MR Brain Segmentation using Decision Trees

Amod Jog, Snehashis Roy, Jerry L. Prince, and Aaron Carass

Image Analysis and Communication Laboratory,
Dept. of Electrical and Computer Engineering,
The Johns Hopkins University
`http://iacl.ece.jhu.edu/`

Abstract. Segmentation of the human cerebrum from magnetic resonance images (MRI) into its component tissues has been a defining problem in medical imaging. Until recently, this has been solved as the tissue classification of the T_1-weighted (T_1-w) MRI, with numerous solutions for this problem having appeared in the literature. The clinical demands of understanding lesions, which are indistinguishable from gray matter in T_1-w images, has necessitated the incorporation of T_2-weighted Fluid Attenuated Inversion Recovery (FLAIR) images to improve segmentation of the cerebrum. Many of the existing methods fail to handle the second data channel gracefully, because of assumptions about their model. In our new approach, we explore a model free algorithm which uses a classification technique based on ensembles of decision trees to learn the mapping from an image feature to the corresponding tissue label. We use corresponding image patches from a registered set of T_1-w and FLAIR images with a manual segmentation to construct our decision tree ensembles. Our method is efficient, taking less than two minutes on a $240 \times 240 \times 48$ volume. We conduct experiments on five training sets in a leave-one-out fashion, as well as validation on an additional twelve subjects, and a landmark validation experiment on another cohort of five subjects.

Keywords: Magnetic resonance, brain segmentation, patches, classification, decision trees, random forest

1 Introduction

The segmentation of magnetic resonance images (MRI) of the whole head into just the primary cerebrum tissues of cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM) has been one of the core challenges of the neuroimaging community for the past twenty years. The majority of existing solutions are conceived as a pipeline, with several preprocessing steps used to isolate the cerebrum before it is segmented. These include inhomogeneity correction—the most well known being N3 [26]—followed or preceded by skull stripping—see Table 1 in [24] for a recent overview—and then either an image intensity standardization technique or directly into the segmentation task. The segmentations approaches that have been employed for this three class problem include: Gaussian distribution based such as Expectation Maximization Segmentation (EMS) [27], unified segmentation [1], and FMRIB's Automated Segmentation Tool (FAST) [28]; Fuzzy c-means (FCM) approaches such as FANTASM [13] and several others [8, 9, 17]; and more recently the Rician based distributions [19]. Newer methods have tended to include one of these distributions at their core while incorporating statistical [1] and topology [3] atlases to help improve their accuracy.

These approaches assume that there are nice reasonable distributions that can approximate all given data, regardless of the patients pathology. In this work we want to explore the possibility of a distribution free model, that can provide rapid tissue segmentations. We have chosen to use random decision forests [11, 4] which provide a model free framework that can learn a complicated distribution that would otherwise be poorly approximated by a fixed distribution choice. Our method uses some existing software tools to isolate the cerebrum in the whole head MRI, by removing the skull [7, 6] and the cerebellum [3]. We then use a decision tree ensemble to generate a hard classification of the tissues in the cerebrum. The approach is inspired in part by the patch matching literature, which has been used for image synthesis [23, 21, 20], super-resolution [20, 14, 16], inhomogeneity correction [18], segmentation [15, 22]. We extend these ideas with the framework of regression based image reconstruction [12] to reconstruct a segmentation of an unseen input image.

2 Method

We use T_1-w and FLAIR images which have been co-registered and bias corrected in our algorithm. We use $\{\mathcal{I}_t^{(T)}, \mathcal{I}_t^{(F)}, \mathcal{I}_t^{(C)}\}, t = 1, \ldots, 5$, to denote the t^{th} training subject, which correspond to the T_1-w, FLAIR, and manual segmentation image respectively. The class image has labels, $1, 2, 3$, which are CSF, GM, and WM respectively. The training data images also have white matter lesions (WML), which have the appearance of GM in $\mathcal{I}_t^{(T)}$, though we wish to segment them as WM.

2.1 Preprocessing Training Data

Fig. 1 provides a flowchart of our algorithm. The training data images are skull stripped and manually labeled using the contour segmentation objects (CSO) tool in MeVisLab. The T_1-w ($\mathcal{I}_t^{(T)}$) images are linearly scaled so that their mean WM intensities are at 1000, the mean WM intensity is found by fitting a three-class Gaussian Mixture Model (GMM) to the intensity histograms. The FLAIR images ($\mathcal{I}_t^{(F)}$) are linearly scaled so that the mode of WM intensities is 1000, the WM mode is obtained from the intensity histogram, smoothed by a kernel density estimator.

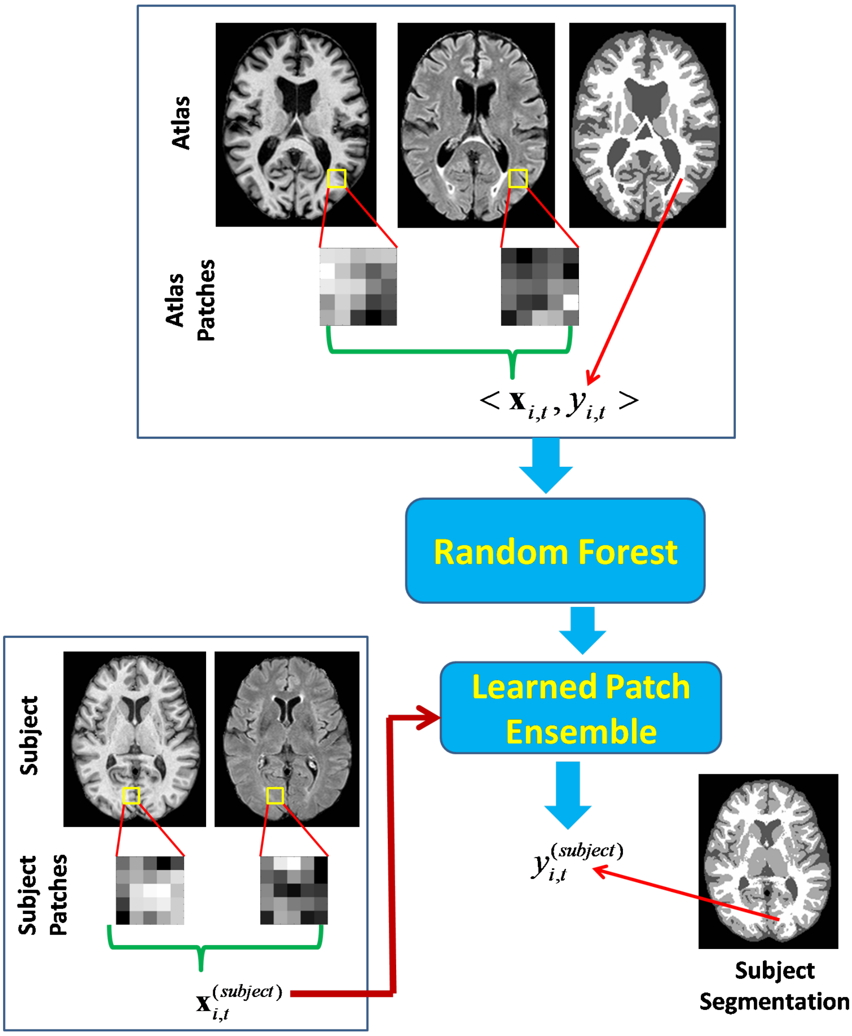

Fig. 1. A overview of our algorithm. The input training data is converted into patches which are fed into our *random forest*. This outputs a *learned patch ensemble* of decision trees, which are used on the test data to predict a subjects segmentation.

2.2 Training and Prediction

At each voxel i of the t^{th} training data $\left(\mathcal{I}_t^{(T)}, \mathcal{I}_t^{(F)}, \mathcal{I}_t^{(C)}\right)$, $p \times q \times r$-sized 3D image patches are defined on $\mathcal{I}_t^{(T)}$ and $\mathcal{I}_t^{(F)}$ and denoted as $\mathbf{x}_{i,t}^{(T)}$ and $\mathbf{x}_{i,t}^{(F)}$, respectively, which reside in a d dimensional space where $d = pqr$. $\mathbf{x}_{i,t}^{(T)}$ and $\mathbf{x}_{i,t}^{(F)}$ are concatenated to form a $2d \times 1$ vector $\mathbf{x}_{i,t}$, which acts as the feature vector for the i^{th} voxel with a corresponding label taken from the i^{th} voxel of

$\mathcal{I}_t^{(C)}$, denoted by $y_{i,t}$. We thus consider $\mathbf{x}_{i,t}$'s as attributes with the dependent variables being $y_{i,t}$'s. We can then construct training pairs of $\langle \mathbf{x}_{i,t}, y_{i,t} \rangle$ for each voxel i in each training subject t. Using all the available data, i.e. all the voxels in all five subjects, leads to an unbalanced training set as each tissue class is not represented equally, thus care is taken to ensure equal proportions of each class in the training data.

We pursue a classification tree solution which enables us to directly use the training algorithm described in [5] to train a bagged ensemble of decision trees. A single decision tree partitions our $2d$-dimensional space by splitting different dimensions using a learned threshold. During training, one third of the attributes are randomly considered for the choice of splitting and the one that best minimizes the Gini impurity criterion, after deciding a threshold, is chosen as the dimension to split upon. A single decision tree is considered as a weak learner and has higher error in general, thus we use a bagged ensemble of decision trees which reduces errors through bootstrap aggregation. In this process, the ensemble consists of n trees, each of which is learned from a bootstrapped data set—which are created by randomly sampling with replacement from the whole training data set, N times where N is the number of training samples in the entire training data. We limit the depth of each tree by fixing the number of samples accumulated at a leaf to be five, thus preventing over-fitting. Prediction is done by passing a test feature vector through each tree and allowing it to traverse the nodes of the tree by observing the splitting criterion and threshold at each node until it reaches a leaf node. The predicted label is calculated by voting between the training data vectors present in the leaf. The training data consists of $\sim 10^6$ samples from the five training subjects, with training done in parallel, we can create a trained ensemble of decision trees takes on average 256 seconds on an 8-core, 2.73GHz machine, image preprocessing which includes normalization of T1 and FLAIR data takes on average 67 seconds, while prediction on a new unseen data set takes on average 31 seconds on the same machine.

3 Results

We perform three experiments to demonstrate the practicality of this new segmentation method. The first is a leave-one-out cross-validation on the training data, the second is an analysis on 12 additional subjects from the same cohort as the training data, and finally a study of the accuracy of the defined CSF/GM & GM/WM boundaries using manually picked landmarks. The training and test data consists of T_1-w and FLAIR images both with resolution of $0.958 \times 0.958 \times 3.0$mm with the manual segmentation being conducted in the same space. Our landmark cohort is made up of five healthy subjects (3 females) with a mean age of 39.4 years (range: 30-49) with the T_1-w and FLAIR images having an isotropic resolution of 1.1mm^3. Two raters (Raters A and B) then placed 10 landmark points upon the inner and outer boundaries of the cortex in each of 21 coarsely selected regions, resulting in each rater picking 210 landmarks per surface for each of the five subjects, more details are available in Shiee et al. [25].

3.1 Cross-Validation

In each round of our cross-validation experiment we removed a single data set from the training sample of five subjects and trained our decisions trees as described in Sec. 2 with the four remaining data sets. The trained decision tree ensemble is then tested on the held out data with evaluation on the three classes including Dice score, 95% Hausdorff distance, and absolute volume difference. The results are reported in Table 1, see Babalola et al. [2] and Dubuisson et al. [10] for an explanation of the metrics used. To provide a baseline for comparison purposes, we computed the same metrics after using FreeSurfer to segment the data, also in Table 1. Fig. 2 has three orientations of a training data set showing the T_1-w, FLAIR, and both the manual segmentation and the result of our algorithm. The red arrow in Fig. 2 denotes a region in the midsagittal plane where our algorithm seems to make a more sensible decision than the human rater by leaving a clear separation between the hemispheres.

Table 1. Cross validation on the five test subjects, performed by training on four data sets and evaluating on the fifth. We report the Dice score, 95% Hausdorff distance (HD), and the absolute volume difference (Abs. Vol. Diff.) as a percentage of the total brain volumes. More details about the computation of these metrics is available from Babalola et al. [2] and Dubuisson et al. [10]. For comparison purposes, we include the results of FreeSurfer on the same data.

	Dice (%)		**95% HD (mm)**		**Abs. Vol. Diff.**	
Structure	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
GM	83.42	1.67	4.11	1.16	2.978	2.013
WM	87.13	2.62	2.18	0.19	1.198	1.010
CSF	81.38	2.55	1.88	0.58	4.466	2.178
Brain	94.71	0.50	7.33	1.45	2.899	1.628
FreeSurfer GM	69.66	1.72	12.02	0.66	2.027	1.128
FreeSurfer WM	79.88	2.52	10.26	0.64	7.377	1.641

3.2 Test Data

For evaluation on the test data, we trained our decision trees on all five subjects in the training data and then used this trained ensemble to predict the segmentation in the test data. The results are shown in Table 2 and some example segmentations for the test data are given in Fig. 3.

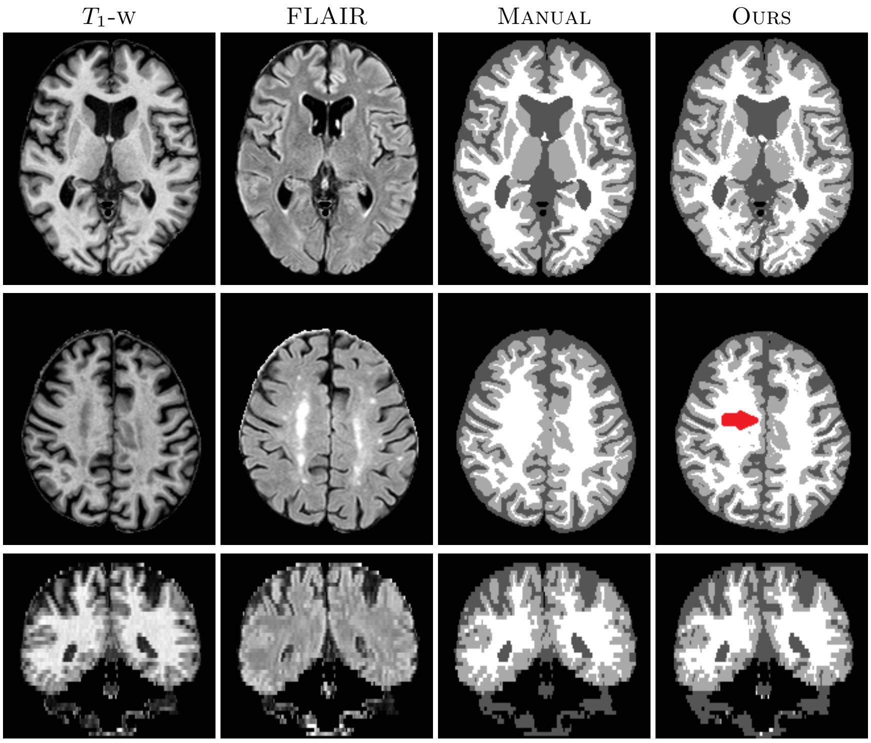

Fig. 2. Each row shows a specific orientation from a training data set. From left to right the columns are: T_1-w, FLAIR, manual segmentation, and the result of our algorithm.

Table 2. Results on 12 test subjects. We report the Dice score, 95% Hausdorff distance (HD), and the absolute volume difference (Abs. Vol. Diff.) as a percentage of the total brain volumes. More details about the computation of these metrics is available from Babalola et al. [2] and Dubuisson et al. [10]. AIS denotes all internal structures.

	Dice (%)		**95% HD (mm)**		**Abs. Vol. Diff.**	
Structure	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
GM	83.46	1.95	2.29	0.44	6.37	4.45
WM	87.01	1.10	3.42	1.03	6.80	5.20
CSF	66.46	2.40	15.67	2.51	13.39	9.21
Brain	94.75	0.60	2.96	0.38	3.32	2.01
AIS	92.53	0.55	27.89	1.49	3.52	1.74

3.3 Landmark Validation

The same trained ensemble that we used on the twelve test data subjects was used on our landmark cohort. With each landmark representing either the CSF/GM or GM/WM interface, we computed the shortest distance from each landmark to the corresponding boundary as defined by our voxel based segmentation. For

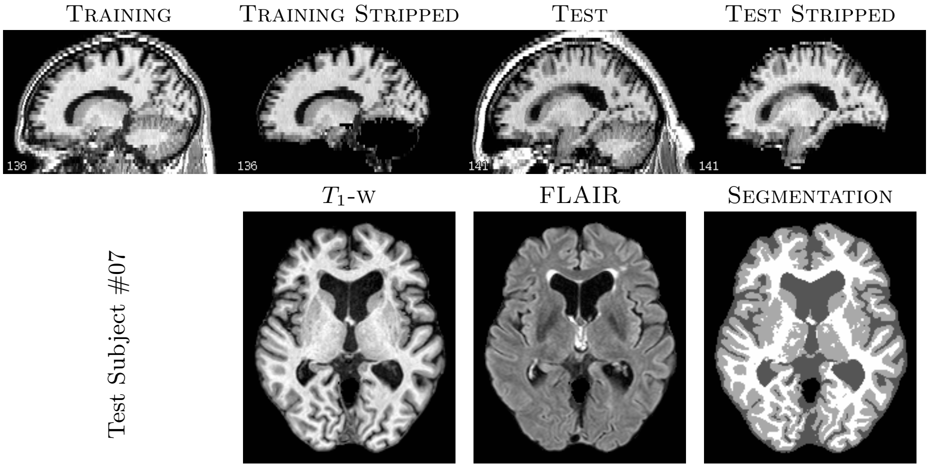

Fig. 3. The top row shows a sagittal view comparison of the skull-stripping on training and test data. The bottom row shows an axial view of the T_1-w, FLAIR, and our segmentation of the same test subject.

Table 3. Landmark results based on five subjects with 420 manually picked landmarks, with 210 landmarks on each of the inner and outer surfaces, by two raters.

	Inner Surface		**Outer Surface**	
	Rater A	**Rater B**	**Rater A**	**Rater B**
Sub. 1	0.52 (0.51)	0.58 (0.61)	1.08 (1.04)	0.99 (1.01)
Sub. 2	0.54 (0.44)	0.62 (0.73)	0.68 (0.72)	0.64 (0.79)
Sub. 3	0.73 (0.97)	0.70 (0.95)	0.65 (0.60)	0.64 (0.59)
Sub. 4	0.41 (0.34)	0.46 (0.38)	0.67 (0.50)	0.64 (0.63)
Sub. 5	0.65 (0.80)	0.70 (0.78)	0.98 (0.69)	1.13 (0.82)
Mean	0.57 (0.66)	0.61 (0.71)	0.81 (0.73)	0.80 (0.78)
FreeSurfer	0.47 (0.38)	0.44 (0.38)	0.51 (0.36)	0.44 (0.38)

a comparison to the state-of-the-art, we also ran FreeSurfer on each of the landmark data sets and computed the shortest distance between each landmark and the appropriate surface generated by FreeSurfer. The results are shown in Table 3.

4 Conclusion

We present a new approach to MR brain segmentation with a focus on speed while achieving very high accuracy. The Cross-Validation and Test Data experiments demonstrate that we can consistently achieve very good results for all three metrics with respect to GM and WM segmentation. In comparison to the hard segmentation generated by FreeSurfer on the Training Data, we are clearly much better for all three metrics. Our inferior results for CSF segmentation on the test data set, are in large part due to the skull stripping differences between the training and text subjects, this is best evidenced by considering both the 95% Hausdorff distance and the absolute volume difference. These metrics show a very large difference in the volumes and the distance between mislabeled voxels for CSF, as our CSF volume extends outside the CSF volume labeled by the manual experts. Our landmark data provide more confirmation that our estimation of the boundaries of WM & GM and GM & CSF are close to the state-of-the-art even though they are just at the voxel level, and not sub-voxel like all surface generation software tools.

5 Acknowledgments

This research was supported by the NIH/NIBIB under grant 1R21EB012765-01A1.

References

1. Ashburner, J., Friston, K.J.: Unified segmentation. NeuroImage 26(3), 839–851 (2005)
2. Babalola, K.O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M., Rueckert, D.: An evaluation of four automatic methods of segmenting the subcortical structures in the brain. NeuroImage 47(4), 1435–1447 (2009)
3. Bazin, P.L., Pham, D.L.: Topology-Preserving Tnumber Classification of Magnetic Resonance Brain Images. IEEE Trans. Med. Imag. 26(4), 487–496 (2007)
4. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 123–140 (1996)
5. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth Publishing Company, U.S.A. (1984)
6. Carass, A., Cuzzocreo, J., Wheeler, M.B., Bazin, P.L., Resnick, S.M., Prince, J.L.: Simple Paradigm for Extra-cerebral Tissue Temoval: Algorithm and Analysis. NeuroImage 56(4), 1982–1992 (2011)
7. Carass, A., Wheeler, M.B., Cuzzocreo, J., Bazin, P.L., Bassett, S.S., Prince, J.L.: A Joint Registration and Segmentation Approach to Skull Stripping. In: 4th IEEE International Symposium on Biomedical Imaging (ISBI 2007). pp. 656–659 (2007)
8. Cavalcanti, N., de Carvalho, F.: An Adaptive Fuzzy C-Means Algorithm with the L_2 Norm. Australian Conf. on Artificial Intel. pp. 1138–1141 (2005)
9. Clark, K.A., Woods, R.P., Rottenber, D.A., Toga, A.W., Mazziotta, J.C.: Impact of acquisition protocols and processing streams on tissue segmentation of T1 weighted MR images. NeuroImage 29(1), 185–202 (2006)
10. Dubuisson, M.P., Jain, A.K.: A modified Hausdorff distance for object matching. In: Intl. Conf. on Pattern Recognition. pp. 566–568 (1994)
11. Ho, T.K.: Random Decision Forest. In: Proc. of the 3rd International Conference on Document Analysis and Recognition. pp. 278–282 (14-16 August 1995)
12. Jog, A., Roy, S., Carass, A., Prince, J.L.: Magnetic Resonance Image Synthesis through Patch Regression. In: 10th IEEE International Symposium on Biomedical Imaging (ISBI 2013). pp. 350–353 (2013)

13. Pham, D.L.: Robust fuzzy segmentation of magnetic resonance images. In: 14th IEEE Symposium on Computer-Based Medical Systems. pp. 127–131 (2001)
14. Rousseau, F.: Brain hallucination. In: Proceedings of the European Conference on Computer Vision (ECCV 2008). pp. 497–508. LNCS (2008)
15. Rousseau, F., Habas, P.A., Studholme, C.: A Supervised Patch-Based Approach for Human Brain Labeling. IEEE Trans. Med. Imag. 30(11), 1852–1862 (2011)
16. Rousseau, F., Studholme, C.: A supervised patch-based image reconstruction technique: Apllication to brain MRI super-resolution. In: 10^{th} IEEE International Symposium on Biomedical Imaging (ISBI 2013). pp. 346–349 (2013)
17. Roy, S., Agarwal, H., Carass, A., Bai, Y., Pham, D.L., Prince, J.L.: Fuzzy C-Means with variable compactness. In: 5^{th} IEEE International Symposium on Biomedical Imaging (ISBI 2008). pp. 452–455 (2008)
18. Roy, S., Carass, A., Bazin, P.L., Prince, J.L.: Intensity inhomogeneity correction of magnetic resonance images using patches. In: Proceedings of SPIE-MI 2011. p. 79621F (2011)
19. Roy, S., Carass, A., Bazin, P.L., Resnick, S.M., Prince, J.L.: Consistent segmentation using a Rician classifier. Medical Image Analysis 16(2), 524–535 (2012)
20. Roy, S., Carass, A., Prince, J.L.: Synthesizing MR Contrast and Resolution through a Patch Matching Technique. In: Proceedings of SPIE-MI 2010. p. 76230j (2010)
21. Roy, S., Carass, A., Prince, J.L.: A Compressed Sensing Approach For MR Tissue Contrast Synthesis. In: $22nd$ Conf. on Inf. Proc. in Medical Imaging (IPMI). pp. 371–383 (2011)
22. Roy, S., He, Q., Jog, A., Carass, A., Calabresi, P.A., Prince, J.L., Pham, D.L.: Example Based Lesion Segmentation. In: Proceedings of SPIE-MI 2014 (2014)
23. Roy, S., Jog, A., Carass, A., Prince, J.L.: Atlas based intensity transformation of brain mr images. In: Multimodal Brain Image Analysis, Lecture Notes in Computer Science, vol. 8159, pp. 51–62 (2013)
24. Shi, F., Wang, L., Dai, Y., Gilmore, J.H., Lin, W., Shen, D.: LABEL: Pediatric brain extraction using learning-based meta-algorithm. NeuroImage 62(3), 1975–1986 (2012)
25. Shiee, N., Bazin, P.L., Cuzzocreo, J.L., Ye, C., Kishore, B., Carass, A., Calabresi, P.A., Reich, D.S., Prince, J.L., Pham, D.L.: Robust Reconstruction of the Human Brain Cortex in the Presence of the WM Lesions: Method and Validation. Human Brain Mapping (*In Press*) (2014), DOI: 10.1002/hbm.22409
26. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imag. 17(1), 87–97 (Feb 1998)
27. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated Model-Based Tissue Classification of MR Images of the Brain. IEEE Trans. Med. Imag. 18, 897–908 (1999)
28. 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. Med. Imag. 20, 45–57 (2001)