30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

COMPARISON OF TISSUE SEGMENTATION ALGORITHMS IN NEUROIMAGE ANALYSIS SOFTWARE TOOLS

On Tsang, Ali Gholipour, Nasser Kehtarnavaz, *Senior Member, IEEE,*
Kaundinya Gopinath, Richard Briggs, and Issa Panahi, *Senior Member, IEEE,*

***Abstract*— Accurate segmentation of different brain tissues is of much importance in magnetic resonance imaging. This paper presents a comparison of the existing segmentation algorithms that are deployed in the neuroimaging community as part of two widely used software packages. The results obtained in this comparison can be used to select the appropriate segmentation algorithm for the neuroimaging application of interest. In addition to the entire brain area, a comparison is carried out for the subcortical region of the brain in terms of its gray matter composition.**

I. INTRODUCTION

THE field of medical imaging has evolved to the point that a wide variety of sophisticated tools is now available to image the human brain in vivo. In particular, the increase in the spatial resolution of magnetic resonance imaging (MRI) and functional MRI (fMRI) has been allowing physicians and scientists to better utilize the functional imaging of different parts of the brain towards prescribing effective therapies.

In applications where functional localization is of importance, knowing the precise location of a particular brain structure is a prerequisite to successful treatment. For instance, analysis prior to brain surgery is done by experts who examine images of the brain and perform a manual segmentation of the structure of interest. However, there exists some disagreement and variability between different independent experts who perform manual segmentation on such images [1], indicating that there is still room for improvement in the segmentation process.

To address inter-expert variability, various algorithms have been introduced in the image processing literature to handle MR image segmentation. A review of such algorithms has been compared by Liew et al [2]. The challenge in tissue segmentation now lies in having a robust classification approach based on image intensity values representing cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM).

At present, two of the most widely used software packages in the neuroimaging community contain automated segmentation routines that utilize iterative approaches to classify brain images into the three tissue classes. These packages are: SPM5 [3], written by the Wellcome Department of Imaging Neuroscience at University College London, UK, and version 4 of FMRIB Software Library (FSL) [4], written by the Analysis Group, FMRIB, Oxford, UK. This paper will provide a quantitative analysis and comparison of the segmentation algorithms in SPM5 and FSL for the whole brain, as well as for the subcortical region, utilizing the new subcortical structure extraction tool from FSL (named FIRST) as a gold standard for the comparison.

Although previous studies have provided comparisons of tissue segmentation algorithms [5][12][12][14] including the ones from SPM and FSL, our comparison utilizes the latest versions of the software packages, a variety of independent datasets, and the most widely used metrics in the literature. Such a comprehensive analysis of this scope has not been previously carried out and can be quite useful to current and future users of these software packages.

The paper is organized as follows. Section II gives a brief summary of the two software packages and their respective segmentation routines. A description of the datasets used for the comparison is provided in section III. The comparison methodology appears in section IV while results and a discussion of the results are found in section V. The conclusion and future work are stated in sections VI and VII, respectively.

II. SEGMENTATION ALGORITHMS

A. *SPM5 - SEGMENT*

The segmentation routine in SPM5 is based on a unified segmentation model that performs tissue segmentation, bias correction, and spatial normalization all in the same model [6] to address the inherent circularity in the voxel-based morphometry (VBM) analysis technique used by researchers. The Expectation-Maximization (EM) algorithm is used to obtain the optimum parameters corresponding to a mixture of Gaussians model representing the tissue classes. Tissue probability maps are used to perform an affine transformation of the brain anatomy to standard International Consortium of Brain Mapping (ICBM) / Montreal Neurological Institute (MNI) space after which the segmentation is performed. Preprocessing of the anatomical brain volumes is not typically required.

Manuscript received April 16, 2008. This work was supported by the Erik Jonsson School of Engineering and Computer Science at the University of Texas at Dallas under the Erik Jonsson Distinguished Assistantship.

O. Tsang, A. Gholipour, N. Kehtarnavaz, and I. Panahi are with the Electrical Engineering Department, University of Texas at Dallas, 800 W. Campbell Rd., Richardson, TX 75080 USA (e-mail: kehtar@utdallas.edu).

K. Gopinath and R. Briggs are with the Dept. of Radiology, University of Texas Southwestern Medical Center, Dallas, TX 75390 USA.

B. FSL - FAST

The FMRIB Automated Segmentation Tool (FAST) is part of the FSL library. The segmentation routine used in the tool is based on a Hidden Markov Random Field (HMRF) model that is optimized using the Expectation-Maximization algorithm [7]. The HMRF model is more robust than similar finite mixture models because it takes into account spatial information in terms of mutual information in local neighborhoods and uses this information to aid in the segmentation of the tissue classes.

The brain volume is first registered to the standard MNI space using apriori tissue probability maps. FAST then segments the brain volume into the three tissue classes and performs bias correction. The required input to this program is a skull-stripped version of the anatomical image and users are expected to provide this information using the FSL's own brain extraction tool or using a third party tool. In our analysis, the `brainmask` utility from the Neuroimage Processing Toolkit (NPTK) 0, developed in the Signal and Image Processing Lab at the University of Texas at Dallas, is used after the skull stripping. In a previous work, the `brainmask` utility was found to augment the output of the brain extraction tool for a more complete image of the brain [9].

C. FSL – FIRST

The FMRIB Integrated Registration and Segmentation Tool (FIRST) is a relatively new addition to the FSL suite [10]. This addition addresses subcortical structures of the brain which are characterized by parameters of surface meshes and point distribution models located in a database. The database was constructed with the help of manually segmented data by the Center of Morphometric Analysis at Massachusetts General Hospital in Boston. T1-weighted images passed to this tool are matched to the database of subcortical structures and the most probable structure is extracted based on the shapes in the image. Specific structures can be extracted as specified by the user.

III. DATA

For the analysis and comparison of the two algorithms, three different datasets were used. The first dataset corresponded to the digital phantom brain provided by the McConnell Brain Imaging Center under the Brainweb module [11]. The advantage of using the Brainweb dataset in the comparison was the availability of a ground truth for the tissue classes (CSF, GM, and WM) from which the digital phantoms were created.

The second dataset consisted of real data acquired from a Siemens Trio 3T magnet of 32 normal subjects with the following two imaging contrasts: T1-weighted spin echo (SE) MRI and T2-weighted turbo-spin echo (TSE) MRI. This protocol specified the acquisition of T1-weighted anatomical MRI at a spatial resolution of 1.0x1.0x1.2 mm^3 and the structural T2-weighted TSE axial scans at a spatial resolution of 0.9×0.9×3.0 mm^3. This dataset, which shall be referred to as the "real dataset" from this point forward, had no ground truth available.

The final dataset used for comparison was the Internet Brain Segmentation Repository (IBSR) provided by Massachusetts General Hospital (MGH) in association with Harvard Medical School. This dataset consisted of 20 normal MR brain image sets with their manual segmentations provided by the Center for Morphometric Analysis at MGH (http://www.cma.mgh.harvard.edu/ibsr/). The manual segmentations included in the dataset had been performed by the experts at MGH and are thus used as the ground truth for validation.

IV. COMPARISON METHODOLOGY

For the Brainweb and IBSR datasets, which included ground truth segmentations, the misclassification rate (MCR) metric [12][13] as well as the sensitivity and specificity metrics were used to assess tissue classification. For the real dataset, the Dice overlap metric [12][14] was used to measure the similarity of the segmented results across the two image types. This cross-validation provided a simple yet effective way to compare the consistency of the segmentation algorithms between different modalities. To generate the data for the MCR and the other metrics, the command line tool `3dOverlap` from the AFNI [15] software was used. This tool performed a voxel by voxel volumetric overlap of two images to determine the total number of voxels that were within their intersection.

The output of the FIRST tool was used to provide a "gold standard" for evaluating the performance of SPM5 and FSL. The segmented gray matter tissue was first found for each subject using SPM5 and FAST. FIRST was subsequently used to extract the subcortical structures from the anatomical image of each subject. Masking the GM tissue volumes obtained from SPM5 and FSL with the subcortical structures labeled by FIRST, a corresponding GM composition of each structure was obtained. Using `3dOverlap`, the number of voxels was counted to calculate the percentage GM composition.

Of the subcortical structures defined in FIRST, all except the brain stem and the lateral ventricles were considered to be gray matter. The brain stem consists of both gray and white matter, and was omitted from the analysis. The lateral ventricles are filled with CSF and devoid of gray matter. The structures that were compared included the nucleus accumbens, amygdale, caudate, hippocampus, pallidus, putamen, and thalamus from both hemispheres of the brain.

It is worth noting a mechanistic challenge encountered in the implementation of this analysis. The spatial normalization of the IBSR dataset proved problematic for the automated unified segmentation algorithm in SPM5. The algorithm was unable to perform a correct initial affine transformation matching the anatomy to the probability maps. To circumvent this issue, the input image had to be reoriented by centering the image on the anterior commissure prior to segmentation.

	Sensitivity	Specificity	MCR
FAST	89.6%	98.7%	10.4%
SPM	90.8%	98.8%	9.1%

Table 1 – Brainweb Dataset: Average metric values for whole brain

This manual reorientation of the input image makes the process less automated. The reason the registration fails for these images is because the tissue probability maps contain no extracortical information and hence the registration model is suboptimal for the images having this information [16]. FAST's normalization routine does not suffer from the same problem despite using similar tissue probability maps indicating that perhaps the FMRIB Linear Image Registration Tool (FLIRT), which is used as part of the FAST routine, is more robust in handling images of different orientations than the corresponding registration algorithm that is included in the SPM segmentation tool.

V. RESULTS AND DISCUSSION

A. *Brainweb*

The Brainweb dataset, while consisting of a single digital phantom, comes with different simulation options pertaining to the amount of noise and the amount of RF inhomogeneity in the simulated image. To verify our analysis, we compared Dice overlap measures with other results from the literature for this particular dataset and obtained similar values [6].

Interestingly enough, the best results for both software packages were obtained when using simulated images with a moderate amount of Rician noise present, as opposed to images without any noise (see **Fig 1**). This is also noted in a recent publication by Ferreira da Silva [17], and bodes well for other real datasets since imaging noise is an inevitable part of image acquisition. Also note that the RF inhomogeneity parameter does not significantly alter the results, as indicated by the curves in **Fig 1,** due to the bias correction algorithms that are implemented within the segmentation tools.

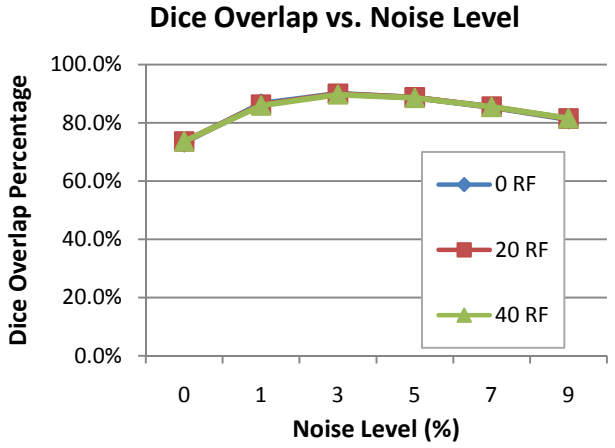

Fig 1 – Dice overlap metric versus noise level in the Brainweb dataset. Note: Different RF inhomogeneity levels negligibly affect outcomes. Graph shows SPM5 segmentation; FAST shows a similar trend.

	GM	WM	CSF
FAST	73.9%	79.5%	60.4%
SPM	75.1%	81.9%	42.1%

Table 2 – Real dataset: Average Dice overlap metric for 3 tissue classes Approximately 5 outliers were removed from the set of 32.

The lower aggregate MCR metric of SPM5 (see **Table 1**) indicates that it performed a more accurate segmentation of the Brainweb volumes, on average, as compared to FAST. The values in **Table 1** are the average metric values of the CSF, GM, and WM tissue classes of the Brainweb phantom for four different noise percentages: 1, 3, 5 and 7% noise at 20% RF inhomogeneity. The noise percentage is representative of the percent ratio of the standard deviation of the white Gaussian noise versus the signal. Higher sensitivity and specificity metrics indicated that the segmentation algorithm in SPM5 correctly identified more tissue voxels than did the FAST algorithm, and also was better at rejecting tissue voxels that were not related to the tissue class of interest.

These three metrics all support the finding that SPM5, with its segmentation algorithm based upon the Gaussian Mixture Model, provides a slightly better segmentation output than the HMRF model used in FAST (see **Fig 2**). The default parameters were used for SPM5's segmentation algorithm, whereas in FAST, it was found that of the various segmentation options, the best performance was obtained using MNI152 tissue probability maps for a priori estimation of initial parameters, option (-a) with a tissue probability map output option (-p). The outputs then had a threshold of 50% applied to them to obtain the final binary image results for each tissue class.

B. *Real Data*

Although a ground truth against which to verify the outputs of the segmentation algorithms was not available for the real dataset, this dataset contained images of the subjects in both the T1-weighted and T2-weighted contrasts. The segmented results of these two images were

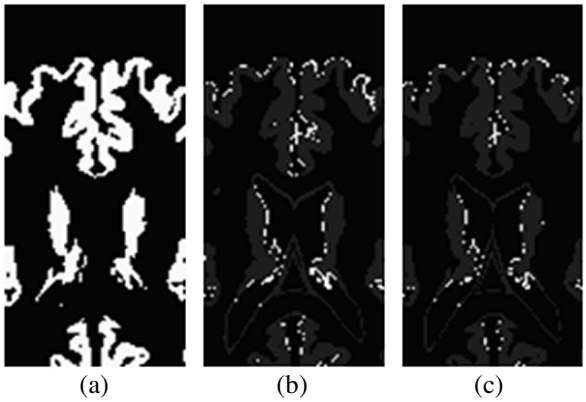

Fig 2 – Axial view of the Brainweb dataset. The GM ground truth shown in white (a), the output of FAST overlaid on the ground truth (b), and the output of SPM5 overlaid on the ground truth (c). Note more white pixels are showing through in (b) than (c) indicating that SPM5's segmentation is more accurate and has less disparity with the ground truth than FSL.

thus used for cross-validation to determine which segmentation algorithm was more consistent. In this segmentation, the best parameters as determined previously with the phantom brain were used.

The results of the segmentation show that the SPM5 GMM based segmentation algorithm was generally more consistent than the FAST algorithm, with Dice overlap measures between T1 and T2 segmented images consistently higher with the exception of the CSF tissue class (see **Table 2**). A Dice overlap measure of 70% or greater indicates a high level of agreement between the two images as stated by Bartko [18].

The large difference in Dice values for the CSF tissue class between the two software packages is surprising but probable since CSF accounts for a small portion of total brain matter. Even slight variations in the algorithms' segmentation of CSF can yield large overlap errors. However, it does appear that SPM5's algorithm favored GM/WM accuracy at the expense of CSF accuracy here.

C. IBSR Data

As noted earlier, segmentation of this dataset using the SPM5 package presented some challenges. The images had to be reoriented prior to segmentation; otherwise, a poor segmentation would result (see **Fig 3**). This problem did not afflict the FAST algorithm, even though it also uses similar tissue probability maps. This may indicate that the spatial normalization tool in the FSL package is more robust.

However, the results of the whole brain segmentation favored SPM5's segmentation algorithm with lower MCR values and higher sensitivity and specificity values when overlaid against the gold standard of the expert segmented tissue images (see **Table 3**). Comparing the results of this dataset against the results of the Brainweb dataset showed that the performance of the SPM5 algorithm was consistent for both datasets. MCR, sensitivity, and specificity values were consistent to the stated results.

Note that the MCR values for the IBSR dataset were quite high but were similar to what Pham et al. [19] obtained using the same dataset. This was somewhat surprising but also understandable since the software packages were optimized using Brainweb during their development. Therefore, they may perform very well on the Brainweb dataset as compared to the IBSR dataset. The large MCR values could also be due in part to the results of the expert segmented images differing from the actual anatomy.

D. Subcortical Region

The subcortical region of the brain has always been a challenge to tissue segmentation algorithms. Although the structures of the region are strictly GM, segmentation algorithms tend to incorrectly classify them as WM. The two software packages' automated segmentation algorithms were tested in the subcortical region to determine their performance in this area. Fortunately, with the ground truth available for both the Brainweb and IBSR datasets, it was possible to compute the absolute performance of the algorithms in terms of their ability to correctly classify the subcortical structures as GM (see **Table 4**) using FIRST's structure label extraction.

From the table, we see that for both the Brainweb and the IBSR datasets the SPM5 segmentation algorithm yields the highest overlap with the ground truth GM tissue maps for both the subcortical area and the whole brain. The real dataset was omitted from the subcortical analysis since the results could not be validated. Again, note the prevalent higher values for Brainweb versus the IBSR dataset.

	Sensitivity	Specificity	MCR
FAST	74.6%	98.6%	25.4%
SPM	78.0%	98.8%	22.0%

Table 3 – IBSR Dataset: Average metric values for whole brain

VI. CONCLUSION

This paper has presented a comparison of the tissue segmentation algorithms of two open-source MRI analysis packages that are currently widely used by researchers: SPM5 and FSL version 4. Three datasets were used to perform the comparison and a number of widely used metrics were used to determine the efficacy of the respective segmentation algorithms. An evaluation of the performance specifically in the subcortical region was also performed with the new subcortical label extraction tool from FSL.

While both algorithms performed quite satisfactorily, the segmentation algorithm in SPM5 was found on average to be more accurate, consistent, and robust than the algorithm implemented in FSL for whole-brain segmentation in our experiments. With regards to the subcortical region, the datasets which had a ground truth available for validation were compared and the SPM5 algorithm was again found to perform better. Therefore, based on the results obtained in this analysis, our recommendation to researchers is to try the segmentation algorithm found in the SPM5 software package in order to see how the outcome compares to their software package of choice for the subcortical region or the whole brain.

Aside from the comparison results that were computed in this analysis and the significance of those results, we believe that the validation methods that were presented here may also be of use to other researchers that seek to repeat these experiments for themselves or to validate their own results.

	Brainweb	IBSR
FAST	92.1%	69.1%
SPM	92.9%	73.7%

(a)

	Brainweb	IBSR
FAST	89.2%	75.6%
SPM	91.5%	79.0%

(b)

Table 4 – Dice overlap measures of the GM composition of the subcortical brain structures (a) and also of the whole brain (b). Values are the computed averages of the datasets.

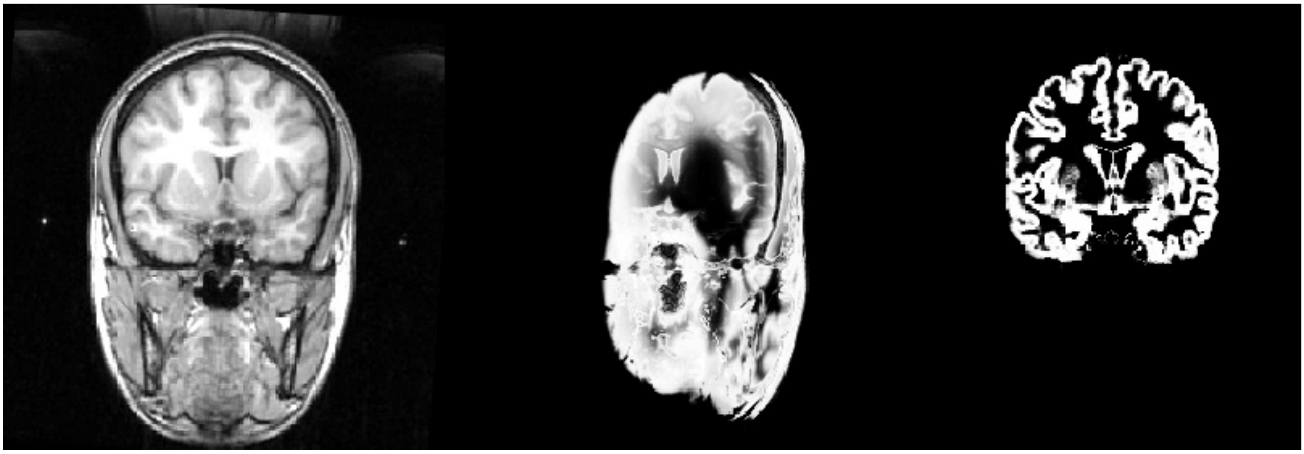

Fig 3 – Example of incorrect segmentation of IBSR data using SPM5. The original image (left), the automated segmentation result for GM (middle), the segmentation output after manually reorienting the data to the approximate orientation of the tissue probability maps (right).

VII. FUTURE WORK

It should be noted that the subcortical GM comparison was performed with the assumption that the output of the FIRST tool was the "gold standard" against which to validate the results against. As this was the initial version of the FIRST tool, subsequent versions should provide more accurate label extracted subcortical structures which would then make such comparisons in the subcortical region more reliable and would lay the foundation of future work.

VIII. REFERENCES

[1] Warfield, S.K., Zou, K.H., Wells, W.M., "Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 23(7), pp. 903-921, 2004.

[2] Liew, A.W.C., Yan, H., "Current Methods in the Automatic Tissue Segmentation of 3D Magnetic Resonance Brain Images," *Current Medical Imaging Reviews*, vol. 2(1), pp. 91-103, 2006.

[3] Ashburner, J., Friston, K.J., Poline, J., et al., "Spatial registration and normalization of images," *Human Brain Mapping,* vol. 2, pp.165– 189, 1995.

[4] Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M. and Matthews, P.M., "Advances in functional and structural MR image analysis and implementation as FSL," NeuroImage, vol. 23(S1), pp. 208-219, 2004.

[5] 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*, vol. 24, pp. 790-795, 2006.

[6] Ashburner, J., Friston, K.J., "Unified Segmentation," *Neuroimage,* vol. 26(3), pp.839–851, 2005.

[7] Zhang, Y., Brady, M., Smith, S., "Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm," *IEEE Trans. on Medical Imaging*, vol. 20(1), pp 45-57, 2001.

[8] Gholipour, A., Kehtarnavaz, N., Gopinath, K., et al., "A Software Tool for Registration Based Distortion Correction in Echo Planar Imaging," *Proc. IEEE Dallas Eng. in Med. Biology Workshop*, Dallas, Texas, Nov. 11-12, 2007.

[9] Tsang, O., Gholipour, A., Kehtarnavaz, N., Gopinath, K., and Briggs, R., "Comparison of brain masking techniques in functional magnetic resonance imaging," *Proceedings of IEEE EMBS Dallas Workshop*, pp. 78-81, Nov 2007.

[10] Patenaude, B., Smith, S., Kennedy, D., et al., "FIRST - FMRIB's integrated registration and segmentation tool," *Human Brain Mapping Conference*, 2007.

[11] Kwan, R.K.-S., Evans, A.C., Pike, G.B., "MRI simulation-based evaluation of image-processing and classification methods," *IEEE Transactions on Medical Imaging*, vol. 18(11), pp 1085-97, 1999.

[12] Huang, A., Abugharbieh, R., Tam, R., et al., "Automatic MRI Brain Tissue Segmentation Using a Hybrid Statistical and Geometric Model," *3rd IEEE International Symposium on Biomedical Imaging*, pp. 394-397, 2006.

[13] Song, T., Jamshidi, M.M., Lee, R.R., Huang, M., "A Modified Probabilistic Neural Network for Partial Volume Segmentation in Brain MR Image," *IEEE Transactions on Neural Networks*, vol. 18(5), pp. 1424-1432, 2007.

[14] Bouix, S., Ungar, L., Dickey, C.C., et al., Evaluating Automatic Brain Tissue Classifiers," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*, vol. 3217, pp. 1038-1039, 2004.

[15] Cox, R.W., "AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages," *Comp & Biomedical Research*, vol. 29, pp.162-173, 1996.

[16] SPM Message Board Archives, January 30, 2008. http://www.jiscmail.ac.uk/cgi-bin/wa.exe?A2=ind0801&L=SPM&P=R68695

[17] Ferreira da Silva, A.R., "A Dirichlet Process Mixture Model for Brain MRI Tissue Classification," *Medical Image Analysis*, vol. 11(2), pp. 169-182, 2007.

[18] Bartko, J.J., "Measurement and reliability: statistical thinking considerations," *Schizophr. Bull.* vol. 17, pp. 483–489, 1991.

[19] Pham, D.L., Prince, J.L., "Robust unsupervised tissue classification in MR images," *IEEE Int'l Symposium on Biomedical Imaging: Macro to Nano*, vol. 1, pp. 109 – 112, 2004.