rottenberg, D., Leahy, R.: Magnetic resonance image tissue classification using a partial volume model. NeuroImage **13** (2001) 856–876
18. Ashburner, J., Friston, K.: Multimodal image coregistration and partitioning - a unified framework. NeuroImage **6** (1997) 209–217
19. Tsang, O., et al.: Comparison of tissue segmentation algorithms in neuroimage analysis software tools. In: IEEE Engineering in Medicine and Biology Society. (2008) 3924–3928