3D Segmentation In The Clinic: A Grand Challenge II at MICCAI 2008 - MS Lesion Segmentation

Release 0.00

Mark Scully[1], Vincent Magnotta[2], Charles Gasparovic[3], Peter Pelligrino[2], Delia Feis[1], and H. Jeremy Bockholt[1]

July 14, 2008

[1]Mind Research Network, Albuquerque, NM
[2]Department of Radiology, University of Iowa
[3]Department of Psychology, University of New Mexico

Abstract

This manuscript examines the application of a new parametric method to the segmentation of MS lesions from structural brain MRI images (sMRI). The data used in this work was provided by the MS Lesion Segmentation Challenge at MICCAI 2008. The method uses the vector image joint histogram, built over a training set, as an explicit model of the feature vectors indicating lesion. The histogram is used to predict lesions in the test data by labeling feature vectors consistent with lesion feature vectors in the training set. The results are evaluated using STAPLE to compare against two separate human raters.

Latest version available at the Insight Journal [http://hdl.handle.net/1926/130]

Contents

Lesion segmentation in magnetic resonance imaging (MRI) presents novel challenges compared to general tissue segmentation. Lesions have been found to present themselves with somewhat homogeneous signal across image contrasts such as ischemic lesion cases. However, lesions in Multiple sclerosis (MS) patients present additional challenges in that they are found to be heterogeneous in itensity variation across image contrasts e.g. enhancing lesions, black holes, etc. It is assumed that the relationships between the intensities of the different sequences, along with the spatial information, provide the necessary information to accurately segment the lesions. The efforts presented here represent an attempt to construct a parametric model of the intensity feature vectors that indicate the presence of MS lesions in the training set and then use that model to predict the location of lesions in the testing data.

1 Data

The MS lesion MRI image data provided for the contest was composed of 54 brain MRI images and represents a range of patients and pathology. Two groups of data were provided consisting of 20 training MRI images and 25 testing images. The image data was acquired separately by Children's Hospital Boston (CHB) and University of North Carolina (UNC). UNC cases were acquired on a Siemens 3T Allegra MRI scanner with slice thickness of 1mm and in-plane resolution of 0.5mm. The data supplied was rigidly registered to a common reference frame and resliced to isotropic voxel spacing using b-spline based interpolation. Multiple sequences were provided, including T1, T2, FLAIR, and DTI.

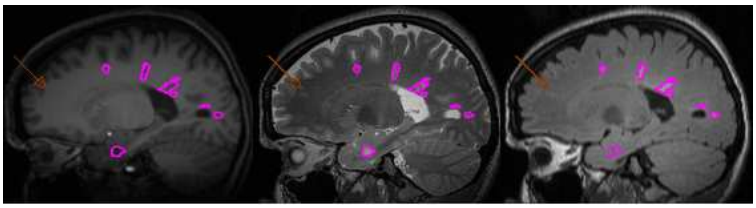

Figure 1: T1, T2, FLAIR, lesion tracing and unlabeled lesion.

2 Data Quality

There were numerous problems with data quality in the training and test data sets. There were a number of artifacts in the scans, including poor motion correction, poor registration, and large variations in the signal to noise ratio. There was also a lack of agreement between experts on lesion location. There were many instances where one expert considered something a lesion that another expert did not label. There were also instances of lesions being inside ventricles, as well as connected lesions being labeled as multiple

separate lesions. Supervised learning approaches are sensitive to the quality of the labeled training set and consequently all of the data quality issues contributed to the reduced accuracy of the parametric method presented here.

3 Preprocessing

Lesion segmentation has been shown to be highly susceptible to variations in field strength both within and across scans [3, 4, 7]. A number of approaches to correct for these problems have been examined in the literature but normally the steps taken involve bias correction and intensity standardization [7].

In order to make working with the images more ameable to consistent tissue classification they were first down sampled to 1mm isotropic resolution. Brain extraction was performed on each subject using BET2 [9] on the down sampled T1 and T2. Samples were taken of the grey, white, and CSF tissues of each subject in order to provide the mean values of each tissue type for bias correction. To reduce variation due to field strength inhomogeniety, bias correction was then performed on each down sampled T1, T2, and FLAIR using methods presented by Styner [2]. In order to reduce inter-subject variability intensity standardization was performed via histogram matching. Histogram matching was accomplished by linearly scaling each bin of the histogram, independently for each image type, to match a reference exemplar [8].

4 Model Building

The overall goal of parametric approaches is to construct a model of MS lesions that has predictive power based on some feature set. The training data consisted of T1, T2, FLAIR, FA, and MD images. While it would have been interesting to extend the analysis to DTI data, only the T1, T2, and FLAIR data was included in this analysis. The model of MS lesion tissue is thus dependent on the intensities of the T1, T2, and FLAIR images along with the spatial information embodied by those voxels in the neighborhood of a given lesion voxel. Parametric methods have shown good results in the past [6], as have spatial methods [3], but stronger results have been shown by combining the two approaches [5]. The motivation behind the approach in this paper was to construct an explicit model of the feature vector intensities that identify lesion tissue. One such model is the joint histogram calculated over the vector image constructed by registering the T2 and FLAIR to the T1, then creating an image where each voxel contains a vector representing the T1 intensity, the T2 intensity, the FLAIR intensity, and the tissue class for that voxel's location. The joint histogram then represents the number of times a given feature vector was labeled lesion, along with the number of times that vector was not labeled lesion. By adding the joint histograms calculated this way for each subject a combined histogram could be created for the entire dataset.

In order to make the histogram calculation more computationally tractable, the T1, T2, and FLAIR values for all training and test subjects were quantized into fifteen bins of width one half standard deviation of the intensities, calculated on a per image basis. There are essentially four tissue class intensity distributions to identify and those distributions are know to overlap some and to have widths well above 0.5 sigma. Thus it can be assumed that quantizing does not grossly distort the data, although it may cause errors in relation to partial volume voxels.

The T1, T2, FLAIR, and lesion label were then treated as a multi-component / vector image and the 4D joint histogram was calculated. For each possible combination of T1 value, T2 value, and FLAIR value the joint histogram contained the number of times that feature vector resulted in a lesion and the number of times it did not. By dividing the number of times a feature vector was labeled lesion by the number of times the

feature vector occurred, the frequency with which that feature vector resulted in lesion was obtained.

5 Prediction

An initial predicted lesion label map was constructed for each test subject by setting the voxel value equal to the frequency of that voxel location's feature vector in the 4D joint histogram model. These values were then thresholded in order to reduce the amount of false positives while keeping the number of false negatives low. The threshold value was selected empirically and could likely be improved through the application of a more rigorous experimental process.

In order to obtain a rough tissue segmentation, KMeans, with initial class means, was performed on the T1 image for each subject. This label map was combined with the thresholded lesion mask to create a label map with different integer values for Grey, White, CSF and Lesion tissue. The generated label map was then filtered so that lesion tissue that was within 2 voxels of CSF tissue was discarded. The label map was then used to pull 200 exemplar points to train a naive Bayesian classifier which then proceeded to classify the vector image composed of the T1, T2, and FLAIR images.

Each independent connected component in the lesion map produced by the Bayesian classifier was then filtered based on a minimum lesion size provided by a local MS expert. Each lesion component had to have a least one dimension with three voxels with the other two dimensions being at least two voxels. While lesions may occur below this size they are generally not labeled by human experts due to high error rates.

In order to submit the data for evaluation the lesion map was upsampled to 0.5mm by 0.5mm by 0.5mm.

6 Results

Table 2 contains the prediction results by comparison to two human raters using the STAPLE algorithm [10]. A score of 95 is equivalent to the performance of another human rater. The overall performance was poor, as can be seen by the total scores and the PPV value under STAPLE. As can be seen from table 2 the method presented here had reasonably good specificity, or true positive rate, and a fairly poor sensitivity, or false negative rate, as was expected based on the initial work. The table also indicates that for some subjects the methods were highly effective while for others they were extremely poor. This could be due to variances in the volume of lesions between subjects. The table also illustrates the variability between the two raters, as evidenced by the large difference between the sets of scores for each site.

7 Discussion

Everything presented here is confounded by the lack of ratings performed by each expert on the entire dataset. Since only UNC cases were rated by the UNC rater and only CHB cases were rated by the CHB rater, it was impossible to examine lesions that both raters agreed on.

Initially the lesions as predicted by the joint histogram model were meant to act as a training set for a Support Vector Machine (SVM). However, when trained on only those feature vectors that had some chance of being lesions the SVM failed to converge even after 48 hours. The failure to converge may be due to there being no optimal decision boundary between the classes, which is implied by the joint histogram data.

The joint histograms are explicit models of the feature vectors that were labeled as lesions by each site's

	CHB	UNC
Number of features	991	481
Intersection	420	420
Union	1052	1052
Jaccard Overlap	0.3999	0.3999
Dice Metric	0.5707	0.5707
Spatial Overlap	0.4238	0.8732

Table 1: Feature Vector Overlap Between Raters

experts. By generating histograms for the UNC and CHB training sets independently the difference in raters can be examined. After thresholding the histograms to form masks various overlap metrics can be calculated. These metrics convey the agreement between the two site's raters on what feature vectors actually represent lesion.

As can be seen from table 1 there exists significant disagreement between the raters from the two sites on what feature vectors indicate lesion. The CHB rater consistently rated more feature vectors as lesion but there were still feature vectors the UNC rater felt were lesion that the CHB rater did not. The fact that the Jaccard Overlap is less than 0.4 indicates extreme differences in the feature vectors considered lesion. It is difficult to believe any automated technique would be able to correctly label all lesions in this situation.

In an ideal world the feature vectors of lesions would never be labeled anything other than lesion. If that were the case, lesion segmentation could be accomplished simply by labeling those feature vectors as lesion. Unfortunately that is not the case in the MS training set. Figure 2 lays out the relationship between the number of times a feature vector was labeled lesion along with the number of times that feature vector occurred. The vectors have been sorted to ascend by frequency of lesion occurrence, calculated as number of times the feature vector was lesion / number of times the feature vector occurred.

One of the things figure 2 expresses is that the majority of lesion feature vectors, 616 out of 1052, were labeled lesion less than ten times for the entire training set. This indicates that over half of the feature vectors indicating lesion are unique, represent collectively only 1925 out of the 114634 voxels labeled as lesion, and most likely can not be labeled lesion without taking into account spatial information. Alternately, it could also indicate that these particular feature vectors are not lesion, and thus represent noise in the expert ratings.

The main point embodied by figure 2 is that the number of times a given feature vector is labeled as lesion is generally dwarfed by the number of times it is not labeled lesion. The feature vector with the highest frequency of being labeled lesion occurred 47 times in the training data set and was labeled lesion only 19 times. If this feature vector was always labeled lesion it would be wrong 60% of the time! The feature vector representing the most lesion voxels, 3687, occurred a total of 195,344 times and so was lesion less than 2% of the time. If this feature vector was labeled as not lesion then 3.2% of the lesions would be labeled as false negatives. If only a few of the feature vectors exhibited these problems the approach could still have merit, but all the feature vectors exhibited this problem. Determining which feature vectors to label lesion becomes a tradeoff between the false positive and false negative rate.

The lack of feature vectors clearly indicating lesion conveys four possibilities: First, the ratings are inconsistent and inaccurate, a position for which some evidence was presented earlier. Second, the quantization resolution was too low to truly distinguish the feature vectors from each other. This is unlikely since the overlap between the lesion distributions and the other distributions is much wider than 0.5 standard devia-

tions. Third, most of the discriminatory information needed to characterize lesions is contained in the spatial relationship of lesion voxels to their neighbors, a position consistent with some of the literature [3], although other MS literature indicates a pure parametric approach as being reasonable [6]. The fourth possibility is that the bias correction and intensity standardization methods used were insufficient or inappropriate. There is a clear case to be made that performing intensity standardization independently on each image is an approach that could introduce more noise than it eliminates. The elements of the feature vectors are themselves not independent so it makes sense to work with vectors of intensities. Most recent approaches to intensity standardization have taken a vector approach [3, 1, 4, 11].

8 Future Work

There are a number of things that could be done to improve the results of the technique presented here. Performing joint histogram based intensity standardization as laid out in [3] would likely improve results since it operates on feature vectors rather than indepedent intensities. Further, experimenting to find the optimal quantization scheme would simultaneously reduce noise and improve recognition.

There are a number of follow on steps likely to prove valuable, such as using the joint histogram technique as a means of focusing on only those features that have some chance of being lesion and using those features as input to an additional classifer. The voxels predicted by the joint histogram as lesion could also be used as seed points for a region growing approach, thus incorporating the spatial information. Alternately, training a Markov Random Field which would then operate over the label image as predicted by the histogram, or the histogram combined with some classifier, would also incorporate spatial information and would likely outperform the techniques presented here. Additionally, further work into eliminating false positives, likely through a heirarchical classification approach, would also improve results.

A Figures

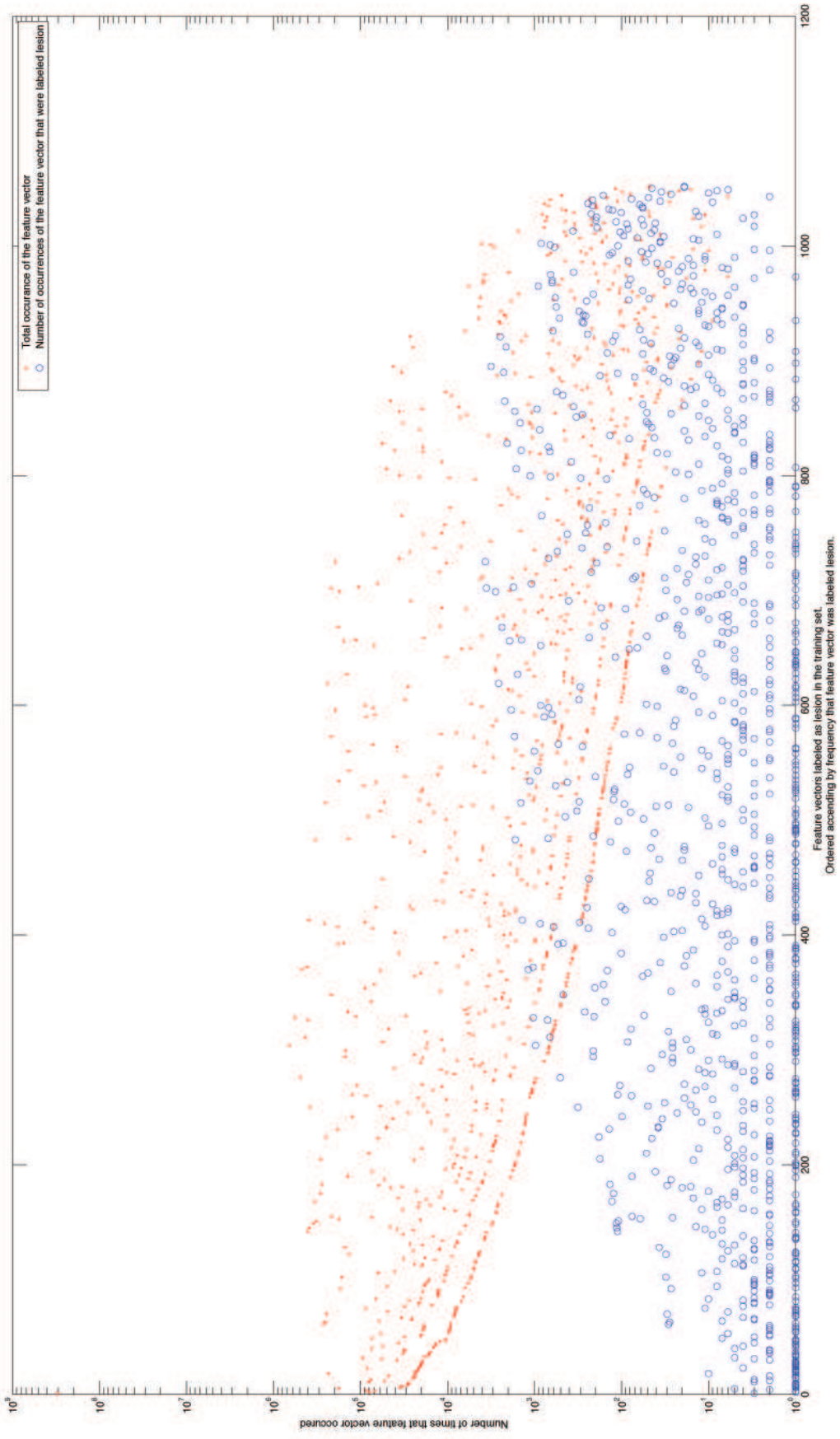

Figure 2: Lesion feature vectors versus occurence.

B Results Table

Ground Truth	UNC Rater								CHB Rater									STAPLE		
All Dataset	Volume Diff.		Avg. Dist.		True Pos.		False Pos.		Volume Diff.		Avg. Dist.		True Pos.		False Pos.		Total	Specificity	Sensitivity	PPV
	[%]	Score	[mm]	Score	[%]	Score	[%]	Score	[%]	Score	[mm]	Score	[%]	Score	[%]	Score				
UNC test1 Case01	33.3	95	33.1	32	2.3	53	93.9	52	94.9	86	32.0	34	3.1	53	93.9	52	57	0.9342	0.0147	0.0104
UNC test1 Case02	102.0	85	11.2	77	22.1	64	88.5	56	73.3	89	8.1	83	15.9	61	77.9	62	72	0.9911	0.4058	0.8711
UNC test1 Case03	26.5	96	2.6	95	47.2	78	38.5	86	5.1	99	2.1	96	50.0	80	34.1	89	90	0.9856	0.7641	0.7458
UNC test1 Case04	16.2	98	3.7	92	50.0	80	67.2	69	30.1	96	2.4	95	55.6	83	71.9	66	85	0.9842	0.8253	0.8110
UNC test1 Case05	16.0	98	48.9	0	0.0	51	100.0	49	89.7	87	43.4	11	0.0	51	100.0	49	49	0.9557	0.0000	0.0000
UNC test1 Case06	25.3	96	50.3	0	0.0	51	100.0	49	233.5	66	44.9	7	0.0	51	100.0	49	46	0.9361	0.0000	0.0000
UNC test1 Case07	55.6	92	39.2	19	0.0	51	100.0	49	3.1	100	35.7	27	0.0	51	100.0	49	55	0.9730	0.0000	0.0000
UNC test1 Case08	288.6	58	21.9	55	14.9	60	88.1	56	535.1	22	26.9	45	44.4	77	83.6	59	54	0.8567	0.2666	0.0505
UNC test1 Case09	420.2	38	58.4	0	0.0	51	100.0	49	633.9	7	74.7	0	0.0	51	100.0	49	31	0.9411	0.0000	0.0000
UNC test1 Case10	397.9	42	18.7	61	25.0	66	95.4	52	1699.9	0	24.1	50	66.7	89	96.3	51	51	0.9183	0.7070	0.2395
CHB test1 Case01	83.5	88	11.9	76	9.3	57	13.3	100	76.4	89	11.6	76	25.8	66	13.3	100	81	0.9989	0.0944	0.6930
CHB test1 Case02	14.9	98	6.6	86	31.8	70	63.2	71	63.7	91	4.5	91	36.8	72	21.1	97	84	0.9954	0.3711	0.8060
CHB test1 Case03	125.7	82	51.7	0	0.0	51	100.0	49	9.1	99	55.5	0	0.0	51	100.0	49	48	0.9655	0.0000	0.0000
CHB test1 Case04	283.9	58	47.1	3	0.0	51	100.0	49	84.7	88	49.7	0	0.0	51	100.0	49	44	0.9240	0.0000	0.0000
CHB test1 Case05	81.7	88	22.4	54	7.4	56	88.1	56	96.5	86	13.8	72	17.4	61	28.6	92	71	0.9996	0.0278	0.7898
CHB test1 Case06	32.4	95	3.6	93	44.4	77	89.0	55	29.4	96	3.5	93	36.4	72	90.5	55	79	0.9861	0.4395	0.6908
CHB test1 Case07	19.1	97	6.0	88	50.0	80	86.7	57	50.8	93	2.5	95	57.9	84	64.2	71	83	0.9937	0.4020	0.8293
CHB test1 Case08	11.5	98	14.1	71	33.3	70	87.7	56	40.8	94	14.5	70	20.6	63	86.4	57	73	0.9809	0.3119	0.4662
CHB test1 Case09	3.8	99	3.1	94	41.6	75	42.7	84	12.5	98	2.3	95	32.7	70	28.1	93	88	0.9922	0.7212	0.8686
CHB test1 Case10	154.7	77	8.1	83	78.9	96	96.2	51	24.6	96	5.0	90	79.3	97	91.9	54	81	0.9795	0.5347	0.5604
CHB test1 Case11	35.5	95	37.2	23	6.8	55	97.4	50	79.2	88	38.3	21	13.8	59	91.2	54	56	0.9879	0.0195	0.0814
CHB test1 Case12	80.9	88	21.6	55	19.3	62	83.2	59	81.0	88	23.5	52	7.7	56	87.5	56	65	0.9840	0.0451	0.1892
CHB test1 Case13	27.2	96	7.0	86	40.0	74	84.6	58	55.4	92	3.9	92	38.1	73	42.3	84	82	0.9969	0.5486	0.9139
CHB test1 Case15	90.6	87	11.5	76	12.3	59	46.9	81	87.6	87	8.7	82	10.6	58	37.5	87	77	0.9948	0.0523	0.5321
All Average	101.1	85	22.5	55	22.4	64	81.3	60	174.6	81	22.2	57	25.5	66	72.5	65	67	0.9690	0.2730	0.4229
All UNC	138.2	80	28.8	43	16.1	61	87.2	57	339.9	65	29.4	45	23.6	65	85.8	57	59	0.9476	0.2984	0.2728
All CHB	74.7	89	18.0	63	26.8	67	77.1	63	56.5	92	16.9	66	26.9	67	63.0	71	72	0.9842	0.2549	0.5301

Table 2: Evaluation table for MICCAI MS lesion segmentation

C Acknowledgements

This research supported by NIH grant U54 EB005149 and DOE grant DE-FG02-99ER6274.

References

[1] Ingemar J. Cox, S. Roy, and Sunita L. Hingorani. Dynamic histogram warping of image pairs for constant image brightness. In *ICIP*, pages 2366–2369, 1995. 7

[2] M. Styner et al. Parametric estimate of intensity inhomogeneities applied to MRI. *TMI*, 19(3):153–165, March 2000. 3

[3] Florian Jäger. A new Method for MRI Intensity Standardization with Application to Lesion Detection in the Brain. 2006. 3, 4, 7, 8

[4] Florian Jäger, Lszl Nyl, Bernd Frericks, Frank Wacker, and Joachim Hornegger. Whole Body MRI Intersity Standardization. In Alexander Horsch, Thomas M. Deserno, Heinz Handels, Hans-Peter Meinzer, and Thomas Tolxdorff, editors, *Bildverarbeitung für die Medizin 2007*, pages 459–463, Berlin, 2007. 3, 7

[5] Rasoul Khayati, Mansur Vafadust, Farzad Towhidkhah, and Massood Nabavi. Fully automatic segmentation of multiple sclerosis lesions in brain mr flair images using adaptive mixtures method and markov random field model. *Comput. Biol. Med.*, 38(3):379–390, 2008. 4

[6] Zhiqiang Lao, Dinggang Shen, A. Jawad, B. Karacali, Dengfeng Liu, E.R. Melhem, R.N. Bryan, and C. Davatzikos. Automated segmentation of white matter lesions in 3d brain mr images, using multivariate pattern classification. *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 307–310, April 2006. 4, 7

[7] A. Madabhushi and J.K. Udupa. Interplay between intensity standardization and inhomogeneity correction in mr image processing. *Medical Imaging, IEEE Transactions on*, 24(5):561–576, May 2005. 3

[8] L.G. Nyul, J.K. Udupa, and Xuan Zhang. New variants of a method of mri scale standardization. *Medical Imaging, IEEE Transactions on*, 19(2):143–150, Feb 2000. 3

[9] S. M. Smith. Fast robust automated brain extraction. *Hum Brain Mapp*, 17(3):143–155, November 2002. 3

[10] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *Medical Imaging, IEEE Transactions on*, 23(7):903–921, July 2004. 6

[11] N.L. Weisenfeld and S.K. Warfteld. Normalization of joint image-intensity statistics in mri using the kullback-leibler divergence. *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*, pages 101–104 Vol. 1, April 2004. 7