

# Automatic Segmentation of Newborn Brain MRI

Neil I. Weisenfeld<sup>a,b</sup>, Simon K. Warfield<sup>b</sup>

<sup>a</sup>*Department of Cognitive and Neural Systems  
Boston University  
Boston, MA, USA*

<sup>b</sup>*Computational Radiology Laboratory  
Radiology Department, Children's Hospital Boston  
and Harvard Medical School  
Boston, MA, USA*

---

## Abstract

Quantitative brain tissue segmentation from newborn MRI offers the possibility of improved clinical decision making and diagnosis, new insight into the mechanisms of disease, and new methods for the evaluation of treatment protocols for preterm newborns. Such segmentation is challenging, however, due to the imaging characteristics of the developing brain. Existing techniques for newborn segmentation either achieve automation by ignoring critical distinctions between different tissue types or require extensive expert interaction. Because manual interaction is time consuming and introduces both bias and variability, we have developed a novel automatic segmentation algorithm for brain MRI of newborn infants. The key algorithmic contribution of this work is a new approach for automatically learning patient-specific class conditional probability density functions. The algorithm achieves performance comparable to expert segmentations while automatically identifying cortical gray matter, subcortical gray matter, cerebrospinal fluid, myelinated white matter and unmyelinated white matter. We compared the performance of our algorithm with a previously published semi-automated algorithm and with expert-drawn images. Our algorithm achieved an accuracy comparable with methods that require undesirable manual interaction.

*Key words:* tissue segmentation, structural MRI, classification, newborn imaging

---

---

*Email addresses:* [weisen@crl.med.harvard.edu](mailto:weisen@crl.med.harvard.edu) (Neil I. Weisenfeld),  
[simon.warfield@childrens.harvard.edu](mailto:simon.warfield@childrens.harvard.edu) (Simon K. Warfield)

## 1. Introduction

Segmentation of newborn brain MRI has important implications for the study and treatment of brain injury and disorder due to prematurity. Shortly after an infant is born, neurodevelopment includes critically important maturational processes which may be measured quantitatively by brain imaging. While the normal developmental process is complex, measures such as relative and absolute brain tissue volumes have been shown to change in a stereotypical manner with age (Huppi et al., 1998). Premature birth exposes the developing brain to potentially harmful environmental factors at a time when critically important maturational processes are occurring. This disruption of the normal early developmental trajectory has been associated with significant adverse neurodevelopmental outcomes in later life (Woodward et al., 2006; Thompson et al., 2008; Woodward et al., 2005; Thompson et al., 2007; Shah et al., 2006; Peterson et al., 2003). We seek to develop better measures of neurodevelopment from MRI and to apply these to clinical decision making and the study of mechanisms of disease in particularly vulnerable children. We present here a novel, automatic algorithm for generating quantitative tissue volume measurements from newborn brain MRI.

Segmentation of newborn brain MRI is different from segmentation of adult brain MRI. In adults, tissue is usually classified as either gray matter, white matter, or cerebrospinal fluid based solely on image-intensity. In newborn MRI, it is necessary to identify additional tissue classes in order to best characterize brain development. In the newborn infant, the process of white-matter myelination progresses rapidly throughout the first year of life and is an important biomarker for brain maturation. Furthermore, cortical and subcortical gray matter are phylogenically and developmentally distinct and warrant separate analysis. Therefore, we seek to differentiate myelinated from unmyelinated white matter and cortical from sub-cortical gray matter. Further parcellation of these tissues into sub-structures is valuable and important, but is not considered here. Figure 1 shows an aligned T1- and T2-weighted MRI from a full-term newborn infant and an associated tissue-class segmentation.

Segmentation of newborn brain MRI is considerably more difficult than segmentation of adult MRI due to reduced contrast and increased noise in images from neonates (Mewes et al., 2006)

as well as inverted contrast between gray matter and white matter(Xue et al., 2007). Intensity-based classification relies on contrast between tissue types adjacent in feature space and adequate signal compared to image noise. In newborn MRI, contrast between gray matter and white matter is reduced compared to adult MRI. Statistical classification identifies an optimal boundary, in feature space, between tissue types. The separability of two tissue types or classes is related to the overlap between classes in feature space. Contrast between tissue classes in newborn MRI is reduced compared to adult MRI because the majority of white matter is as-yet unmyelinated and has a water content closer to that of gray matter than in adults and adolescents. Furthermore, this contrast changes with age reaching a point at about 9 months of age when gray matter and white matter are roughly isointense (Barkovitch, 2005) on MRI and images therefore cannot be classified based on intensity alone. The inversion of contrast between white matter and gray matter, compared to adult MRI, is also a problem given the limited resolution of neonate MRI. Limited resolution, compared to the size of the structures being imaged, leads to significant partial volume effects. With the intensity of unmyelinated white matter being somewhere between gray matter and cerebrospinal fluid, partial volume averaging between these two tissues is often misclassified as unmyelinated white matter(Xue et al., 2007).

In MR imaging there is a tradeoff between scan duration, voxel size, and signal to noise (SNR). Each of these factors is significantly more challenging in newborn brain imaging. Because infants' brains are smaller, it is necessary to create higher resolution images in order to appreciate the fine structure of the newborn brain. Smaller voxels lead to noisier images, and the reduced contrast in newborn images compared to adults then leads to reduced contrast to noise ratio (CNR) in images. This can be mitigated by increasing scan duration, however this too is undesirable. Due to concerns about the risks associated with anesthesia, our research MR studies are performed on unsedated infants who have been fed and then swaddled to prevent movement (e.g. Mathur et al., 2008). There is usually a narrow window after feeding during which the infant will sleep, so it is important to keep scan times to a minimum. Parallel imaging and the next generation of multi-channel imaging coils will help greatly in this regard allowing for an increase in SNR and a decrease in scan time. Sensitive analysis methods are needed presently to overcome these limitations.

Segmentation in adult MRI has been the topic of a great deal of study over the past fifteen years or so, with most successful algorithms employing voxelwise, intensity based classification. The basic strategy is usually based on statistical classification theory, with a “feature vector” consisting of signal intensities from one or more types of MR image. Given a categorical random variable  $S$  referring to tissue class label and continuous random variable  $G$  referring to grayscale image intensity, we define the distributions  $\Pr(S = s) = p(s)$ ,  $\Pr(S = s|G = \mathbf{g}_i) = p(s|\mathbf{g}_i)$  and  $\Pr(G = \mathbf{g}_i|S = s) = p(\mathbf{g}_i|s)$ . A Bayesian formulation of voxelwise, intensity-based classification can then be posed as follows:

$$p(s|\mathbf{g}_i) = \frac{p(\mathbf{g}_i|s)p(s)}{\sum_{s'} p(\mathbf{g}_i|s')p(s')} \quad (1)$$

where the value  $\mathbf{g}_i$  is a vector of image intensities representing a point in the feature space for classification. In our case, the vector is a sample, from the same coordinate  $i$ , from an aligned T1-weighted MRI and a corresponding T2-weighted MRI. The probabilities over all tissue types sum to one at each voxel coordinate:  $\sum_s p(s|\mathbf{g}_i) = 1$ . A “soft” segmentation is  $p(s|\mathbf{g}_i)$  or the probability of each tissue class  $s$ , at a given voxel coordinate  $i$ , given the measured image intensities  $\mathbf{g}_i$ . A “hard” labeling of the image can be made by choosing the most probable of the tissue classes at each coordinate  $i$ :  $\hat{s} = \arg \max_s p(s|\mathbf{g}_i)$ . In adults and adolescents, it is possible to distinguish gray matter, white matter, and CSF based on T1 and T2 weighted scans. Vannier and colleagues demonstrated as early as 1985 such a scheme with both supervised and unsupervised classification on MRI brain data using systems designed for satellite imaging (Vannier et al., 1985).

Aside from the associated pre- and post- processing machinery, the key difference in classification strategies is the estimation of the likelihood  $p(\mathbf{g}_i|s)$  and the prior model  $p(s)$ . Wells et al. introduced an algorithm suitable for images corrupted by a spatially varying intensity artifact (e.g. due to RF sensitivity variations throughout the imaging volume) and devised an Expectation-Maximization (EM) (Dempster et al., 1977) algorithm for simultaneously estimating the posterior probabilities  $p(s|\mathbf{g}_i)$  and the parameters of a model of the intensity artifact. They modeled the likelihoods both parametrically as Gaussians and non-parametrically using Parzen windowing (Wells et al., 1996). Warfield introduced a non-parametric strategy based on Cover’s density estimation

(Cover, 1968) and used an atlas of spatially varying features to augment the classification feature space (Warfield et al., 2000). Van Leemput extended Wells' EM scheme to also estimate the means and variances of tissue class Gaussians and also to include both a spatially varying prior and a Markov random field (MRF) spatial homogeneity constraint, replacing the global prior  $p(s)$  with the product of a spatially varying prior  $p_i(s)$  and a prior based on the MRF neighborhood  $p_{\partial i}(s)$  (Van Leemput et al., 2001). Such MRF techniques are frequently used in brain segmentation to help remove isolated voxels of misclassification by altering the prior probabilities of tissue classes to favor assigning a voxel to the same class as its neighbors. More recent work has focused on further subdivision of the brain into functional regions with some work on surface-based methods for parcellating the cortex and others working on volumetric, whole-brain techniques (e.g. Dale et al., 1999; Pohl et al., 2007; Makris et al., 2005). Often these techniques require an intensity-based classification as a starting point.

Updating the model to include a spatially varying prior and a Markov random field prior model results in the following Bayesian formulation of voxelwise, intensity-based classification:

$$p(s|\mathbf{g}_i) = \frac{p(\mathbf{g}_i|s)p_i(s)p_{\partial i}(s)}{\sum_{s'} p(\mathbf{g}_i|s')p_i(s')p_{\partial i}(s')} \quad (2)$$

Despite the success of these methods in adult imaging, less attention has been directed toward the segmentation of images from newborns, but several authors have introduced automatic and semi-automatic segmentation schemes in recent years. Prastawa and colleagues introduced a method that generates a map of gray matter, myelinated white matter, unmyelinated white matter, and cerebrospinal fluid in newborn brain MRI (Prastawa et al., 2005). The authors use an EM-estimator similar to the method of Van Leemput. Since these methods are sensitive to initial estimates of the Gaussian tissue class model parameters, the authors use a robust estimation scheme which uses atlas-guided tissue samples, outlier rejection, and a clustering approach for generating two tissue classes (myelin and unmyelinated white matter) from the single white-matter tissue class in their atlas. The resulting segmentation is then used to seed a second segmentation step based on kernel density estimation and a minimum-spanning tree clustering scheme based on the work of Cocosco et al. for outlier rejection (Cocosco et al., 2003). The authors present an analysis

of their work comparing results in a single slice from four subjects to the results of two raters' manual segmentations for gray matter, CSF, and myelinated and unmyelinated white matter.

More recently, Anbeek and colleagues presented a method which generates segmentations of CSF, white matter (combined), and cortical and sub-cortical gray matter (Anbeek et al., 2008). For a cohort of 13 subjects, two raters manually thresholded T1-weighted and T2-weighted scans in order to generate binary segmentations for each tissue class, which were then combined into a multi-label segmentation. An automatic segmentation strategy based on  $k$ -NN classification was applied to each of these subjects with a feature space comprising both MR image intensity values and voxel coordinates. The 13 subjects were segmented in a leave-one-out fashion with twelve subjects acting as training data for segmentation of the 13th. The authors report results with this within-cohort training strategy.

Xue and colleagues presented a method for segmentation of gray matter, white matter (combined), and CSF (Xue et al., 2007). Their method is based on parametric density estimates and an EM algorithm with the addition of several steps addressing problems with whole-brain global density estimates. Within the EM estimation, the authors add a Markov random field prior in order to enforce a spatial homogeneity constraint. Since the cortical ribbon is so thin in newborn imaging, the so-called "partial volume averaging" MRI effect presents a more significant problem than in adult brain segmentation at the same resolution. Xue presents a novel technique for removing the effects of partial volume averaging by exploiting the knowledge that misclassification occurs in a predictable way (e.g. CSF and gray matter "average" into an intensity similar to white matter). The authors also address the difficulty of performing intensity-based classification using global statistics by introducing a local parcel structure which is used to constrain Gaussian parameter estimates to a region of the brain.

The motivation for development of our algorithm was two-fold. First, there is an increased prevalence of high-quality newborn MRI data which makes it desirable to have a fully-automatic method that is capable of taking data from scanner to segmentation without manual intervention. Second, it is necessary to differentiate myelinated white matter, unmyelinated white matter, cerebrospinal fluid, cortical gray matter and subcortical gray matter. Myelination has been shown to be an important biomarker for neurodevelopment and disease (Barkovitch, 2005) and we therefore

have an interest in differentiating it from unmyelinated white matter. Furthermore, the distinct role and architecture of subcortical gray matter make differentiating between cortical and subcortical tissues desirable. Since none of the methods discussed previously met all of these criteria, we were motivated to develop the method discussed here.

## **2. Methodology**

### *2.1. Overview*

We propose here a novel, adaptive classifier fusion algorithm for segmenting newborn brain MRI. Our goal is to take aligned T1- and T2-weighted MR images of a subject and to assign a label of either cortical gray matter, subcortical gray matter, unmyelinated white matter, myelin, cerebrospinal fluid, or background/non-brain to each voxel in an image. Accuracy in this task requires good estimates of the probability distributions of intensities associated with each tissue class. Our algorithm estimates tissue class intensity distributions non-parametrically and is therefore able to represent complex distributions. Construction of such density estimates requires a set of samples of representative tissue MR intensity values, called prototypes, for each tissue class.

Our algorithm starts with a library of template MRI images representing typical subjects of similar developmental age to the study subject. Each template is associated with a large number of manually selected tissue class prototypes. Prototypes are tissue labels, at a particular coordinate in the image, that an expert has chosen as representative of a tissue class. Typically there are several thousand such prototypes per template image. Each template image is then registered to the subject being segmented (the “study subject”), and each template’s prototype labels are transferred, through the registration process, to the study subject. The algorithm is initialized with one prototype list for each template image, tailored to the study subject’s anatomy through registration. Prototype lists each lead directly to non-parametric density estimates and these, in turn, lead to segmentations of the study subject. These segmentations are then fused in order to improve their accuracy.

The fused classification provides a probabilistic labeling of the study subject’s tissues, but may contain errors due to structural differences between the templates and the study subject that cannot

be accounted for by registration. The original prototype lists are edited in order to reject prototypes that are likely to have an erroneous label, given the probabilities of the fused classification. The edited prototype lists are then used to generate a new set of segmentations which, in turn, leads to a new fused classification. This cycle of edit prototypes, segment, and fuse forms an iterative process which our algorithm repeats until convergence. Figure 3 gives a graphical overview of this process.

## 2.2. Classification

We refer to the true underlying segmentation as  $T$  and refer to its value at a particular voxel index  $i$  as  $T_i$ . A probabilistic segmentation is therefore  $p(T_i = s)$  for all tissue labels  $s$ .  $p(T_i)$  is estimated through a classifier fusion strategy and we introduce several more variables to support this. A number of estimated segmentations are to be fused. These segmentations are indexed with  $j$  and voxel coordinates within a segmentation with  $i$ , so that  $D_{ij}$  is the segmentation label at coordinate  $i$  in estimated segmentation  $D_j$ . We refer to the collection of segmentations  $D_j$  as  $\mathbf{D}$ .

Our algorithm fuses these estimates in an optimal way such that less-reliable estimates are given less weight than more reliable ones. The probabilities of the true, unknown segmentation are estimated as:

$$p(T_i = s | \mathbf{D}, \Theta) = \frac{p(T_i = s) \prod_j p(D_{ij} | T_i = s, \theta_j)}{\sum_{s'} p(T_i = s') \prod_j p(D_{ij} | T_i = s', \theta_j)} \quad (3)$$

The variable  $\Theta$  is a collection of matrices  $\theta_j$ , one for each noisy segmentation estimate. Each  $\theta_j$  is a matrix of probabilities associated with the types of errors that the estimate might contain with respect to the true, unknown segmentation. The elements of  $\theta_j$  are  $\theta_{js'}$  =  $p(D_j = s' | T = s)$ . The diagonal elements where  $s = s'$  are the probability of correctly identifying each label, while the off-diagonal elements ( $s \neq s'$ ) represent each possible disagreement between the segmentation  $D_j$  and the true, unknown segmentation. The STAPLE algorithm (Warfield et al., 2004) is used to solve Equation 3 without *a priori* knowledge of the true, unknown segmentation  $p(T)$  and below we address how we estimate the various components of this equation.

Each of the estimated segmentations  $D_j$  is the result of classification with a set of class-conditional probability densities  $p_j(\mathbf{g}_i|s)$  arising from separate prototype lists. Each segmentation is:

$$D_{ij} = \arg \max_s p_j(s|\mathbf{g}_i) \quad (4)$$

$$= \arg \max_s \frac{p_j(\mathbf{g}_i|s)p_i(s)p_{\partial i}(s)}{\sum_{s'} p_j(\mathbf{g}_i|s')p_i(s')p_{\partial i}(s')} \quad (5)$$

Where  $\mathbf{g}_i$  is a 2-vector of MR intensities, from an aligned T1-weighted and T2-weighted image, at a given coordinate  $i$ .  $p_j(\mathbf{g}_i|s)$  is the class-conditional probability density function estimates arising from the prototypes registered from template subject  $j$ ,  $p_i(s)$  is a spatially varying prior from an atlas of prior probabilities, and  $p_{\partial i}(s)$  is a Markov random field prior in a neighborhood centered around coordinate  $i$  in order to model spatial homogeneity.

What differs between each estimated segmentation  $D_j$  is the estimate of the class-conditional probability density function  $p_j(\mathbf{g}_i|s)$ . Our algorithm uses a non-parametric density estimate for this task. This strategy requires a list of prototypes for each tissue class and generates a density estimate based on this list. Prototypes are of the form  $(i, s, \mathbf{g}_i)$  and are triplets of coordinate, tissue class label, and associated intensity samples. In a supervised classification strategy, these could be points hand chosen by an expert. Here we choose them by non-rigid registration of template subject images to our study subject.

Each template has an associated list of prototype positions and labels  $(i'_k, s_k), k = 1 \dots n$  that were hand-chosen by an expert as representative locations of each tissue class. Several hundred points were chosen for each tissue class, on average, from each template image. Each template image is then registered to the study subject using mutual-information based registration between the template's structural image (usually a T1-weighted MRI) and the study subject's T1-weighted MRI. First an initial rigid registration is performed, then affine, and finally a non-rigid B-spline based free-form deformation. The registration of template to the study subject yields a transformation of voxel coordinates  $\mathcal{T}$ . This transformation is then applied to the list of prototype coordinates associated with that template subject. If  $i = \mathcal{T}\{i'\}$  so  $i$  is coordinate  $i'$  transformed, then  $\mathcal{T}$  can be used to transform the prototypes  $(i'_k, s_k)$  to  $(i_k, s_k)$  for all points  $k = 1 \dots n$ .

Through registration, our algorithm projects the prototype coordinates of a template subject onto the study subject. The grayscale image data from the study subject is sampled using the transformed coordinates and their associated labels. Sampling turns pairs of prototype coordinates and labels  $(i_k, s_k)$  into triplets of coordinates, labels, and grayscale intensity values  $(i_k, s_k, \mathbf{g}_{i_k})$ . A list of such prototypes generated from each template subject  $L_j = \{(i_1, s_1, \mathbf{g}_{i_1}) \dots (i_n, s_n, \mathbf{g}_{i_n})\}$  is formed. These prototypes are used to directly estimate  $p_j(\mathbf{g}_i|s)$  using the method due to Cover (Cover, 1968). The resulting  $p_j(\mathbf{g}_i|s)$  is tailored to the grayscale T1 and T2 data from the study subject and varies based on how well template subject  $j$  could be registered to the study subject. Segmentations  $D_j$  are formed by Equation 5 and combined into a single, estimated segmentation with Equation 3.

At this stage, the algorithm has an estimated segmentation and it is therefore possible to revisit each of the prototype lists  $L_j$  and remove prototypes that are inconsistent with the current estimate of the underlying subject anatomy. Such an edited prototype list will improve our density estimates and the resulting segmentations. For each prototype list  $L_j$ , and for each prototype  $k$ , a prototype  $(i_k, s_k, \mathbf{g}_{i_k})$  is evaluated. A random number is generated  $r \in [0, 1]$  and if  $r > p(T_i = s|\mathbf{D}, \Theta)$ , then that prototype is removed from its list. So if a prototype at coordinate  $i$  indicates  $s$  is gray matter and  $p(T_i|\mathbf{D}, \Theta)$  specifies gray matter probability at 0.8 at that coordinate, then we generate a random number such that there is a  $p = 0.8$  chance of keeping that prototype and a  $1 - p = 0.2$  chance of its removal.

Given our edited lists of prototypes  $L_j$ , our algorithm generates new segmentation estimates  $D_j$  and creates a new estimate of the true, underlying segmentation. This process of editing prototypes, generating new segmentations, and combining segmentations is repeated until convergence, usually about 4 or 5 iterations. We measure convergence by monitoring the root-mean-squared difference between successive segmentation estimates and stop once this falls below a threshold.

### 2.3. Partial Volume Correction

Because of the small size of the cortex relative to the voxel size, partial volume averaging is a more significant problem in newborn brain segmentation than in adults. Xue and colleagues described the nature of this problem in newborn brain MRI and implemented a targeted solution

(Xue et al., 2007). Similarly, our algorithm includes a relaxation labeling strategy (Rosenfeld et al., 1976) designed to identify and correct the specific combination of events described by Xue. For each voxel, the probability mass was summed in the 27-voxel (radius=1) neighborhood around the center voxel. If the probability mass was both greater than 20% gray matter and 20% CSF, then the probability of unmyelinated white matter at the center voxel was reduced by a tunable fraction. For our experiments, this was set to 0.1 per iteration and the relaxation scheme set to run for 30 iterations.

#### 2.4. *Spatial Priors and Spatial Homogeneity*

When tissue classes we seek to differentiate coincide in intensity, but differ in spatial arrangement, our algorithm resolves this discrepancy using an atlas of spatially varying prior probabilities. Specifically, cortical and sub-cortical gray matter appear isointense on MRI and myelin can be similar in intensity with partial volume averaging between gray matter and unmyelinated white matter. Equation 5 includes a spatially-varying prior probability as  $p_i(s)$ .

Our term newborn atlas is shown in Figure 2. For the construction of this atlas, scans were taken from fifteen 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 (Warfield et al., 2000). The resulting segmentations were brought into alignment using affine transformations derived with the method outlined in (Weisenfeld and Warfield, 2007). Once aligned, an atlas of voxelwise probabilities was created by calculating the frequency that each tissue class label appears at each voxel coordinate. An image of tissue-class frequency was constructed for each of our tissue classes as well as for a brain/non-brain classification. Additionally, the aligned T1-weighted MR scans were standardized to have equal intensity mean and variance and then averaged in order to generate a mean structural MRI used for alignment to the study subject. The use of an atlas generated from appropriate-for-gestational-age preterms, when analyzing full term children at the same gestational age, should not be an issue as there is an affine head shape difference (Mewes et al., 2007) that is discounted by the registration methods employed.

Separately, our algorithm models spatial homogeneity as a Markov Random Field in a neigh-

borhood  $\partial i$  around each coordinate  $i$ . This is included in Equation 5 as  $p_{\partial i}(s)$ . Our algorithm solves a Mean Field approximation (Elfadel, 1993) to this problem using the method in (Warfield et al., 2004). We solve this estimator after partial volume correction, on the final estimated segmentation. This serves to remove small, isolated islands of misclassification due to residual image noise.

### 2.5. Noise Correction

Segmentation using global image intensity statistics requires that the tissue of the same composition will produce similar intensities everywhere within the image. By definition,  $p(\mathbf{g}_i|s)$  is therefore considered stationary and does not depend on  $i$ . This assumption does not hold, in practice, because MR imagers are not equally sensitive to signal arising from everywhere within the field of view of the magnet and because of thermal noise in the measured signal. We assume the following model of image formation:

$$\Gamma(i + \delta(i)) = \beta(i) [g(i) + \epsilon(i)] \quad (6)$$

where  $\Gamma(i)$  represents the measured image intensities at image coordinate  $i$  and  $g(i)$  represents the desired, noise-free signal. Since our T1 and T2 weighted images may not be aligned, we introduce a mapping between voxel coordinates  $\delta(i)$ . A spatially varying, multiplicative change in signal due to RF sensitivity issues and field inhomogeneity is represented by  $\beta(i)$  and additive noise by  $\epsilon(i)$ .

Wells' EM-based segmentation scheme, and subsequently others, have attempted to factor out  $\beta$  from Equation 6 by solving for a gain or bias correction field during the classification process (Wells et al., 1996). Several authors have also developed pre-processing methods that operate on the data prior to and separate from classification (e.g. Sled et al., 1998; Likar et al., 2001). We have implemented and extended one such method proposed by (Viola, 1995) and further developed in (Mangin, 2000). This technique assumes that the intensity non-uniformity is smoothly varying compared to the contrast of interest within the images and that it can be modeled as a field of multiplicative correction factors. The key observation made by Viola is that such a smoothly varying gain serves to increase entropy in the image by enlarging the range of intensity values associated with a given tissue type. Correction involves solving for a multiplicative adjustment

field that, subject to constraints on its form, reduces entropy in the global image histogram. This step is performed prior to knowledge of the distribution of specific tissue classes and we've found that, in extreme cases, correction also aids in registration tasks involving these images.

Thermal noise in MR images is modeled as additive, positionally dependent noise ( $\epsilon$  in Equation 6), and is another effect which we address during preprocessing. For reasons discussed earlier, brain MR images from newborns tend to be noisier than images from adult brains. Simple low-pass filtering of the images would be effective, but would also blur the edges between adjacent tissues of differing types making segmentation difficult. Anisotropic diffusion has been used as a method for smoothing images while preserving edges represented by strong intensity gradients and we adopted this strategy for filtering our images (Perona and Malik, 1990; Gerig et al., 1992) prior to segmentation.

## 2.6. Registration and Atlas Alignment

Since the subjects are unседated, it is possible that the child will move, or be moved, between the acquisition of the T1 and T2 images. It is necessary to align the images so that each voxel coordinate  $i$  addresses the same underlying tissue in each image. We assume that the T1 image is fixed ( $\delta(i) = 0$ ) and rigidly align the T2 image using maximization of mutual information (Mattes et al., 2001).

It is necessary to align the atlas to the study subject and this was done by aligning the average (intensity mean) atlas T1-weighted MRI with the subject's own T1-weighted MRI. The final atlas registration was the composition of three mutual information-based registration (Mattes et al., 2001) processes with increasingly complex transform properties. An initial rigid registration was used to generate a rigid body alignment of the atlas and subject and this result was used to initialize a similar registration which derived a best affine fit between the two scans. Affine registration is enough to account for gross size and shape differences between the subjects and the atlas, but does not account for local shape differences. A final non-rigid registration step based on B-spline interpolation (e.g. Rueckert et al., 1999) was used to capture these local shape differences.

### 2.7. Data Acquisition

Data were acquired on a 1.5T clinical MR scanner from GE Medical Systems (Wakeusha, WI, USA) using an 8-channel receiver coil designed for adult head imaging. The present study utilized 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. Both sequences were coronally acquired with a matrix size of 256x256 and a field of view of between 18 and 22cm. Slice thickness was 1.5mm for the SPGR sequence and 2.0mm for the FSE sequence. The SPGR parameters were TR30/TE9 and a flip angle of 45 degrees. For the FSE, the parameters were TR3000/TE140 and the echo train length was 15. The subjects were unседated infants wrapped with a blanket and their heads were fixed with a vacuum pillow. All subjects had clinical MRIs which were read as normal and were either pre-term infants scanned at term equivalent age or term infants scanned shortly after birth.

### 2.8. Expected Run-Time

On a typical 2.0GHz Intel PC, our algorithm takes approximately 20 minutes for pre-processing and 25 minutes for rigid, affine, and non-linear registration of template subjects to the target, per template subject. We typically run on multi-core, multi-processor machines with one registration per core. Segmentation takes approximately one hour, with an additional fifteen minutes for post-processing.

## 3. Experiments and Results

In order to illustrate the effect that prototype editing has on the distribution of tissue class prototypes in featurespace, we plotted in Figure 4(a) the results of projecting the prototypes  $L_j$  from a single template subject by registration (left) and the resulting segmentation  $D_j$  (right). Below, in Figure 4(b), we show the featurespace after four iterations of the editing process. The initial projected prototypes show clusters that severely overlap one another. Even with the Bayes optimal decision boundary, a number of prototypes are misclassified and the class-conditional density estimates will suffer, as illustrated by the misclassifications in the corresponding segmentation.

		train	
		$\mathcal{M}$	$\mathcal{A}$
test	$\mathcal{M}$	$0.95 \pm 0.02$	$0.93 \pm 0.02$
	$\mathcal{A}$	$0.95 \pm 0.01$	$0.97 \pm 0.01$

Table 1: mean $\pm$ standard deviation across subjects of average predictive values (PV) for various combinations of training and test sets for classification based on Cover density estimation.  $\mathcal{M}$  is a manually generated (hand-chosen) training set while  $\mathcal{A}$  is the automatically generated data set from the algorithm described herein. When the training and test set are the same, a leave-one-out cross validation scheme was used. The PVs show a high level of internal consistency in both the manual and automatically generated data sets as well as excellent agreement between the data sets.

Figure 4(b) shows the result after editing. The tissue types have been largely resolved into individual, separable clusters and the segmentation is greatly improved. The final result shown at right in Figures 1 and 5 is due to fusion of 13 such candidate segmentations.

The impact of the partial volume correction is shown in Figure 5. The arrow in each image points to an area of concern. On the T2-weighted MRI shown at left, cortical gray matter is shown “averaged” with the surrounding cerebrospinal fluid to yield signal between that of gray matter or CSF. This signal has the intensity properties of non-myelinated white matter and leads to a rim of misclassification in segmentations. The middle image shows the intensity-based classification without a Markov random field prior or partial volume correction. The misclassification is shown as a rim of tissue marked red, for unmyelinated white matter, outside of the cortex. At right, the image shows the result after addition of partial volume correction and a Markov random field prior as discussed in the previous section. The red rim of misclassification around the cortex and within the ventricles is largely eliminated.

We sought to validate the prototype choices that our automated method makes by comparison with hand-chosen tissue class prototypes. We performed a series of classification experiments where our classifier is trained with one data set (the “training” set) and is used to classify a different set of points (the “test” set). Classification involved Cover density estimation using the training set and labeling each point in the test set with the most probable tissue type as in 1. For each of

Subject	$\mathcal{M}$	test				
		$\mathcal{A}^0$	$\mathcal{A}^1$	$\mathcal{A}^2$	$\mathcal{A}^3$	$\mathcal{A}^{\text{FINAL}}$
1	0.98	0.66	0.77	0.86	0.91	0.93
2	0.96	0.65	0.77	0.87	0.92	0.94
3	0.96	0.66	0.79	0.89	0.94	0.95
4	0.96	0.65	0.78	0.87	0.92	0.94
5	0.94	0.66	0.78	0.87	0.93	0.95
6	0.94	0.67	0.80	0.90	0.94	0.95
7	0.94	0.66	0.79	0.89	0.94	0.96
8	0.94	0.70	0.82	0.91	0.96	0.97
9	0.94	0.67	0.79	0.89	0.93	0.95
10	0.96	0.69	0.80	0.89	0.94	0.96
mean±sd	$0.95 \pm 0.02$	$0.67 \pm 0.02$	$0.79 \pm 0.02$	$0.88 \pm 0.02$	$0.93 \pm 0.01$	$0.95 \pm 0.01$

Table 2: The effect of the training data editing process. Expert hand-chosen (manual) training data  $\mathcal{M}$  is compared first with itself in a leave-one-out manner. Reported is average predictive value across tissue classes. The initial automatic training set generated by registration alone is compared to  $\mathcal{M}$  in column  $\mathcal{A}^0$ . Each subsequent column  $\mathcal{A}^1, \dots, \mathcal{A}^{\text{FINAL}}$  shows an iteration of the training prototype editing process with a subsequent improvement in PV. The result is a final classifier training set consistent with the hand-chosen training set. Summary values are presented as mean and standard deviation (sd).

ten subjects, we performed manual training of the density estimator by hand-choosing one to two hundred training points per tissue class. For each subject, from the manual training set  $\mathcal{M}$  and the final, automatically generated training set  $\mathcal{A}$ , we generated measurements of predictive value (PV) by training with one set and testing with the other. For comparison purposes, we also performed a “leave-one-out” style classification experiment where the hand-chosen prototypes  $\mathcal{M}$  were used as both the training and test set, however the training set was altered for each classification in order to remove the particular point being classified. In this way, the training set never included the data being classified. We also performed a similar leave-one-out experiment using the automatically generated prototypes  $\mathcal{A}$ . These experiments give both a measure of internal consistency in each of these sets and, in the case of the hand chosen prototypes, a standard with which to compare. The results are shown as mean $\pm$ standard deviation (SD) across subjects for all combinations of training and test set in Table 1. Predictive values are averaged across tissue types in order to create a summary value. Only tissue types distinguishable based on intensity alone (all gray matter, unmyelinated white matter, cerebrospinal fluid) were tested here. The distinction between cortical and subcortical gray matter and the identification of myelin is tested in later experiments where the classification procedure incorporates spatial prior probabilities.

In order to demonstrate the impact of the prototype editing process, compared with training by registration alone, a similar analysis was performed by training with the hand-chosen data  $\mathcal{M}$  and classifying the automatically generated data at each stage in the editing process. As a reference,  $\mathcal{M}$  is compared to itself in a leave-one-out manner. The results for each subject are shown in Table 2. Column  $\mathcal{A}^0$  shows the agreement (PV) between  $\mathcal{M}$  and  $\mathcal{A}$  immediately after the registration process and before any prototypes have been edited out. Each of the following columns shows a subsequent iteration of the editing process, ending with the final training set in the last column, labeled  $\mathcal{A}^{\text{FINAL}}$ . The result clearly improves with each iteration of the algorithm.

Next, the ten subject MRI scans were segmented using our complete algorithm. The pre-processed data were then segmented using a previously validated, non-iterative scheme (Warfield et al., 2000) whereby classifier training data is chosen manually, by an expert. Table 3 shows Dice (Dice, 1945) overlap measures comparing the results of the fully automatic segmentation to results achieved with the previously published, semi-automated technique.

Subject	CGM	CSF	myelin	UMWM	SCGM
1	0.95	0.82	0.63	0.95	0.92
2	0.89	0.94	0.80	0.89	0.86
3	0.89	0.93	0.71	0.91	0.89
4	0.84	0.95	0.70	0.85	0.81
5	0.92	0.93	0.77	0.93	0.88
6	0.89	0.98	0.77	0.93	0.85
7	0.93	0.96	0.70	0.96	0.89
8	0.91	0.97	0.79	0.93	0.87
9	0.87	0.94	0.66	0.91	0.80
10	0.94	0.80	0.67	0.95	0.88
mean±sd	0.90 ± 0.03	0.92 ± 0.06	0.72 ± 0.06	0.92 ± 0.03	0.86 ± 0.04

Table 3: Dice overlap measure comparing automatic segmentation to a segmentation generated with hand-chosen prototypes. Summary values are presented as mean and standard deviation (sd).

Finally, in order to investigate the level of agreement between automatically-generated segmentations and segmentations *drawn* manually by expert raters, a study was performed where a single, mid-coronal slice was chosen for each of four subjects under study. This same slice was segmented by each of three experts who each repeated the exercise five times. The experts were instructed to move from subject to subject, rather than repeating the same segmentation five times in a row, in order to avoid any training effect. The mid-coronal slices were chosen to illustrate segmentation of each of the five tissue types being investigated. For each of four subjects, three experts segmented each scan five times. This led to fifteen repeat manual segmentations of each subject.

Our segmentation strategy was used to generate a full, three-dimensional segmentation of each of four subjects under study and the slice corresponding to the manual segmentation for that subject was extracted for comparison with the experts. The atlas of prior probabilities and the template segmentations and their associated example prototypes, were made from a group of thirteen newborn infants scanned at approximately 40 weeks gestational age. None of the subjects under study was included in the atlas. An expert consensus segmentation was generated using the method in (Warfield et al., 2004) and for each subject, the manual segmentation was compared with the expert consensus. For comparison purposes, each of the experts was, in turn, compared with the expert consensus and the results are reported as predictive value for the manual segmentation and mean $\pm$ standard deviation predictive value for the expert drawn segmentations. The results are listed in Table 1.

#### **4. Discussion**

The primary contribution of this work is a novel algorithm for segmenting newborn brain MRI into several tissue types. Our method operates by training a supervised classifier using spatial information through a registration and subsequent prototype editing process. A library of template subjects is used to represent the variation in anatomy between subjects. Each template subject is registered to the study subject and that transformation is used to project prototype labels used to train a supervised classifier and, in turn, generate a separate candidate segmentation of the test subject. The candidate segmentations are fused and the resulting estimated true segmentation

Subject		CGM	CSF	myelin	UMWM	SCGM
1	experts	$0.86 \pm 0.06$	$0.89 \pm 0.05$	$0.81 \pm 0.11$	$0.85 \pm 0.05$	$0.86 \pm 0.08$
	automatic	0.75	0.96	0.86	0.79	0.96
2	experts	$0.87 \pm 0.06$	$0.93 \pm 0.02$	$0.96 \pm 0.05$	$0.87 \pm 0.06$	$0.90 \pm 0.12$
	automatic	0.77	0.98	0.96	0.72	0.74
3	experts	$0.90 \pm 0.04$	$0.91 \pm 0.02$	$0.77 \pm 0.06$	$0.88 \pm 0.03$	$0.91 \pm 0.03$
	automatic	0.77	0.97	0.81	0.78	0.95
4	expert	$0.87 \pm 0.08$	$0.91 \pm 0.02$	$0.81 \pm 0.06$	$0.87 \pm 0.04$	$0.94 \pm 0.04$
	automatic	0.84	0.95	0.69	0.70	0.94

Table 4: Intra-rater mean $\pm$ standard deviation of predictive values (PVs) by study subject and tissue class compared with the PV for the automatic classification result. Tissue classes are cortical gray matter (cgm), cerebrospinal fluid (CSF), myelin, unmyelinated white matter (umwm), and sub-cortical gray matter (scgm).

is used to guide removal of prototypes inconsistent with the study subject’s anatomy. The final stages of the process (prototype refinement, segmentation, and combination) are repeated until the process converges to a final result.

Prototypes that are incorrectly labeled with respect to the underlying anatomy lead to poor density estimates and misclassification. Our algorithm estimates an initial set of tissue class prototypes and then refines these estimates. Figures 4(a) and 4(b) illustrate the effect that prototype editing has on the featurespace for classification. Unlike the method used in Prastawa et al. (Prastawa et al., 2005), editing is not performed in featurespace. Our algorithm uses a consensus estimate of the underlying segmentation in order to accept or reject prototypes based on their consistency with the anatomy. This is in contrast to clustering methods, such as Cocosco et al. (Cocosco et al., 2003), which ignore anatomical information and instead make assumptions about the geometry of the tissue class distributions. By not making such assumptions, our algorithm is capable of dealing with complex multi-modal distributions and complicated decision boundaries within featurespace.

Our strategy for choosing prototypes compares favorably with hand-chosen prototypes. In a series of classification experiments, we compared hand-chosen prototypes with those selected by

our algorithm during the course of segmenting data from ten test MRI subjects. Our experiments summarized in Table 1 indicate that automatically selected prototypes are an effective substitute for expert hand-chosen prototypes. Table 2 illustrates such a comparison subject-by-subject and iteration-by-iteration for our algorithm. Registration alone yields prototypes which are a poor match for hand-chosen chosen prototypes (column  $\mathcal{A}^0$ ) with predictive values less than 0.7. The editing process consistently improves the result until the prototypes are similar to those that were hand-chosen, with predictive values greater than 0.9 (column  $\mathcal{A}^{\text{FINAL}}$ ).

In our algorithm, prototypes lead to density estimates which, in turn, lead to segmentations. Our segmentations compare favorably both to a previously published algorithm (Table 3) and hand-drawn segmentations (Table 4). Average Dice overlap is generally quite high, with the lowest value (mean  $\pm$  sd =  $0.72 \pm 0.06$ ) found in the myelin tissue class. Zijdenbos notes that Dice values above 0.7 are considered excellent agreement (Zijdenbos et al., 1994). Comparison with hand-drawn segmentations allowed us to test the entire imaging pipeline from acquisition to classification, but suffers from large inter- and intra-rater variability and bias.

Because our algorithm depends on an initial registration process, it may generate poor results if a subject has anatomy that differs so significantly from the template subjects that registration fails. If a consensus of template subjects can not be reasonably registered to the study subject, a large number of prototypes might be incorrectly labeled. The resulting poor density estimates and possibly poor atlas alignment would likely lead to a large number of misclassifications. This can be mitigated by inclusion of template subjects more closely matching the study subject and by improved registration. In practice, we have not encountered this when segmenting normal newborn anatomy. Future work may include testing on a variety of gestational ages beyond newborn and weighting of segmentations due to template scans by a direct measure of how well each template subject is brought into register with the study subject. Furthermore, we have not attempted in this work either cortical or sulcal reconstruction. Given the topology of the cortex, and contiguity of the underlying white matter, a technique such as (Grau et al., 2004) may be of benefit.

We have presented and validated a strategy for the segmentation of newborn brain MRI into several tissue types. This strategy uses a number of proven technologies, but introduces a novel method for training a supervised classifier. Unlike previous methods for improving the classifica-

tion feature space by clustering or pruning, our editing algorithm does not operate in the feature space itself, but rather learns the anatomy of the underlying subject. This algorithm also differs from previously presented newborn segmentation methods by differentiating cortical from sub-cortical gray matter and myelinated from unmyelinated white matter. Finally, because the prior knowledge used is stored in the form of labels on template subjects, the method does not rely on prior knowledge of the intensity characteristics of the underlying scans.

## 5. Acknowledgments

The authors would like to thank Dr. Andrea Mewes and Ms. Hoai-Huong Tran for drawing manual segmentations. This investigation was supported in part by a research grant from CIMIT, grants RG 3478A2/2 and RG 4032A1/1 from the NMSS, and by NIH grants R03 CA126466, R01 RR021885, R01 GM074068, R01 EB008015 and P30 HD018655.

## References

- Anbeek, P., Vincken, K. L., Groenendaal, F., Koeman, A., van Osch, M. J. P., van der Grond, J., 2008. Probabilistic brain tissue segmentation in neonatal magnetic resonance imaging. *Pediatr Res* 63 (2), 158–163.
- Barkovitch, A. J., 2005. Magnetic resonance techniques in the assessment of myelin and myelination. *Journal of Inherited Metabolic Disorders* 28, 311–343.
- Cocosco, C. A., Zijdenbos, A. P., Evans, A. C., 2003. A fully automatic and robust brain MRI tissue classification method. *Med Image Anal* 7 (4), 513–527.
- Cover, T. M., 1968. Estimation by the nearest neighbor rule. *IEEE Trans Information Theory* 14 (1), 50–55.
- Dale, A. M., Fischl, B., Sereno, M. I., 1999. Cortical surface-based analysis. I. segmentation and surface reconstruction. *Neuroimage* 9 (2), 179–194.
- Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1), 1–38.
- Dice, L. R., 1945. Measures of the amount of ecologic association between species. *Ecology* 26, 207–302.
- Elfadel, I. M., 1993. From random fields to networks. PhD thesis, MIT, Cambridge, MA, USA.
- Gerig, G., Kubler, O., Kikinis, R., Jolesz, F., 1992. Nonlinear anisotropic filtering of MRI data. *IEEE Trans Med Imag* 2, 221–232.
- Grau, V., Mewes, A. U. J., Raya, M. A., Kikinis, R., Warfield, S. K., 2004. Improved watershed transform for medical image segmentation using prior information. *IEEE Trans Medical Imaging* 23 (4), 447–458.

- Huppi, P. S., Warfield, S., Kikinis, R., Barnes, P. D., Zientara, G. P., Jolesz, F. A., Tsuji, M. K., Volpe, J. J., 1998. Quantitative magnetic resonance imaging of brain development in premature and mature newborns. *Ann Neurol* 43 (2), 224–235.
- Likar, B., Viergever, M. A., Pernus, F., 2001. Retrospective correction of MR intensity inhomogeneity by information minimization. *IEEE Trans Med Imaging* 20 (12), 1398–1410.
- Makris, N., Schlerf, J. E., Hodge, S. M., Haselgrove, C., Albaugh, M. D., Seidman, L. J., Rauch, S. L., Harris, G., Biederman, J., Caviness, V. S. J., Kennedy, D. N., Schmahmann, J. D., 2005. MRI-based surface-assisted parcellation of human cerebellar cortex: an anatomically specified method with estimate of reliability. *Neuroimage* 25 (4), 1146–1160.
- Mangin, J.-F., 2000. Entropy minimization for automatic correction of intensity non-uniformity. In: *Math Method Biomed Imag Anal*. IEEE Computer Society, Los Alamitos, California, pp. 162–169.
- Mathur, A. M., Neil, J. J., McKinstry, R. C., Inder, T. E., 2008. Transport, monitoring, and successful brain MR imaging in unsedated neonates. *Pediatr Radiol* 38 (3), 260–264.
- Mattes, D., Haynor, D. R., Vesselle, H., Lewellen, T., Eubank, W., 2001. Nonrigid multimodality image registration. In: *Medical Imaging 2001: Image Processing*. pp. 1609–1620.
- Mewes, A. U. J., Huppi, P. S., Als, H., Rybicki, F. J., Inder, T. E., McAnulty, G. B., Mulkern, R. V., Robertson, R. L., Rivkin, M. J., Warfield, S. K., 2006. Regional brain development in serial magnetic resonance imaging of low-risk preterm infants. *Pediatrics* 118 (1), 23–33.
- Mewes, A. U. J., Zollei, L., Huppi, P. S., Als, H., McAnulty, G. B., Inder, T. E., Wells, W. M., Warfield, S. K., Jul 2007. Displacement of brain regions in preterm infants with non-synostotic dolichocephaly investigated by mri. *Neuroimage* 36 (4), 1074–1085.
- Perona, P., Malik, J., 1990. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans Pat Analy Mach Intel* 12, 629–639.
- 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., 2003. Regional brain volumes and their later neurodevelopmental correlates in term and preterm infants. *Pediatrics* 111 (5 Pt 1), 939–948.
- Pohl, K. M., Bouix, S., Nakamura, M., Rohlfing, T., McCarley, R. W., Kikinis, R., Grimson, W. E. L., Shenton, M. E., Wells, W. M., 2007. A hierarchical algorithm for mr brain image parcellation. *IEEE Trans Med Imaging* 26 (9), 1201–1212.
- Prastawa, M., Gilmore, J. H., Lin, W., Gerig, G., 2005. Automatic segmentation of mr images of the developing newborn brain. *Med Image Anal* 9 (5), 457–466.
- Rosenfeld, A., Hummel, R., Zucker, S., June 1976. Scene labeling by relaxation operations. *IEEE Trans Systems, Man, and Cybernetics* 6 (6), 420–433.
- Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., Hawkes, D. J., 1999. Nonrigid registration using

- free-form deformations: application to breast MR images. *IEEE Trans Med Imaging* 18 (8), 712–721.
- Shah, D. K., Guinane, C., August, P., Austin, N. C., Woodward, L. J., Thompson, D. K., Warfield, S. K., Clemett, R., Inder, T. E., 2006. Reduced occipital regional volumes at term predict impaired visual function in early childhood in very low birth weight infants. *Invest Ophthalmol Vis Sci* 47 (8), 3366–3373.
- Sled, J. G., Zijdenbos, A. P., Evans, A. C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Trans Med Imaging* 17 (1), 87–97.
- Thompson, D. K., Warfield, S. K., Carlin, J. B., Pavlovic, M., Wang, H. X., Bear, M., Kean, M. J., Doyle, L. W., Egan, G. F., Inder, T. E., 2007. Perinatal risk factors altering regional brain structure in the preterm infant. *Brain* 130 (Pt 3), 667–677.
- Thompson, D. K., Wood, S. J., Doyle, L. W., Warfield, S. K., Lodygensky, G. A., Anderson, P. J., Egan, G. F., Inder, T. E., 2008. Neonate hippocampal volumes: prematurity, perinatal predictors, and 2-year outcome. *Ann Neurol* 63 (5), 642–651.
- Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Trans Med Imaging* 20 (8), 677–688.
- Vannier, M. W., Butterfield, R. L., Jordan, D., Murphy, W. A., Levitt, R. G., Gado, M., 1985. Multispectral analysis of magnetic resonance images. *Radiology* 154 (1), 221–224.
- Viola, P. A., 1995. Alignment by maximization of mutual information. PhD thesis, MIT, Cambridge, MA, USA.
- Warfield, S. K., Kaus, M., Jolesz, F. A., Kikinis, R., 2000. Adaptive, template moderated, spatially varying statistical classification. *Medical Image Analysis* 4 (1), 43–55.
- Warfield, S. K., Zou, K. H., M, W., 2004. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Trans. Med. Imag* 23, 903–921.
- Weisenfeld, N. I., Warfield, S. K., 2007. Simultaneous alignment and central tendency estimation. Paper read at workshop “Statistical Registration: Pair-wise and Group-wise Alignment and Atlas Formation” at MICCAI: Medical Image Computing and Computer Assisted Intervention 2007 in Brisbane, Australia.
- Wells, W. M., Grimson, W. L., Kikinis, R., Jolesz, F. A., 1996. Adaptive segmentation of MRI data. *IEEE Trans Med Imaging* 15 (4), 429–442.
- Woodward, L. J., Anderson, P. J., Austin, N. C., Howard, K., Inder, T. E., 2006. Neonatal MRI to predict neurodevelopmental outcomes in preterm infants. *N Engl J Med* 355 (7), 685–694.
- Woodward, L. J., Edgin, J. O., Thompson, D., Inder, T. E., 2005. Object working memory deficits predicted by early brain injury and development in the preterm infant. *Brain* 128 (Pt 11), 2578–2587.
- Xue, H., Srinivasan, L., Jiang, S., Rutherford, M., Edwards, A. D., Rueckert, D., Hajnal, J. V., 2007. Automatic segmentation and reconstruction of the cortex from neonatal MRI. *NeuroImage* 38 (3), 461–477.
- Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., Palmer, A. C., 1994. Morphometric analysis of white matter lesions in MR images: Method and validation. *IEEE Trans Med Imag* 13 (4), 716–724.

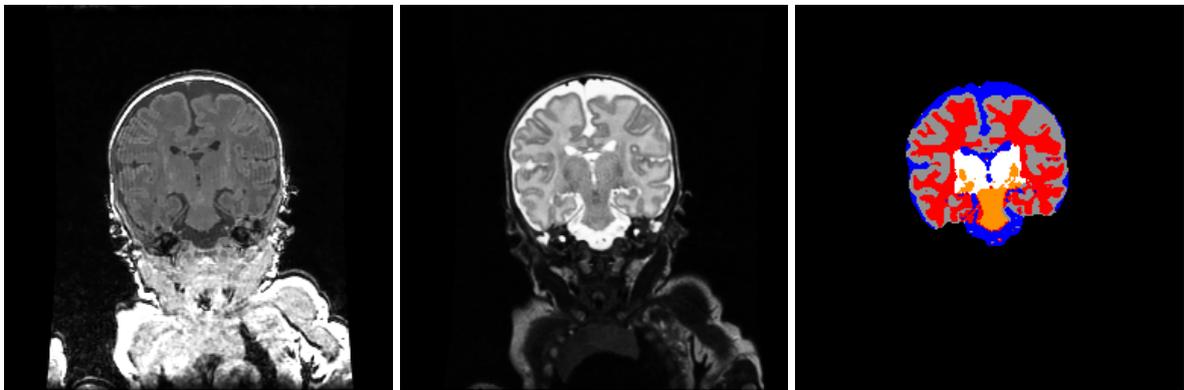


Figure 1: An example T1-weighted MRI from a term newborn (left) and the corresponding T2-weighted MRI (middle) and segmentation (right). The segmentation is shown with cortical gray matter in gray, subcortical gray matter in white, unmyelinated white matter in red, myelin in orange, and cerebrospinal fluid in blue.

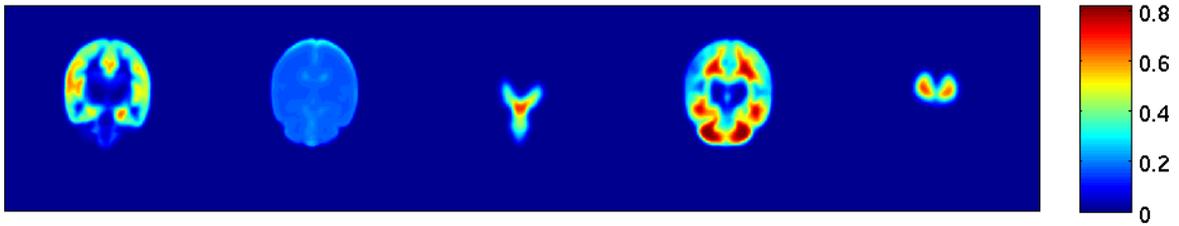


Figure 2: Atlas constructed from 15 newborn segmentations (see text). A representative mid-coronal slice was chosen for each channel. Shown from left to right is the prior probability of cortical gray matter, cerebrospinal fluid (CSF), myelin, unmyelinated white matter, and subcortical gray matter. The atlas was smoothed with a Gaussian filter ( $\sigma = 2.0\text{mm}$ ) and the prior probability of cerebrospinal fluid was boosted slightly everywhere within the brain in order to account for the variability of sulcal CSF. The color bar is calibrated to show prior probability between zero and one.

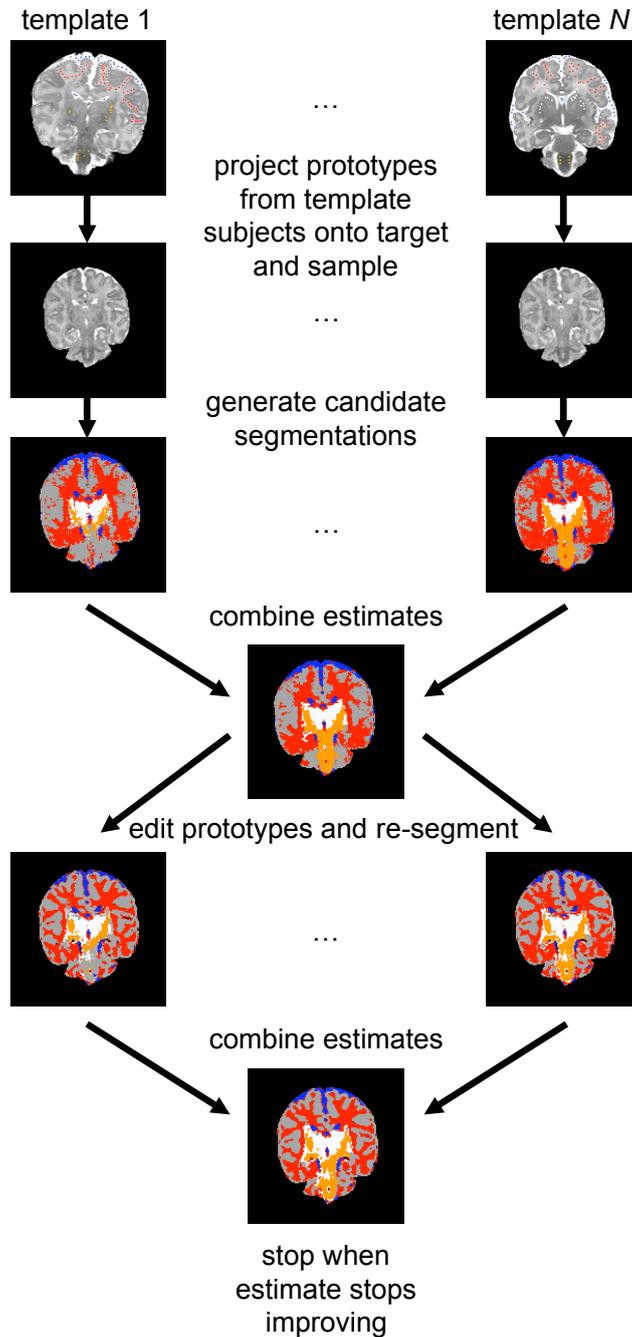
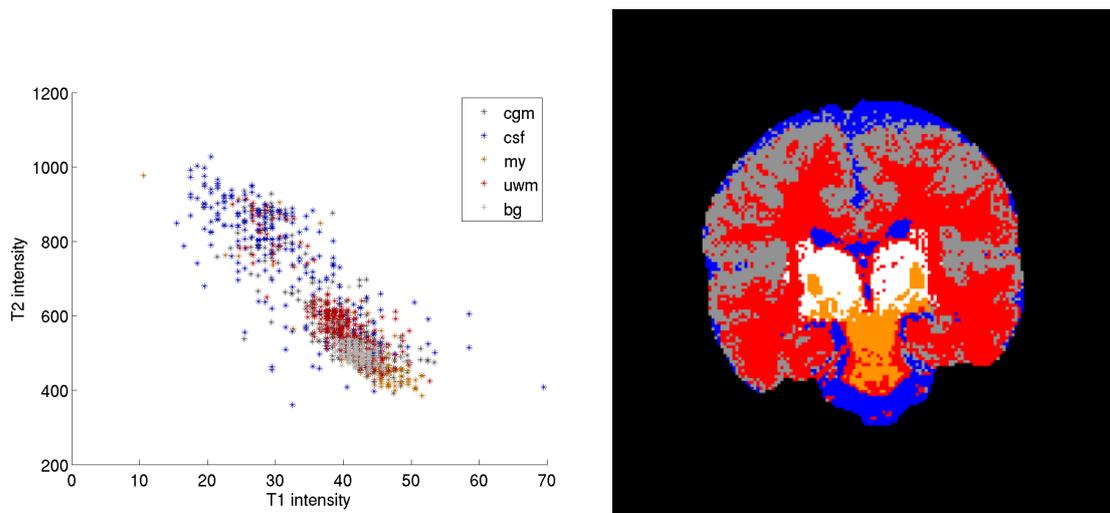
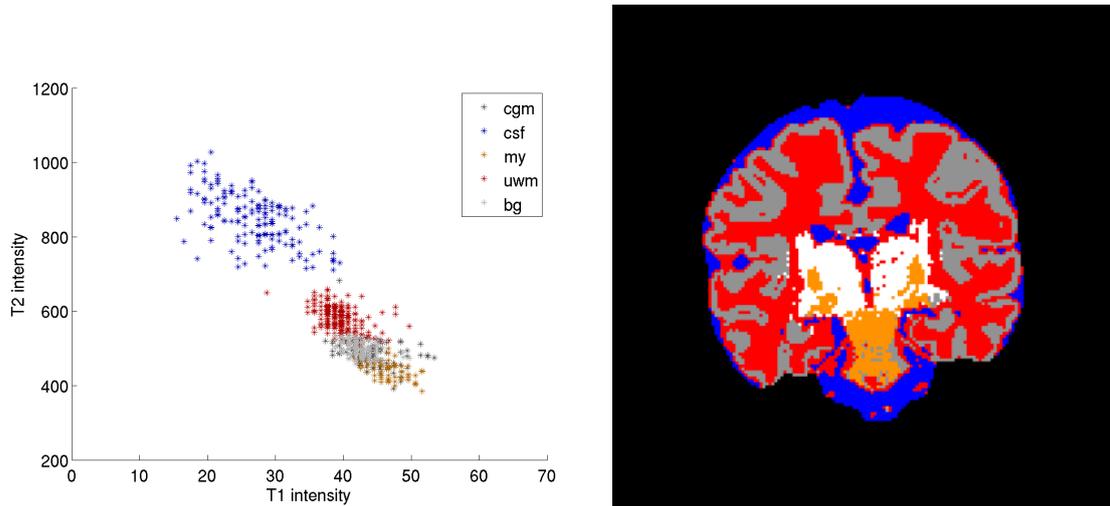


Figure 3: Schematic of the classification process. Prototypes from each template subject are projected onto the target subject under study. Image intensity values are sampled and each set of prototype points generates a different candidate segmentation. The segmentations are combined into an estimated true segmentation using Equation 3. The estimated true segmentation is then used to edit the prototypes and a new segmentation is generated. This process is iterated until convergence, when a final estimate of the segmentation is found.



(a) Prototype samples and the resulting segmentation after registration of a single template image.



(b) The same prototype samples after four iterations of the editing process and the improved segmentation of the subject from a single template image.

Figure 4: Demonstration of the impact of editing on the feature-space and the resulting segmentations, prior to fusion. Top left is a set of prototypes built by registration of a single template to the study subject. The tissue classes are shown as colors and there is a great deal of overlap between the classes making classification difficult. Any decision boundary drawn would have a large number of prototypes misclassified. This leads to poor density estimates and the resulting segmentation, top right, is noisy. After four iterations of editing, the prototype list has been edited to remove labels considered inconsistent with the underlying anatomy. The tissue classes are now largely separated. The resulting segmentation, on the right, is greatly improved and is one of a number of segmentations  $D_j$  that are fused to make the final result shown in Figure 1.

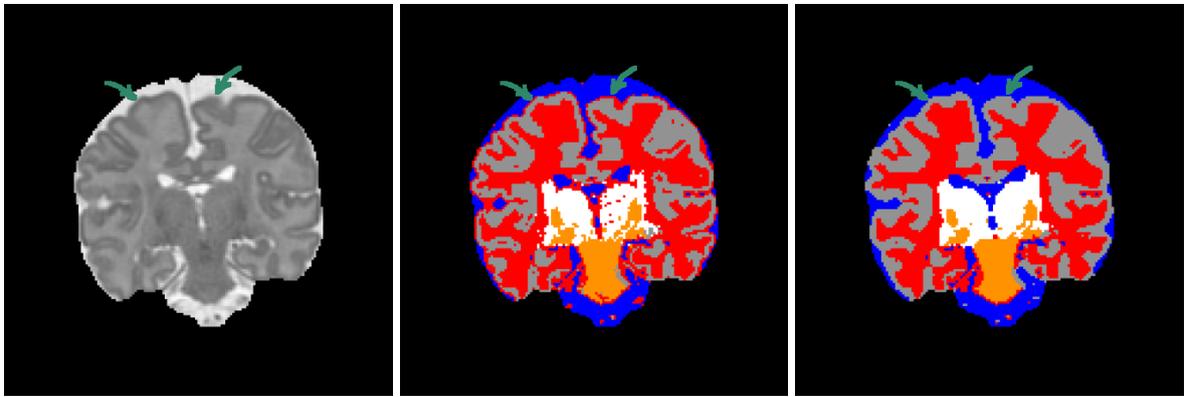


Figure 5: A T2-weighted MRI (left) and segmentations (center,right). CSF shown in blue, cortical gray matter in gray, unmyelinated white matter in red, subcortical gray matter in white, and myelin in orange. The arrow highlights the rim of misclassification (center) due to partial volume averaging between cortical gray matter and CSF, the corresponding anatomy in the T2 image (left), and the partial-volume corrected image (right). The corresponding T1-weighted image is not shown here.