

# Highly Accurate Segmentation of Brain Tissue and Subcortical Gray Matter from Newborn MRI

Neil I. Weisenfeld<sup>1,2</sup>, Andrea U.J. Mewes<sup>1</sup>, and Simon K. Warfield<sup>1</sup>

<sup>1</sup> Computational Radiology Laboratory  
Brigham and Women's and Children's Hospitals  
Harvard Medical School, Boston, MA

<sup>2</sup> Department of Cognitive and Neural Systems  
Boston University, Boston, MA

**Abstract.** The segmentation of newborn brain MRI is important for assessing and directing treatment options for premature infants at risk for developmental disorders, abnormalities, or even death. Segmentation of infant brain MRI is particularly challenging when compared with the segmentation of images acquired from older children and adults. We sought to develop a fully automated segmentation strategy and present here a Bayesian approach utilizing an atlas of priors derived from previous segmentations and a new scheme for automatically selecting and iteratively refining classifier training data using the STAPLE algorithm. Results have been validated by comparison to hand-drawn segmentations.

## 1 Introduction

The number of preterm births in the U.S. has risen more than 30% since 1981 and the overall incidence of children born with low birthweight was 7.9% in 2003. Furthermore, “preterm birth is a leading cause of infant death and is associated with nearly one-half of all congenital neurological defects (e.g. cerebral palsy)” [1, page 17]. Regional brain tissue volume, as assessed by magnetic resonance (MR) imaging, is a biomarker for structural maturation that can be used to assess the effectiveness of interventions and for screening candidates likely to have adverse outcomes and directing them toward appropriate treatments.

MR images from the brains of newborn children are particularly difficult to segment. Aside from the occasionally unavoidable artifacts due to the motion of a non-sedated subject, neonatal brains exhibit different signal intensity characteristics than fully developed brains. The minimal brain segmentation process seeks to identify cortical gray matter with most investigators also seeking to at least identify white matter and cerebrospinal fluid. Doing so requires at least two channels of MR data acquired from different pulse sequences and it is customary to use a T1-weighted sequence and a T2-weighted sequence to differentiate CSF, gray matter, and fully-myelinated white matter.

In the developing brain of the normal newborn, myelination progresses in a stereotyped fashion starting with the cortico-spinal tracts. Shortly after delivery, white matter elsewhere in the brain is largely unmyelinated. Researchers studying these processes are interested in differentiating and quantifying both myelinated and unmyelinated white matter as well as cortical and sub-cortical gray matter. Differentiation of these tissue classes is not possible based solely on image intensities, so we have chosen to use an atlas of prior probabilities derived from previous segmentations. We present here a fully-automatic method for image segmentation including a new, atlas-based scheme for selection of training points for supervised classification. A number of candidate segmentations is generated for each subject, STAPLE is used to generate an optimal linear combination of candidate segmentations, and the training points for each subject are edited based on the STAPLE estimate of the true segmentation. The process is then repeated with the edited training points and iterated until a stable, final segmentation is achieved.

## 2 Materials and Methods

### 2.1 Data Acquisition

Data was acquired on a 1.5T clinical MR scanner from GE Medical Systems (Waukesha, WI, USA) using a head-surface RF receiver coil. The present study utilizes MR scans from two pulse sequences: a T1-weighted SPGR (spoiled gradient recall) sequence and a T2-weighted scan from an FSE (fast spin echo) sequence. Additional data were acquired during these scans and not used for the present study. The scans were coronally acquired and are 256 x 256 voxels in-plane, each covering approximately 0.7mm x 0.7mm of tissue. Slice thickness was 1.5mm for the SPGR sequence and 3.0mm for the FSE sequence.

### 2.2 Segmentation Process

**Preprocessing** Preprocessing consisted of intensity non-uniformity correction, anisotropic diffusion filtering, and intra-subject rigid registration to account for subject motion between scans. For intensity non-uniformity correction we have implemented and extended the method proposed by [3] and fully developed in [4]. This method is virtually parameter-free and we have found it to be very robust. To reduce thermal noise in scans, we use a standard anisotropic diffusion [5,6] implementation provided in the open-source Insight Toolkit (ITK) [7]. Rigid registration is performed using a mutual-information implementation, also in ITK.

**Atlas Construction/Registration** As was mentioned previously, differentiation of sub-cortical gray matter, such as basal ganglia, and myelinated white matter from csf, unmyelinated white matter, and cortical gray matter requires information beyond image intensities alone. We have chosen to use an atlas of prior probabilities for the existence of a given tissue class at a given location (e.g. [8]). For the construction of this atlas, scans were taken from thirteen

healthy preterm infants born after 28 weeks GA and scanned at approximately 42 weeks. These scans were segmented using a previously-published semi-automated segmentation technique [8]. A model subject was chosen and scans from the remaining twelve subjects were registered to the model using an affine, mutual-information-based registration algorithm from ITK [7]. For each atlas subject, a transformation to the space of the model individual was estimated using either the T2-weighted or SPGR scan from both the model and the subject. This transformation was then applied to the individual segmentations to yield a set of aligned segmentations. Once aligned, an atlas of voxelwise probabilities was created by calculating how frequently each tissue class label appears at each voxel coordinate. In order to apply the atlas to the segmentation of a new individual, a second affine transformation is calculated to register the atlas to the new subject. This registration is performed in a similar manner as during atlas construction, only the pair-wise registration is performed using an average SPGR or T2 generated from scans of the subjects that went into the atlas itself.

**Classification** The goal of our segmentation process in newborns is to assign tissue class labels (background, gray matter, unmyelinated white matter, myelinated white matter, cerebrospinal fluid, and basal ganglia) to each voxel in an image volume. Even for gray matter, csf, and white matter, this is not possible based on intensities in a single image, so we consider joint intensities in both T1 and T2 weighted images and attempt to differentiate tissue classes. In order to achieve this, we generate a non-parametric estimate of the class-conditional probability density functions for each of the relevant tissue classes.

We have previously demonstrated that supervised statistical classification (classification based on the use of positive and negative examples) has advantages in terms of robustness and accuracy when compared with unsupervised classification (techniques that attempt to identify tissue classes solely from algorithmic principles) [9]. The generation of training data for supervised classification generally involves a human expert selecting a number of prototype voxels representing each tissue class and recording samples of image intensities at these locations. While this process is often not terribly time consuming, the availability of large quantities of data and larger studies makes it desirable to have a more automated method. Furthermore, manual selection of tissue class prototypes requires a trained expert, is subjective, and suffers from inter- and intra-rater variability.

We have previously [10] presented segmentation in terms of a Bayesian framework and we review the key points here and present our method for automatic selection of training points. Given a total of  $W$  tissue classes, a tissue classification can be considered the assignment of tissue class labels  $w_i; i = 1..W$  to each voxel coordinate. Ultimately, we seek the posterior probability of a tissue class given the measured intensities:  $p(w_i|\mathbf{x})$  where  $\mathbf{x}$  is a vector of image intensity samples from each channel. Bayes' rule can be stated as  $p(w_i|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|w_i)p(w_i)$ . A priori, we consider  $p(\mathbf{x})$  to be uniform across intensities and it therefore represents a uniform scale factor. We estimate the posterior probability  $p(w_i|\mathbf{x})$  by

first estimating the likelihood  $p(\mathbf{x}|w_i)$  and then multiplying by the statical prior  $p(w_i)$  provided by the atlas for each individual voxel location.

The likelihood  $p(\mathbf{x}|w_i)$  is the conditional probability density function (pdf) for class  $w_i$ . The training data represents a set of samples drawn from this pdf and the process of estimating an unknown pdf from a collection of such samples is covered in detail in [11] and we summarize those results here. The probability that a sample  $\mathbf{x}$  will fall in a region of the feature space  $R$  is given by  $P = \int_R p(\mathbf{x}')d\mathbf{x}'$ .  $P$  therefore gives a locally averaged (smoothed) estimate of  $p(\mathbf{x})$  and converges to  $p(\mathbf{x})$  as the region,  $R$ , gets smaller. We can estimate  $p(\mathbf{x})$  therefore by estimating  $P$ . Consider a set of  $n$  samples drawn independently from  $p(\mathbf{x})$  and  $k$  samples fall inside region  $R$ , the expected value for  $k$  is  $E[k] = nP$  and a good estimator for  $P$  is  $\frac{E[k]}{n}$ . Assuming that region  $R$  is small so that  $p(\mathbf{x})$  doesn't vary appreciably within the region and that  $p(\mathbf{x})$  is continuous, then  $P = \int_R p(\mathbf{x}')d\mathbf{x}' \approx p(\mathbf{x})V$  where  $V$  is the volume of feature space enclosed by region  $R$ . For a given trial, we assume that our  $k$  is an estimate for  $E[k]$  and that an estimate for the locally averaged probability density function is  $p(\mathbf{x}) = \frac{k/n}{V}$ .

Furthermore, if  $k_i$  of our  $k$  samples that fall inside region  $R$  belong to class  $w_i$ , then an estimator for the joint probability is  $p(\mathbf{x}, w_i) = \frac{k_i/n}{V}$ . The posterior probability of any tissue class can then be estimated by [11]:

$$p(w_i|\mathbf{x}) = \frac{p(\mathbf{x}, w_i)}{\sum_{j=1}^W p(\mathbf{x}, w_j)} = \frac{k_i}{k}$$

This also give a robust and effective estimator for the likelihood  $p(\mathbf{x}|w_i) = p(\mathbf{x}, w_i)/p(w_i)$ . At this stage of analysis, the prior  $p(w_i)$  is considered stationary across the image and is uniform due to a similar number of training samples in each class. The probabilities resulting from this classification are therefore a set of likelihood estimates for each class at each voxel. Simple multiplication of these likelihood estimates  $p(\mathbf{x}|w_i)$  by the spatial priors encapsulated in the atlas  $p(w_i)$  yields the posterior probability  $p(w_i|\mathbf{x})$ , since we still consider  $p(\mathbf{x})$  to be uniform. A maximum aposteriori labeling of the volume is achieved by picking the class with the largest posterior value at each voxel. Prior to choosing the maximum posterior at each voxel, a Markov random field model can be utilized to provide a spatial homogeneity constraint to further refine the statistics (see, for instance, [12]).

We now turn to the problem of automatically selecting training points for a given subject. The atlas is derived from a collection of individual scans and segmentations and we'll call a subject used in the construction of the atlas an "atlas subject" and the subject being segmented the "test subject." Each atlas subject has an associated affine transformation  $T_a$  that was derived to bring the subject into alignment with a single, model subject. A second transformation  $T_s$  aligns the atlas itself to a test subject whose data is to be segmented. The composition of these two brings a coordinate from "atlas subject space" to "test subject space" and we'll call this transformation  $T_{as}$ . For each of our thirteen atlas individuals, we have a set of hand-chosen prototypes (classifier training data) that are sample locations of a given tissue class in that individual. The

prototypes were chosen to provide an effective supervised classification for that subject. A prototype is a pairing of a voxel coordinate  $\mathbf{c}$  and a tissue class label  $w_i$  such as  $(\mathbf{c}, w_i)$ . We can transform our voxel coordinates from atlas subject space to test subject space to yield test-subject-specific prototypes  $(T_{as}\{\mathbf{c}\}, w_i)$ .

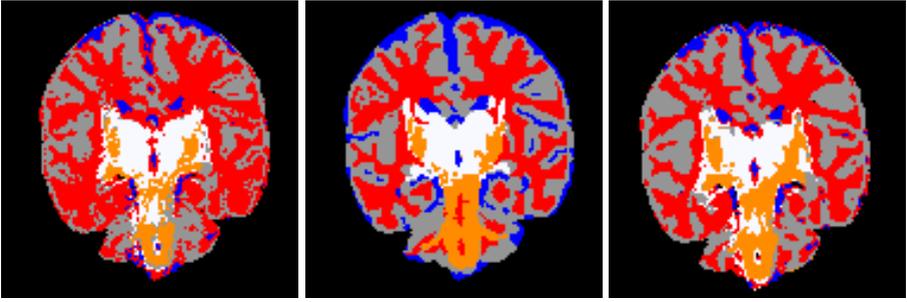
For each of our thirteen atlas individuals, we transform each of our prototype coordinates into the space of our new subject and use these as training points to generate a supervised classification as described above. In this manner, we generate thirteen candidate segmentations for each test subject to be segmented. The STAPLE algorithm is used to generate an optimal linear combination of these segmentations to provide a soft-labeling for each tissue class at each location in the test subject’s scan. The thirteen sets of prototype points are then edited in order to reject coordinate/label pairs that the STAPLE estimate considers improbable. For each prototype point, the probability  $p$  that the true segmentation at that coordinate matches the label in the prototype is taken from the STAPLE estimate and the prototype point is either kept or rejected, accordingly. Rejection could be based on a threshold on  $p$  or, as we do here, by generation of a uniform random number between  $[0, 1]$  and comparison to  $p$ . Training data that is inconsistent with the STAPLE reference standard is ultimately eliminated. This allows new estimation of the segmentations from the reduced training data, and the entire process is iterated to convergence. When a stable estimate is achieved, the STAPLE estimate can be turned into a “hard” decision by choosing the highest scoring label at each voxel location.

### 3 Results

Validation was performed by comparing the resulting segmentation to ground truth estimates in each of five cases. Five subjects were randomly selected from an institutional review board-approved study of neonatal brain development and none of these subjects was included in the atlas. Each subject was born premature at or before 28 weeks gestational age (GA). Two subjects exhibited no white-matter injury, two had evidence of moderate white-matter injury, and

**Table 1.** Results of automatic segmentation compared to single-slice manual segmentations using Dice similarity measure for cortical gray matter (cgm), myelinated white matter (myelin), unmyelinated white matter (uWM), and basal ganglia (bg). Values greater than 0.7 indicate excellent agreement.

| child   | cgm       | csf       | myelin    | uWM       | bg        |
|---------|-----------|-----------|-----------|-----------|-----------|
| child1  | 0.76      | 0.64      | 0.69      | 0.77      | 0.70      |
| child2  | 0.59      | 0.66      | 0.62      | 0.71      | 0.73      |
| child3  | 0.76      | 0.77      | 0.68      | 0.77      | 0.70      |
| child4  | 0.77      | 0.63      | 0.77      | 0.74      | 0.64      |
| child5  | 0.67      | 0.71      | 0.73      | 0.70      | 0.75      |
| mean±sd | 0.71±0.08 | 0.68±0.06 | 0.70±0.05 | 0.74±0.03 | 0.70±0.04 |



**Fig. 1.** Comparison of previously-published semi-automated method (left), manual segmentation (middle), and new fully-automatic method (right)

a fifth subject had severe white-matter injury. The scans used in this study were performed at term-equivalent age or approximately 42 weeks GA. These data provide a challenging test for segmentation algorithms.

In order to provide a credible estimate of ground truth, segmentations were performed by an expert rater using a combination of simple image processing operations (e.g. thresholding) and manual voxel-by-voxel editing of the results. Due to the labor intensive nature of this endeavor, a manual segmentation was generated for a single mid-coronal slice for each subject.

The segmentation quality was assessed by generation of the similarity coefficient due to Dice [13] comparing each child’s automatic segmentation to the single slice manual segmentation. Table 1 shows results for the single-slice comparisons. Dice values range from zero, indicating no overlap, to one, indicating perfect overlap, but cannot be interpreted as a percentage of overlap. Zijdenbos notes that values greater than 0.7 are considered excellent agreement [14].

Quite good results were achieved for all classes including the clinically-important differentiation of myelinated and unmyelinated white matter. The worst performance was seen in the cerebrospinal fluid (CSF) tissue classes, however this is not surprising for any intensity-based classifier. In newborn images, the amount of CSF can be quite low and small regions of CSF yield image samples with large “partial-volume” effects due to the mixing of signal with that from nearby tissues.

## 4 Discussion and Conclusion

Automated segmentation of brain MRI has been pursued at length with promising results (e.g. [15,16,17]). Relatively few groups, however, have attempted effective segmentation algorithms for MRI of newborns. Peterson and colleagues performed a slice-by-slice manual sampling of representative points to generate tissue class averages and then thresholded and manually edited the results [18]. Prastawa and colleagues report on an automated method using a three-subject atlas for gray matter, csf, and myelinated and unmyelinated white matter, but did not attempt sub-cortical gray matter segmentation [19]. We have previously

shown a method that generates accurate segmentations, but requires interactive training [10].

Cocosco has shown [20] a method that generates prototypes for each tissue class by randomly selecting coordinates corresponding to locations where the class prior is larger than a threshold. These prototypes are then refined (pruned) using a minimum spanning tree strategy. This methodology is used in [19]. Since this method produces simplified decision boundaries, it may not fully exploit the available contrast in the images. Grau also used an atlas scheme to automatically generate prototypes, but applied this solely to adult MRI [21].

We have presented a Bayesian framework for segmentation using a supervised classifier and a strategy for automatically generating training points from an atlas of previously-generated supervised classifications. The STAPLE algorithm was used to generate an optimal linear combination of repeated segmentations in order to reduce the possibility of random chance leading to a particularly poor choice of prototype samples and to aid in the editing of samples suited to individual anatomy that, despite affine registration, still differs from the anatomy under consideration. We assessed the quality of the resulting segmentations using single-slice results drawn by an expert. Overall the strategy worked well on very challenging data. Future work should include an assessment of this technique considering the inter- and intra-rater variability of human segmentations of the same data as well as an analysis of the robustness of the method to segmentation errors in the atlas and registration failures.

## Acknowledgments

This investigation was supported in part by NSF ITR 0426558, a research grant from the Whitaker Foundation, a research grant from CIMIT, grant RG 3478A2/2 from the NMSS, and by NIH grants R21 MH67054, P41 RR013218 and U41 RR019703.

## References

1. Martin, J.A., Hamilton, B.E., Sutton, P.D., Ventura, S.J., Menakcker, F., Munson, M.L.: Births: Final data for 2003. *National Vital Statistics Reports* **54**(2) (2005)
2. 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 Trans. Med. Imag.* **23** (2004) 903–921
3. Viola, P.A.: Alignment by maximization of mutual information. PhD thesis, MIT, Cambridge, MA, USA (1995)
4. Mangin, J.F.: Entropy minimization for automatic correction of intensity non-uniformity. In: *Math. Methods in Biomed. Imag. Anal.*, Los Alamitos, California, IEEE Computer Society (2000) 162–169
5. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. *IEEE Pat. Anal. Mach. Intel.* **12** (1990) 629–639
6. Gerig, G., Kubler, O., Kikinis, R., Jolesz, F.: Nonlinear anisotropic filtering of MRI data. *IEEE Trans. Med. Imag.* **2** (1992) 221–232

7. Yoo, T.S.: Insight into Images. A. K. Peters (2004)
8. Warfield, S.K., Kaus, M., Jolesz, F.A., Kikinis, R.: Adaptive, template moderated, spatially varying statistical classification. *Medical Image Analysis* **4**(1) (2000) 43–55
9. Warfield, S.K.: Segmentation of Magnetic Resonance Images of the Brain. PhD thesis, The University of New South Wales, Australia (1997)
10. Weisenfeld, N.I., Mewes, A.U.J., Warfield, S.K.: Segmentation of newborn brain MRI. In: Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging, Piscataway, NJ, IEEE (2006) Conference held April 6-9, 2006, Crystal City, Virginia, USA.
11. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley and Sons, New York, NY, USA (1973)
12. Leemput, K.V., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. *IEEE Trans. Med. Imag.* **18** (1999) 897–908
13. Dice, L.R.: Measures of the amount of ecologic association between species. *Ecology* **26** (1945) 207–302
14. Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C.: Morphometric analysis of white matter lesions in MR images: Method and validation. *IEEE Trans. Med. Imag.* **13**(4) (1994) 716–724
15. Wells, W.M., Kikinis, R., Grimson, W.E.L., Jolesz, F.: Adaptive segmentation of MRI data. *IEEE Trans. Med. Imag.* **15**(4) (1996) 429–442
16. Zhang, Y., Brady, M., Smith, S.M.: Segmentation of brain MR images thorough a hidden markov random field model and the expectation maximization algorithm. *IEEE Trans. Med. Imag.* **20**(1) (2002) 45–57
17. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M.: Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**(3) (2002) 341–355
18. Peterson, B.S., Anderson, A.W., Ehrenkranz, R., Staib, L.H., Tageldin, M., Colson, E., Gore, J.C., Duncan, C.C., Makuch, R., Ment, L.R.: Regional brain volumes and their later neurodevelopmental correlates in term and preterm infants. *Pediatrics* **111** (2006) 939–948
19. Prastawa, M., Gilmore, J.H., Lin, W., Gerig, G.: Automatic segmentation of MR images of the developing newborn brain. *Medical Image Analysis* **9** (2005) 457–466
20. Cocosco, C.A., Zijdenbos, A.P., Evans, A.C.: A fully automatic and robust brain MRI tissue classification method. *Medical Image Analysis* **7** (2003) 513–527
21. Grau, V., Mewes, A.U.J., Raya, M.A., Kikinis, R., Warfield, S.K.: Improved watershed transform for medical image segmentation using prior information. *IEEE Trans. Med. Imag.* **23**(4) (2004) 447–458