Brain Tissue Classification of Magnetic Resonance Images Using Conditional Random Fields

Ameet Soni
Department of Computer Sciences
University of Wisconsin - Madison
soni@cs.wisc.edu

May 14, 2007

Abstract

In this project, I propose the application of a discriminative framework for segmentation of a T1-weighted magnetic resonance image(MRI). The use of Gaussian mixture models (GMM) is fairly ubiquitous in processing brain images for statistical analysis. This generative framework makes several assumptions that restricts its success and application. GMM assumes there is no spatial correlation when classifying tissue type, and also assumes each class of tissues is described by one Gaussian distribution. While the model is simple and quick, both assumptions are often violated leading to poor performance. Conditional Random Fields (CRF) avoid these assumptions by learning to discriminate between different tissue classes without assumptions on the class conditional density distribution. Additionally, CRFs allow arbitrary features of the data to be instantiated, allowing spatial correlations to be added in addition to other rich features other than voxel intensity. Results show that future work on the model is needed to enhance the model and improve the quality of training data

1 Introduction

Magnetic Resonance Imaging (MRI) is at the heart of medical imaging technology, providing high resolution three dimensional images of soft tissue. The large quantity of data provided by MRI for brain imaging in particular aids statisticians and medical professionals in disease diagnosis and functional understanding of the human brain. This large quantity of data, however, requires sophisticated processing algorithms to help analyze the images. Several of these processes remain as imperfect bottle necks in MRI analysis. One such bottleneck is tissue segmentation of the brain. Of particularly interest to this study is the identification of three main tissue types in the brain: grey matter(GM), white matter(WM) and cerebrospinal fluid(CSF). This task has implications on a wide array of exploratory studies and visualization tools of the brain.

Manual segmentation of the brain by experts is an extremely time-consuming process and arbitrary. In addition, inconsistencies between domain experts are "non-reproducible, and also are impractical for the large amounts of data required for meaningful statistical analysis" [9]. This motivates the need for computational methods to automatically segment the brain by classifying each voxel as being a member of one of the three main tissue types. The range of automated methods can be roughly grouped into two camps: surface based geometric methods and intensity based statistical methods. Geometric methods attempt to model the surface between tissue layers directly as opposed to classifying voxel by voxel. Methods in this category include deformable surface models[21], thin plate splines[6], and level set methods[19][20].

Intensity based methods center around the classification of individual voxels and include methods such as neural network classifiers[22][23], k-nearest neighbor classifier[9], and Gaussian mixture modeling (GMM)[7]. The Gaussian mixture model method is widely used and is implemented in the SPM software pack-

age[1]. The package uses the clustering algorithm in which the intensity values are fit to set of Gaussian distributions, each representing a class of tissue. The model suffers from an assumption of spatial independence of voxel intensities as well as a requirement to register the brain image to a spatial prior map. Spatial correlation was encoded by extending GMM to Hidden Markov Random Field (HMRF) model[5]. This model maintains the assumptions on the Gaussian class conditional density, but conditions the classification of a voxel on it's neighbors in the image. While an improvement over the GMM, HMRF still maintains a few shortcomings. First, HMRF model is particular sensitive to initial parameter estimation since noisy data can pose many local maxima in the parameter search space. Second, HMRFs assume a class conditional normal distribution for the intensity values. This assumption seems ill advised with problems such as partial volume effects and the bias field problem confounding the distribution of intensities. Third, HMRFs still assume the intensities are generated via a normal function dependent only on the label at a voxel. While tissue classes are spatially correlated, the data is not. This limitation is seen in the fact that HMMs (and HMRFs by extension) are unable to model overlaps and non-independent features of the classes and data. Partial volume effects and bias fields lead to such non-independence and overlaps. Lastly, HMRFs belong to a set of generative classifier algorithms that model the generation of the data. While this leads to intuitive models with neat parameter estimation functions under criteria such as maximum likelihood, they in general have higher asymptotic error than algorithms that directly learn the posterior distribution of the classes given the data(see [10] and [2] for a discussion on generative vs. discriminative algorithms).

This paper explores the use of Conditional Random Fields (CRF)[1][2][3] in brain tissue segmentation. CRFs discriminate the tissue types without making assumptions about the distribution of the data. In addition, by not modeling dependencies in the data, CRFs allow us to incorporate more information than just intensity values of a voxel in learning tissue classifications, such as correlating the data with neighboring voxel classification and intensities. Lastly, CRFs learn to discriminate classes using training data - previously labeled images. This replaces a spatial prior map and can accurately portray the high amount of variation among images. Prior work has show CRFs to outperform generative models in tasks related to natural language processing[1][2], bioinformatics[4], and vision[11][12][13]. The results of this project show that the formulation is incomplete, but that CRFs perform fairly well statistically relative to HMRFs but leave more to be desired upon visual inspection. Further work on developing feature space and normalizing the training data should show improvements in performance.

[1]URL: http://www.fil.ion.ucl.ac.uk/spm/

2 Related Work

2.1 Finite Gaussian Mixture Model

GMMs embody one of many similar parametric statistical approaches. They assign probability values on the class of an individual voxel, usually using only the intensity value of that voxel. All of these statistical methods use probabilistic inference criteria such as the maximum likelihood estimate (MLE) or maximum *a posteriori* (MAP) to classify. The difference lies in how the density function for the pixel intensities is formulated[14].

The Gaussian mixture model employed by SPM assumes that MR images consist of a finite number of different tissue types, each represented by a cluster from which each voxel in the image is generated. Each of these clusters is described by a normal distribution with its own mean and variance. The density function is thus formulated as a mixture of normal distributions, each with a mean and variance. Each individual Gaussian represents a cluster - a tissue type in this application[15]. Let us begin with a mathematical definition of the framework. We will define statistical models in a probabilistic framework. The task is to model the joint probability of the data and tissue class.

$$p(\mathbf{y}, \mathbf{x}) = \prod_{s \in I^3} p(y_s, x_s) \tag{1}$$

where s is an index over voxels on image of dimensions I by I by I (just to denote that we

are in 3D, the dimensions are not restricted to be the same) and $y_s \in \mathcal{L}$ where $\mathcal{L}$ is the set of possible labels $GM, WM, CSF, other$ of voxel s and x_s is the feature set of voxel s. As the notation indicates, the joint probability breaks piecewise into the number of voxels. In the GMM, we assume that each intensity value is generated by it's class type, so we condition on y_s

$$p(y_s, x_s) = p(x_s|y_s)p(y_s) \tag{2}$$

The first probability is the class conditional density function, which for GMMs is a Gaussian function. The second probability is a prior on each tissue class, or the mixing parameter. We calculate the marginal probability over x_s

$$p(x_s|\theta) = \sum_{\ell \in \mathcal{L}} p(Y_s = \ell) \cdot p(x_s|Y_s = \ell) \tag{3}$$

where the last term is the class conditional density function dependent on the parameters θ

$$p(x_s|Y_s = \ell) = f(x_s; \theta_\ell) \tag{4}$$

In this case, the function is a Gaussian, $\theta_\ell = (\mu_\ell, \sigma_\ell)$ and

$$f(x_s; \theta_\ell) = \frac{1}{\sqrt{2\pi\sigma_\ell^2}} \exp\left(\frac{(x_s - \mu_\ell)^2}{2\sigma_\ell^2}\right). \tag{5}$$

To classify, we simply take the maximum value for y_s

$$\hat{y_s} = \arg\max_{\ell \in \mathcal{L}} p(Y_s = \ell|x_s) \tag{6}$$

where

$$p(Y_s = \ell|x_s) = \frac{p(x_s|Y_s = \ell, \theta_\ell)p(Y_s = \ell)}{\sum_{\ell' \in \mathcal{L}} p(x_s|Y_s = \ell', \theta_{\ell'})p(Y_s = \ell')} \tag{7}$$

This algorithm is considered to be *unsupervised* - it does not have previously labeled examples to learn the classes. Therefore, we cannot directly employ MLE or MAP to determine the best parameters for our mode. Ashburner and Friston[7] describe their use of the Expectation-Maximizationem algorithm. EM is an iterative algorithm which alternates between estimating the class variables y_s and determining the model parameters θ by maximizing the likelihood of the model. EM is guaranteed to converge to a local maximum for the parameters, but will not necessarily find the correct solution. The algorithms simplicity lends to its popularity.

GMM, however, suffers from not utilizing all of the information available. It essentially throws all of the data into a histogram, and thus loses any information in the local neighborhood that could identify noise versus signal. To distribute the clusters close to the tissue class, the image needs to be registered to a prior probability map of what tissue type a voxel belongs to based on its location. This prior map helps introduce spatial correlation. But it also assumes the brain image is normal and will fit the probability map. If there is any deviation, tissue classification will be incorrect[5][9].

2.2 Gaussian Hidden Markov Random Field

To address the problem of ignoring spatial information, a Hidden Markov Random Field[5] model was applied to the segmentation task. An HMRF is a stochastic MRF generated process whose state sequence is unobserved. It is more general than a Hidden Markov Model, whose state process is a Markov chain. The FSL[2] package has an implementation of the HMRF which outperforms the GMM. The intuition behind introducing spatial correlations is that neighboring voxels are expected to have the same classification and similar intensities. GMMs assume the second assumption holds and thus implies that neighbors will get the same label. In noisy images, however, neighboring voxels may have the same label, but GMMs won't capture this since they do not have similar intensities. HMRFs model the distribution of labels for a voxel conditional upon its neighboring set in addition to the data. This additional set of parameters place the voxels features in context of its surroundings.

The HMRF model differentiates from GMM in eq. 2. We extend to constrain parameters on neighbors

$$p(y_s, x_s) = p(x_s|y_s)p(y_s|y_{\pi(s)}) \tag{8}$$

[2]URL: http://www.fmrib.ox.ac.uk/fsl/

where $\pi(s)$ is the set of neighbors of pixel s. The marginal of x_s from eq. 3 becomes

$$p(x_s|y_{\pi(s)},\theta) = \sum_{\ell\in\mathcal{L}} p(Y_s = \ell|Y_{\pi(s)})\cdot f(x_s;\theta_\ell) \quad (9)$$

where $Y_{\pi(s)}$ is a configuration over the neighbors of s. Notice the only change is in the first factor, where we now have a different mixing parameter for each configuration of neighbors for each label. The density for the generation of data $f(x_s;\theta_\ell)$ is the same as for the GMM. This model is the Gaussian hidden Markov random field(GHMRF) that is implemented for FSL. It should be clear that a degenerate form of the GHMRF where $\pi(s) = \emptyset$ yields the GMM.

For classification, we extend eqs 6 and 7 to incorporate the constraints on y_s

$$\hat{y_s} = \arg\max_{\ell\in\mathcal{L}} p(Y_s = \ell|x_s, Y_{\pi(s)}) \quad (10)$$

where

$$p(Y_s = \ell|x_s, Y_{\pi(s)}) =$$

$$\frac{p(x_s|Y_s = \ell,\theta_\ell)p(Y_s = \ell|Y_{\pi(s)})}{\sum_{\ell'\in\mathcal{L}} p(x_s|Y_s = \ell',\theta_{\ell'})p(Y_s = \ell'|Y_{\pi(s)})} \quad (11)$$

Zhang et al.[5] extend the EM framework to learn parameters for the model similar to before, where we add in parameters for conditional probability of the tissue class given the neighbors. They prove that the HMRF model is much less sensitive to noise than the GMM. Empirical studies show that incorporating spatial correlations does create smoother results[8].

One difficulty with the algorithm lies in how much to weight the neighboring pixels and which pixels to include. Too high of a weight and images are overly smooth, too low and they became equivalent to a model lacking spatial constraints. Another problem with the HMRF model is the initialization of parameters. In noisy data, the EM search will hit many local maxima that can easily be fitting the noise as opposed to the signal. This means the initialization of the cluster parameters is a critical step. The use of a prior probability map, as discussed above, creates many issues when dealing with non-standard populations or deviations in brain structure. The HMRF algorithm uses a discriminant threshold technique[17] to maximize interclass variance and minimize intraclass variance. The problem is that the classes are unknown, and thus the prior probabilities become equivalent to the simple histogram based techniques HMRF is trying to improve upon. This is demonstrated in results obtained from other studies by the high variability in results from HMRF[18]. Another limitation of the model is that it assumes there is one normal distribution for each label. In addition to evidence that some tissue class intensity values do not behave normally but Rician[7], partial volume effects (the presence of multiple tissue types in a voxel) resulting from discrete sampling of the image blur the line between tissue types and cannot be explained with one Gaussian distribution. In addition, the lack of normalization across an image means that the distribution of white matter in one slice maybe vastly different than in a slice further away. Many models attempt to solve this with bias field correction, but methods are not perfect[5]. In light of these concerns, a Gaussian class conditional density assumption may not be optimal.

3 Model

3.1 Conditional Random Fields

An extensive review of CRFs is available in[2]. CRFs are a conditional distribution over the classification given the input data. The goal of the segmentation task is to determine the posterior probability of the tissue classes in the image. This is given in eq. 11, where we see that the model constraints are built into a distribution on the observed data. This is a generative framework since we attempt to model functions of the class labels that most likely generated the data $p(x_s|y_s)$. Rather than learning the parameters for the data density function, discriminative algorithms learn the posterior directly since that is what is really required.

CRFs are an example of this discriminative framework. Conveniently, every generative algorithm has a discriminative pair. The simplest CRF is a linear-chain model - a pair to the Hidden Markov Model, which holds a first order Markov assumption in which the label $\mathbf{y}_t$ is de-

pendent on the label at $\mathbf{y}_{t-1}$.

To generalize for medical imaging, we will extend the model to an arbitrary random field by converting the HMRF model. We will refer to two dimensions, although the notation and model is generalizable to a three dimensional model.

To convert an HMRF to a CRF, begin with eq. 8. Let us convert each of these parameters into a binary feature represented as $f_k(y, y', x)$. Each feature function has an associated weight λ_k. We will constrain features to be have three inputs: a scalar y_s, a vector $y_{\pi(s)}$, and a vector $\mathbf{x}$. Note that each input is an instantiation over that variable, and that the last feature represents an arbitrary set of intensity values not just x_s. The first term of eq. 8 is the probability of the data given the class. This can be represented as

$$f_{\ell i}(y_s, y_{\pi(s)}, x_s) = \delta(y_s, \ell)\delta(x_s, i) \tag{12}$$

where δ is the delta function, returning 1 if the parameters are equal and 0 otherwise. For a given voxel, the only feature function of this set that will be true is when the intensity value is $i \in \Omega$ where Ω is the range of values for intensities. $\lambda_{\ell i}$ gives of the weight of label ℓ with intensity i. Note that there will $|\mathcal{L}| \times |\otimes|$ number of such feature functions. The second half of eq. 8 is the distribution of class labels give the neighborhood of class labels. This can be broken down into

$$f_{\ell\ell'}(y_s, y_{\pi(s)}, x_s) = \delta(y_s, \ell)\delta(y_{\pi(s)}, \ell') \tag{13}$$

each with an associated $\lambda_{\ell\ell'}$. By plugging in these factors and modifying to traditional mathematical notation we obtain the feature function defined HMRF model equivalent to eq. 8

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z_{joint}} \exp\left\{\sum_{s\in I^3}\sum_{s'\in\pi(s)}\sum_{k=1}^{K}\lambda_k f_k(y_s, \mathbf{y}_{s'}, \mathbf{x}_s)\right\} \tag{14}$$

where Z_{joint} is a normalization constant. To transform our model into a discriminative framework, we condition on $\mathbf{x}$ by dividing by the marginal of x.

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp\left\{\sum_{s\in I^3}\sum_{s'\in\pi(s)}\sum_{k=1}^{K}\lambda_k f_k(y_s, \mathbf{y}_{s'}, \mathbf{x}_s)\right\}}{\sum_{y'}\exp\left\{\sum_{s\in I^3}\sum_{s'\in\pi(s)}\sum_{k=1}^{K}\lambda_k f_k(y'_s, \mathbf{y}'_{s'}, \mathbf{x}_s)\right\}} \tag{15}$$

which reduces to our CRF model

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\exp\left\{\sum_{s\in I^3}\sum_{s'\in\pi(s)}\sum_{k=1}^{K}\lambda_k f_k(y_s, \mathbf{y}_{s'}, \mathbf{x}_s)\right\} \tag{16}$$

There are two key differences in eqs 8 and 16. The first is that we are no longer modeling the joint probability of the data and class. The probability of observing the data is irrelevant since we want to produces a model of the likelihood of class combinations. The second is that the feature functions can be defined in an arbitrary sense. HMRFs are constrained two features - the class conditional density and the label correlation to its neighbors. CRFs allow us to introduce new features such as parameters tied to multiple intensity values or intensity functions dependent on neighboring values and labels. This allows complex dependencies to be modeled.

3.2 Features

The key asset to a discriminative framework is that modeling additional dependencies is not intractable as is the case with most generative algorithms. For this project, only a small subset of features was introduced, but many more are suggested in the future work that could vastly improve performance of the model. The following sets of features were defined on the dataset:

- Correlation of label to intensity at position s (from GMM)
- Correlation of label to neighbor labels (from HMRF). For this project, neighbors were defined on a 2D plan as the voxel to the left (i-1), right (i+1), up(j+1) and down(j-1). Diagonally connected labels were removed for simplicity. In addition, the large size of the graph and presence of a large amount of "non-matter" lead to the omitting of features including positions with no intensity value.

- Correlation of label neighbor intensity values. Neighbors defined as above.

In addition, the following features were planned for testing, but could not be included due to tractability and project time constraints:

- Correlation of label to gradient values.
- Correlation of labels across modalities
- Correlation of labels to slices above and below the current plane.
- Other features to help account for unnormalized data and biases in the intensity distribution.

3.3 Outputs

The output labels of the CRF model were one of four possibilities, as based on the training set. The labels for y_s are chosen from:

- Other - the default state of the system, describes voxels that do consist of brain matter or are unidentified in the image.
- GM - Gray Matter tissue.
- WM - White Matter tissue.
- CSF - Cerebral Spinal Fluid.

3.4 Learning and Inference

While the addition of a large set of features can aid in learning patterns of brain tissue distribution, they also increase the computational complexity of the model. In particular, learning the parameters λ becomes difficult. Sutton and McCallum (2006)[2] describe many approximate learning methods that reduce complexity. L-BFGS is a limited memory version of a quasi-Newton approximation method that maximizes the parameters according to an approximation of the actual problem. In practice, such a learning procedure is tractable and quite effective, and was used in this project. Inference is done via a standard Viterbi algorithm whereby the most likely path is chosen through the model by choosing the maximum score at each voxel step.

3.5 Implementation Issues

MALLET[3] provided the implementation of the CRF graphical model as well as the necessary learning and inference procedures required of this project. Specifically, templates for abstract graphical models is available in the GRaphical Models in MALLET (GRMM) distribution. The codes was extended to handle two dimensional data (although tractability become an issue as the optimization algorithms are designed with sequential data in mind). In addition, to simplify learning, intensity values were discretized into one of 20 bins equally dividing the intensity value range.

4 Experimental Methodology

The 20 normal MR brain data sets and their manual segmentations were provided by the Center for Morphometric Analysis at Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/. This is the same dataset used in validation of many segmentation techniques including thin plate spline[6] and HMRF[5], although the authors of the HMRF paper do not provided quantitative analysis since they claim there are flaws in the data. From the data description: "The coronal three-dimensional T1-weighted spoiled gradient echo MRI scans were performed on two different imaging systems. Ten FLASH scans on four males and six females were performed on a 1.5 tesla Siemens Magnetom MR System (Iselin, NJ) with the following parameters: TR = 40 msec, TE = 8 msec, flip angle = 50 degrees, field of view = 30 cm, slice thickness = contiguous 3.1 mm, matrix = 256x256, and averages = 1. Ten 3D-CAPRY scans on six males and four females were performed on a 1.5 tesla General Electric Signa MR System (Milwaukee, WI), with the following parameters: TR = 50 msec, TE = 9 msec, flip angle = 50 degrees, field of view = 24 cm, slice thickness = contiguous 3.0mm, matrix = 256x256, and averages = 1." Voxel intensities were squeezed into the range from 0 to 255 and then binned for this experiment. The skull was removed from each image. Also provided with the IBSR dataset was a manually seg-

[3]URL: http://mallet.cs.umass.edu/index.php/Main_Page

mented map, used for training. The positions of the brain were normalized, but otherwise left untransformed.

5-fold cross validation was performed to analyze the quality of the CRF model. The images were divided into groups of 4 and then each group was tested on a model built with the other 4 groups providing the training data. Since the FSL algorithm is unsupervised, each image was processed on its own. The middle horizontal slice of the brain was analyzed for both algorithms. Each image was 256 along one dimension and between 50-65 in the other. FSL's Fast algorithm was run with default parameters(3 tissue types). Images 17 and 18 were removed due to clear defects in the scans. To compare the methods, I used two statistical measure of similarity to measure the overlap between the algorithm output and the "gold standard" segmentation. The first measure is the kappa index which ranges from 0 to 1 with 1 being complete agreement

$$\kappa(S1, S2) = 2 * \frac{|S_1 \cap S_2|}{|S_1| + |S_2|} \tag{17}$$

where S_1 and S_2 are the set of voxels with a certain property. In this experiment, the property will be GM or WM. $|S|$ it size of set S. The second index is the Jaccard index which ranges from 0 to 1, with 1 being complete agreement.

$$J(S1, S2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{18}$$

To obtain the discrete classification of both algorithms, we use the Viterbi algorithm for CRFs and the binary maps produced by FSL for HMRFs.

Provided with the dataset was an unpublished evaluation of common algorithms. The performance can be seen in the Jaccard index values for GM 1 and WM 2. Of importance is the performance of the experts; these low values indicate that the gold standard segmentation should be taken with a grain of salt since there was a large amount of variation between experts.

5 Results & Discussion

The results from the experiments are shown in Table 1 and 2. As can be seen in the data, CRFs outperform HMRFs across the board in identifying gray matter in the images. The reverse is true for white matter. Table 1 highlights in bold the higher performing algorithm, and CRFs outperform in 19 out of 20 for GM while HMRFs outperform in the same ratio in WM. It should be noted that in comparison to figures 1 and 2, both algorithms outperform those tested by the IBSR. Similar results are obtained when looking at the Kappa values in Table 2.

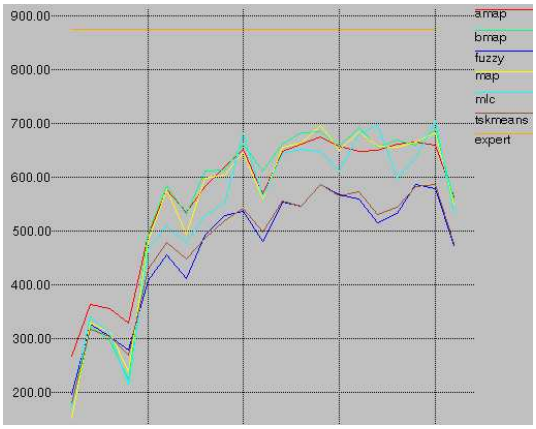

Figure 1: Jaccard Index plot ranked from hardest to easiest subject for multiple segmentation algorithms, Gray Matter

There are several different explanations for the results of the experiment. First, in general WM is harder to identify based on empirical results. All algorithms achieve lower Jaccard Indexes in identifying the white matter. This could be due to the fact that it is present in lower quantities and thus any errors are magnified, whereas in gray matter, a misclassification is not as severe. Second, since the data contains a greater proportion of GM points (about 3 to 1), the CRF model is overfitting the data and thus learning to err on the side of GM when in doubt. This can be solved by adding a penalty cost to the Viterbi algorithm, something that has worked in past projects to get a better understanding of the tradeoff between two classes. Rather than a hard-max, we classify a voxel as GM only if its probability is greater than the probability of WM by a certain constant that can be tuned. Last, this could be a case of oversmoothing, where the

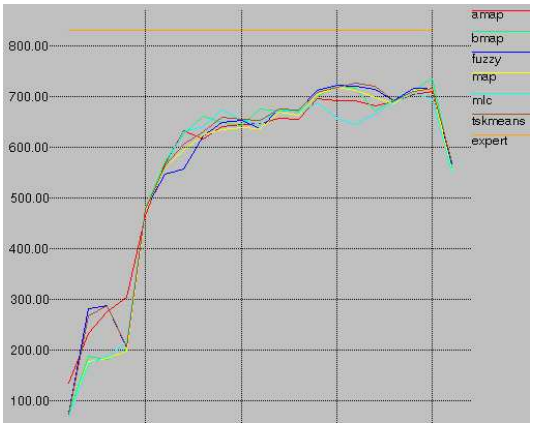

Figure 2: Jaccard Index plot ranked from hardest to easiest subject for multiple segmentation algorithms, White Matter

Sub. No	GM		WM	
	CRF	HMRF	CRF	HMRF
1	0.712	**0.756**	0.392	**0.622**
2	**0.700**	0.646	0.447	**0.489**
3	**0.670**	0.538	0.522	**0.548**
4	**0.688**	0.625	0.453	**0.532**
5	**0.811**	0.653	0.450	**0.542**
6	**0.701**	0.635	0.576	**0.619**
7	**0.699**	0.611	0.621	**0.623**
8	**0.621**	0.480	0.418	**0.537**
9	**0.684**	0.588	0.309	**0.513**
10	**0.745**	0.516	0.389	**0.423**
11	**0.736**	0.693	**0.451**	0.431
12	**0.602**	0.577	0.532	**0.543**
13	**0.776**	0.754	0.508	**0.571**
14	**0.746**	0.713	0.480	**0.541**
15	**0.651**	0.562	0.311	**0.461**
16	**0.675**	0.530	0.444	**0.499**
19	**0.633**	0.569	0.357	**0.436**
20	**0.540**	0.515	0.502	**0.592**
Mean	**0.688**	0.562	0.453	**0.500**

Table 1: Jaccard Index for each subject, Gray Matter(GM) and White Matter(WM)

constraints introduced force the labels towards a state of lower entropy. In this case, classification would tend towards GM since they form larger cluster broken apart by tracks of white matter in the images used. This seems to be the most likely answer since the number of retrieved instances of WM is roughly equal the actual number of WM voxels.

Figures 3 and 4 show that oversmoothing is a problem in both HMRFs and CRFs. In inspection of visual images, the CRF protocol seems to fail to produce quality segmentations. In many cases, the first few iterations match the data well, but then successive iterations tend to become globular and actually achieve worse qualitative performance despite an improvement in quantitative statistics. CSF isn't even identified in most images. This indicates that the CRF in the current framework is inadequate to handle image segmentation. There are several reasons to explain poor performance.

The first set of problems arises from the dataset. As noted in other papers [5], the quality of data provided by the IBSR is fairly poor. 2 images were already noted for clear defects in the scan images. The intensity values were nowhere near uniform throughout the images - a pitfall that CRFs would experience since they are trying to generalize across images as opposed to HMRFs which attempt only to generate the current images. Second, the segmentation provided is assumed to be ground truth. In actuality, the agreement between segmenters (compare the 6 amongst each other) yields a Jaccard Index of 0.876 and 0.832 for GM and WM respectively. Thus, training on these imperfect classifications is risky.

The second set of problems comes from the CRF algorithm itself. The features used were discretized binary features, thus destroying any relative information between intensities. Many extensions exist to make CRF conducive to numeric features, but further work on modeling these distributions needs to be done. Most CRF algorithms have been validated on text datasets and thus are rarely posed with ordered features. Second, the rich feature set capability CRFs that are often cited as the main attraction for the model were ill-exploited in this project. A better understanding of common vision techniques and texture extraction would aid in developing a better set of descriptors for images. Many successful vision techniques using CRFs use meta information such as multiscale features to aid in connecting multiple regions of an images[13].

Sub. No	GM		WM	
	CRF	HMRF	CRF	HMRF
1	0.832	0.861	0.563	0.767
2	0.823	0.785	0.618	0.657
3	0.803	0.700	0.686	0.708
4	0.815	0.770	0.623	0.694
5	0.895	0.790	0.620	0.703
6	0.824	0.777	0.731	0.765
7	0.823	0.759	0.766	0.768
8	0.766	0.649	0.589	0.699
9	0.812	0.741	0.472	0.678
10	0.854	0.681	0.560	0.595
11	0.848	0.819	0.622	0.602
12	0.752	0.732	0.694	0.704
13	0.874	0.860	0.674	0.727
14	0.855	0.833	0.649	0.702
15	0.789	0.720	0.475	0.631
16	0.806	0.693	0.615	0.666
19	0.775	0.725	0.526	0.607
20	0.701	0.680	0.668	0.743
Mean	0.814	0.754	0.620	0.690

Table 2: Kappa Index for each subject, Gray Matter(GM) and White Matter(WM)

Probably the biggest contributor to deficient performance is the size of images. The graph for one images is extremely large, while the feature space is manageable. This is in contrast to NLP tasks where the graph is relatively small (on the order of words or sentences) but the feature space is large (huge dictionary). This makes inference a difficult task and favors oversmoothing of the data in iterative maximum likelihood steps. When a graph is highly connected, inventive algorithms need to be put in place to prevent overfitting. There exists literature on loopy belief propagation techniques for images that could aid in learning parameters and inference[24]. In addition, sampling techniques such as particle filtering[25] could be used to provide better inference and classification than Viterbi which is a dynamic programming approach. In addition, partitioning the image space into small overlapping regions using a hierarchical graphical modeling approach would allow local inference to be exhaustive, but limit the information transgressing the rest of the graph to be infrequent and manageable.

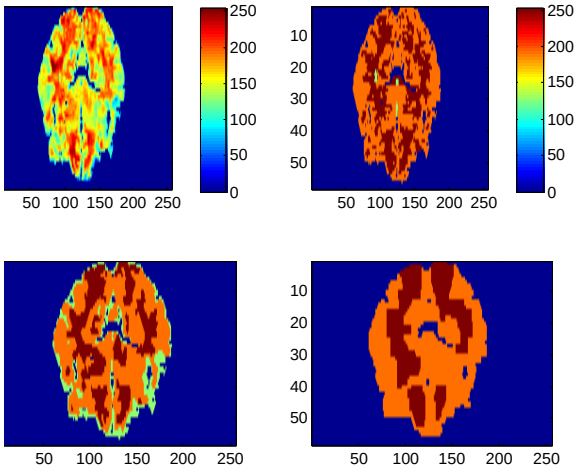

Figure 3: Segmentation image of Subject 2. From top-left clockwise: original skull removed intensities, gold segmentation with WM being dark red(254) and GM orange(192), CRF segmentation, Fast segmentation

6 Conclusions and Future Work

As the results indicate, there is room for improvement. While the features used help develop information, other features and structural changes may help model the problem better. Many of these improvements are discussed above and include gradient orientation and magnitude features as well as features to help handle bias fields and partial volume effect. Also, work needs to be done to speed up training and inference. In the experiments run, it took 24 hours to run just 2 or 3 iterations of one fold on a modern high speed computer. Many of the suggestions for inference (sampling and loopy belief propagation) and reduction of scale(hierarchical models, partitioning of images) would definitely improve run time.

Processing of the images needs to improved to normalize features across subjects and to also account for variability in voxel thickness. A neighbor along a dimension with 1.00mm thickness is closer than a neighbor on a dimension of 3.00mm thickness and thus should have more weight in the clique of a voxels neighbors. In addition, new training sets thare are more stable and modern should be utilized. The BrianWeb simulator[4] provides images simulated by an ad-

[4]URL: http://www.bic.mni.mcgill.ca/brainweb/

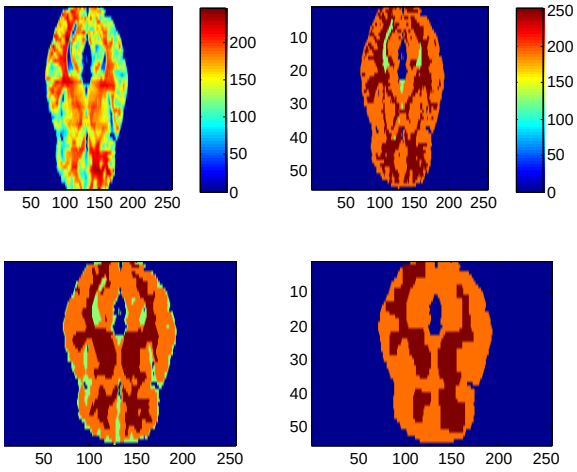

Figure 4: Segmentation image of Subject 3. From top-left clockwise: original skull removed intensities, gold segmentation with WM being dark red(254) and GM orange(192), CRF segmentation, Fast segmentation

vanced algorithm that should have been the basis test for this model before looking at real data. The worry is that simulated data isn't ideal for a trained classifier, but it still should provide a foothold for further exploration into improving the model. In addition, the IBSR recently added 18 new brain images whose quality is better. Testing was not completed in time for the project, but the results will be available shortly. The resolution is much higher thus making training even longer for these images.

Conditional Random Fields provide a discriminative framework for which we can segment brain tissue. Compared to similar methods such as GMM and HMRF, they provide several improvements. They do not assume any density function for the data and thus are free from the errors of assuming a Gaussian distribution when factors such as partial volume effect and bias fields affect the data. Second, CRFs can use training data to infer probability estimates and thus do not have to rely on templates and heuristics for prior probabilities that can error prone with unnormalized images. CRFs also do not need to be registered like some GMMs do. Lastly, since CRFs do not need to model dependencies in the data like generative algorithms do, we are free to add additional features to our dataset to uncover more modalities than just one Gaussian per tissue class. Future testing should be carried out on segmentation of objects in brain tissue, since all prior success with CRFs has come with globular objects that the model tends to prefer.

7 Acknowledgments

The 20 normal MR brain data sets and their manual segmentations were provided by the Center for Morphometric Analysis at Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/. I would also like to thank members of the course who helped me with handling and transforming images - something that unfortunately took up most of the project time.

References

[1] J. Lafferty, A. McCallum, F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001).*

[2] C. Sutton, A. McCallum. An Introduction to Conditional Random Fields for Relational Learning. *Introduction to Statistical Relational Learning.* Edited by Lise Getoor and Ben Taskar. MIT Press. 2006.

[3] H. Wallach. 2004. Conditional Random Fields: An Introduction. In: *University of Pennsylvania CIS Technical Report MS-CIS-04-21.*

[4] B. Settles. 2005. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191-3192.

[5] Y. Zhang, M. Brady, S. Smith. 2001. Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45-57.

[6] X. Xie, M.K Chung, G. Wahba. 2006. Magnetic resonance image segmentation with thin plate spline thresholding. TR 1105. Department of Statistics, University of Wisconsin-Madison.

[7] J. Ashburner, K.J. Friston. Image Segmentation. *Human Brain Function.* 2nd edition. Academic Press. 2003

[8] D.L. Pham, C. Xu, J.L. Prince. A Survey of Current Methods in Medical Image Segmentation. 2000. *Annual Review of Biomedical Engineering* 2:315-337.

[9] C.A. Cocosco, A. P. Zijdenbox, A.C. Evans. 2003. A Fully Automatic and Robust Brain MRI Tissue Classification Method. *Medical Image Analysis*. 7(4):513-527.

[10] A.Y. Ng, M.I. Jordan. 2002. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes. *In Advances in Neural Information Processing Systems(NIPS), 14*.

[11] K. Murphy, A. Torralba, W.T.F. Freeman. 2004. Using the forest to see the trees: a graphical model relating features, objects and scenes. *In Advances in Neural Information Processing Systems(NIPS), 16*.

[12] S. Kumar, M. Hebert. 2004. Discriminative Fields for Modeling Spatial Dependencies in Natural Images. *In Advances in Neural Information Processing Systems(NIPS), 16*

[13] X. He, R. Zemel, M.. Carreira-Perpin. Multiscale conditional random fields for image labeling. 2004. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) 2004*.

[14] C.M. Bishop, *Neural Networks for Pattern Recognition.* Oxford, U.K.:Oxford Univ., 1995.

[15] W.M. Wells, E.L. Grimson, R. Kikinis, F.A. Jolesz. Adaptive segmentation of MRI data. 1996. *IEEE Transactions in Medical Imaging*. 15:429-442.

[16] A.P. Dempster, N. Laird, D.B. Bubin. 1977. Maximum likelihood for incomplete data via EM algorithm. *J.Roy. Stat.Soc*, ser. B, 39(1):1-38.

[17] N. Otsu. 1979. A threshold selection method from gray-level histogram. *IEEE Trans. Syste. Man Cybern*. 9:62-66.

[18] E. Angelini, T. Song, B. Mensh, A. Laine. 2007. Brain MRI Segmentation with Multiphase Minimal Partitioning: A Comparative Study. *International Journal of Biomedical Imaging*. (2007), Article ID 10526.

[19] S. Osher, N. Paragios (Eds.). 2003. Geometric level set methods in imaging, vision and graphics. Springer Verlag, New York.

[20] J.A. Sethian. 1996. Level Set Methods and Fast Marching Methods. Cambridge Univ. Press, Cambridge, UK.

[21] J.D. MacDonald, N. Kabani, D. Avis, A.C. Evans. 2000. Automated 3-D Extraction of Inner and Outer Surfaces of Cerebral Cortex from MRI. *NeuroImage* 12:340356.

[22] E. Gelenbe, Y. Feng, K.R.R. Krishnan. 1996. Neural network methods for volumetric magnetic resonance imaging of the human brain. *Proc. IEEE*. 84:1488-1496.

[23] L.O.Hall et al. 1992. A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. *IEEE T.Neural Networks*. 3:672-682.

[24] P.F. Felzenszwalb, D.P. Huttenlocher. 2006. Efficient Belief Propagation for Early Vision. *International Journal of Computer Vision*. 70(1).

[25] S. Arulampalam, S. Maskell, N. J. Gordon, T. Clapp. 2002. A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking. *IEEE Transactions of Signal Processing*. 50(2):174-188.