
Multiple Sclerosis Lesion Segmentation in MRI Images

Gautham Vasan

Department of Computing Science
University of Alberta
Edmonton, T6E2H2
vasan@ualberta.ca

Jared Rewerts

Department of Computer Engineering
University of Alberta
Edmonton, T6E2H2
rewerts@ualberta.ca

Megha Panda

Department of Computing Science
University of Alberta
Edmonton, T6E2H2
meghaanu@ualberta.ca
email

Parisa Mohebbi

Department of Computing Science
University of Alberta
Edmonton, T6E2H2
mohebbi@ualberta.ca

Reza Sobhannejad

Department of Computing Science
University of Alberta
Edmonton, T6E2H2
sobhanne@ualberta.ca

Sina Jalali

Department of Computing Science
University of Alberta
Edmonton, T6E2H2
jalalil@ualberta.ca

Abstract

Medical Imaging is an emerging field that can help reduce workload for medical professionals and increase outcomes for afflicted individuals. Our focus is on comparing classification techniques for Multiple Sclerosis lesions in order to find the most successful techniques. We also attempt to combine some techniques to improve detection success. Our results confirm the standard that Random Forest typically provides the best results, but that there are other options useful in segmenting lesions. Interestingly, Random Forest also tends to classify more quickly than other algorithms, particularly when compared to SVM. This makes it the most practical in terms of both accuracy and speed.

1 Introduction

Multiple Sclerosis is a debilitating disease that affects the brain and spinal cord. Technically, MS is an autoimmune disease; the immune system attacks parts of the body, in this case, the protective covering of nerve fibers. If occurrences are bad enough, the nerves can build up scar tissue, which impedes nerve conduction. This scar tissue is what we refer to as lesions.

One of the most shocking aspects of Multiple Sclerosis is that it primarily affects individuals between the ages of 15 and 40. Women are also twice as likely to be affected by MS. Early detection is key to managing symptoms and controlling disease outcomes. Affected individuals can use exercise, cool temperatures and a balanced diet to help treat symptoms.

To detect Multiple Sclerosis lesions, scans are made of patient brains. Medical professionals analyze these scans, looking for signs of MS. These manual segmentation techniques have several drawbacks, not least among them the time input required by the medical professionals. Manual

segmentation also has a high rate of error; only 70% of lesions are properly detected. Automatic segmentation techniques hope to alleviate these issues.

One of the main challenges of automatic segmentation is the low lesion volume in patient brain matter. Typically, only a small percentage of the patients brain has lesions, which can make detecting them difficult. This also poses a problem when gathering positive examples of the lesion. The positive to negative ratio is much lower than in many other segmentation problem spaces. Further, since most patients have very little lesion tissue, over classification can be a problem with automated segmentation. If a healthy brain is scanned and processed, ideally, we would show no lesions. In practice, this can be very difficult.

1.1 Related Work

Segmentation of Lesions in MRI images is an active area of research. The 3 main objectives are:

1. Extracting features that can differentiate healthy tissues from scar tissues,
2. Selecting the most relevant features that help achieve the task, and
3. Improving the performance of the classification.

1.1.1 Review of Segmentation Methods

In MS lesion segmentation, the typical features are the intensity of each voxel in different modality images. Additionally, some methods like [1], combined K-NN based on intensities with a template-driven segmentation method to reduce false positives. Others like the method presented in [6] used the probability of a voxel belonging to tissue class with help of an atlas. Intensity of the six neighbor voxels can be added to feature vector of a point which has been used with an ANN in [5]. In [3], they used up to 255 features derived from applying different filters on images, and then with the help of principal component analysis (PCA), data is transformed to new orthogonal coordinates so the first column covers the greatest variance between data; hence, a simple thresholding on the first component of PCA can classify lesions. In [2], random decision forest (RDF) is used with local (like intensity) and context-rich (will be discussed later) features. RDF has the advantage of automatically selecting the best features over other methods.

1.1.2 Exploring Combinations [PLEASE GIVE YOUR COMMENTS ON THIS PART]

Lots of methods have been employed for lesion segmentation as mentioned before but no valid comparison is done to measure performance of different classifiers using different features. In this project we aim to compare various algorithm's efficiencies on the same data and feature sets to explore their properties in lesion segmentation. We use Support Vector Machine (SVM), Neural Networks, K-Nearest Neighbors, Random Forest, Markov Random Field, and Logistic Regression as classifiers. For features, we use Haar-Like, image filters, LM filters, entropy, gaussian based, and atlas. In a broad sense, we tried to extract all possible features to see how the well-known classifiers operate based on them.

2 Feature Detection

Image analysis aims at reducing information to a subset that is relevant to the task in hand. Information reduction often happens gradually with information being reduced until the desired result is extracted from the data. [4] The first level of reduction computes local features that are assumed to pertain to objects of interest. We use 4 different classes of features here in order to segment the MRI images. The features we use for the given task were selected based on the literature survey for MS Lesion segmentation described in the Related Work section. The general characteristics and reasons for why the particular class of features work is described below.

2.1 Context (Haar-Like Features)

In the detection phase of the Viola-Jones object detection framework [?], a window of the target size is moved over the input image, and for each subsection of the image the Haar-like feature is

calculated. This difference is then compared to a learned threshold that separates non-objects from objects. Because such a Haar-like feature is only a weak learner or classifier (its detection quality is slightly better than random guessing) a large number of Haar-like features are necessary to describe an object with sufficient accuracy. Context-rich features exhibit high variability. This variability combined with their ability to highlight regions which differ from their neighborhood explains why they were chosen. Together with local features (voxel intensities), context-rich features learn a multichannel appearance model conditioned by tissue spatial priors.[2]

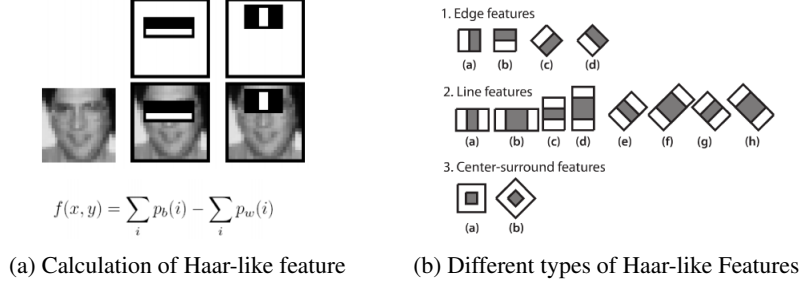


Figure 1: Haar-Like Features

2.2 Image Filters

Filtering is a technique used for modifying or enhancing an image. It can be used to emphasize certain features or remove other features in an image. In image processing filters are mainly used to suppress either the high frequencies in the image, i.e. smoothing the image, or the low frequencies, i.e. enhancing or detecting edges in the image. We extract the features through Linear filtering of the MRI image using convolution.

2.2.1 Leung-Malik (LM) Filter Bank

We use the filter bank provided by the Visual Geometry Group at the University of Oxford [] to obtain a set of features. The LM set is a multi scale, multi orientation filter bank with 48 filters. It consists of first and second derivatives of Gaussians at 6 orientations and 3 scales making a total of 36; 8 Laplacian of Gaussian (LOG) filters; and 4 Gaussians. The filters occur at the basic scales $\sigma = \sqrt{2}, 2, 2\sqrt{2}, 4$. The Gaussians occur at the four basic scales while the 8 LOG filters occur at σ and 3σ . The algorithm proposed by Leung and Malik [?] demonstrated excellent results for recognizing 3D textures from a single image under any lighting and viewing geometry. Since different MRI scanners produce different quality of images, the LM filters should be able to recognize different textures, i.e., lesions in the brain.

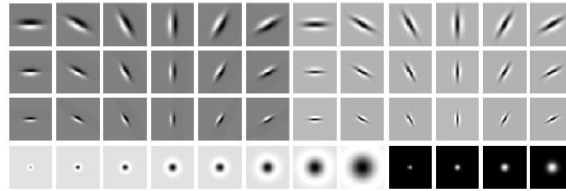


Figure 2: The LM filter bank has a mix of edge, bar and spot filters at multiple scales and orientations. It has a total of 48 filters

2.3 Entropy & Gaussian based Features)

2.4 Atlas Features

All the features need to explained briefly with good images.

3 Classifiers

Reasons for using each classification method. Brief description about them. No need for images here I guess.

3.1 Support Vector Machines

SVMs try to split the positive and negative data using a hyper-plane which applies very well on complex tasks when implemented with kernels. The model is both powerful and unlikely to overfit, but it is sensitive to ratio of positive and negative data (pos/neg) as it weights them equally. In our case, we need a model to learn lots of negative data, healthy brain patches, and a few positive data; hence, we should sample training data not unbalanced than ratio of 2 to 1.

3.2 Neural Networks

The most important property of neural networks is the ability to identify a good estimation even for complex functions. Basically in Neural Networks we have a few layers. The first layer is our features (input) and the last layers is related to the labels (output). We also have a few hidden layers between these two. Considering this structure we can make almost all functions, even the most complex ones. That's the reason for choosing this method. In this project we tried Neural Networks with 2 and three hidden layers.

3.3 k-Nearest Neighbours

Nearest Neighbours approach decides on the label of a certain data-point by counting votes of its k closest points. The model is completely dependent on the distance measure used to find close points. It is an easy to implement model and does not require training. To make a decision on each voxel we take the majority of votes from nearest points in the dataset.

3.4 Random Forests

The model is based on taking votes of a set of decision trees on each voxel. One of the best properties of RFs is that they are less sensitive of imbalanced ratio of positive and negative data compared to other models. Also, it provides the best features which split the classes better and scores the features based on how helpful they are in classification. Furthermore, it is hard for RFs to overfit while they fit complex models as a result of using different trees to make the overall decision. We learn the random forest and create the trees by distributing points in their leafs and predict by merging votes of each tree to label the data.

3.5 Markov Random Fields

The basic principle of Markov Random Field is to treat the input image as a graph in which each voxel is a node and all the neighboring voxels are interconnected. In the case of foreground segmentation, two spatial nodes of foreground and background are added to the graph and edges between voxels. Each pair of nodes are assigned a weight equal to the probability that a given voxel belongs to either foreground or background. In graph theory, a cut is a partition of the vertices of a graph into two disjoint subsets. An st-cut is a graph cut comprising all directed edges from a node in a vertex set A to a node in its complement, \bar{A} , such that $s \in A$, $t \in \bar{A}$, and A is connected. The cost of a cut is simply defined as the sum of edge weights for the edges removed by the cut. The aim of this algorithm is to use a graph-cut to divide graph into two classes subject to the Min-Cut problem, i.e., to find the st-cut with minimal cost. The advantage of this method is that it considers the probabilities of its neighbours as a feature.

In this project, we trained a random forest (Claimed to be the best algorithm for lesion segmentation according to literature survey [?]) and used it to obtain initial probabilities of each voxel classified as a lesion. Subsequently, MRF is applied on the graph constructed using the weights, i.e., probabilities obtained from the classifier.

3.6 Logistic Regression

The model fits a line to separate the classes of data. It is easily implemented and if it has sufficient features, is capable of doing well in our problem space. We learn the hyper-plane through our train data and classify test data based on the direction of the hyper-plane they fall into. A particular area of interest is how MRF improves logistic regression when used in sequence.

4 Experiment Design

Pipeline Diagram is required here

4.1 Validation Measures

For each pixel of the brain, our classifiers must label it as either lesion or healthy brain matter. Depending on the data set used, our ground truth source may be different. The MICCAI data set is provided by doctors and has also been cross referenced several times. The hope is that this increases the accuracy of the results. The MICCAI scan represent real brains. Brain Web on the other hand uses a simulated brain scan and places lesions in the simulated scans.

When comparing our output to the ground truth, we come up with 4 possibilities for each region of the brain scan:

1. True Positive
The ground truth and our detection agrees the region is a lesion.
2. True Negative
The ground truth and our detection agrees the region is healthy.
3. False Positive
Our detection finds a lesion in a brain region, but the ground truth disagrees.
4. False Negative
The ground truth finds healthy brain, but our detection disagrees.

Taken together, these values can be used to analyze the performance of different classifiers.

4.1.1 Dice Score

Dice is the primary metric we used to analyze algorithm performance.

$$\frac{2 * TruePositive}{FalsePositive + FalseNegative + 2 * TruePositive}$$

One of the major benefits of Dice is that it heavily weights true positives. This makes it a very strict measure for a problem space like ours, where the amount of positive tissue is very low.

As an example, consider scanning a brain that has a single tiny lesion. To get a perfect dice score, you must classify that lesion and only that lesion. Failing to classify that lesion will result in a score of 0. If you successfully classify the lesion, but also falsely classify surrounding tissue, the dice score will decrease rapidly. Since the healthy tissue heavily outweighs the lesions, it is easy to falsely classify a small amount of this healthy tissue as lesion.

There is one caveat to keep in mind when using Dice however. If we test against a healthy brain scan, Dice cannot be used to analyze our results. This is because there is no possibility of getting a true positive and so the result will always be 0.

4.1.2 Accuracy

Accuracy is good measure of how much brain matter we properly classify. Unlike Dice, accuracy weights true negatives evenly with true positives.

$$\frac{TruePositive + TrueNegative}{FalsePositive + FalseNegative + TruePositive + TrueNegative}$$

For our problem, accuracy usually isn't very useful. The reason for this is that true negatives are too common to be considered a reliable metric. For most brain scans, you could just classify the brain as healthy and the resulting accuracy would be good. We still use it though, as it can be useful when scanning healthy brains, which is where Dice fails.

4.1.3 Sensitivity

Sensitivity is similar to Dice in that it heavily weights true positives.

$$\frac{TruePositive}{FalseNegative + TruePositive}$$

Sensitivity affords no penalty to misclassifying healthy brain as lesion. This means a brain scan that is output as entirely lesion will do well on it. Despite this drawback, it can still be useful in situations that Dice is in.

4.1.4 Detections

As a simple evaluation measure, we count the number of properly classified lesions. A lesion is considered classified if we find even a single pixel of lesion within it. This makes it unreliable in terms of overall performance, but it is a valuable metric for basic testing of classifiers.

4.2 Training and Test Data

The quantitative analysis of the MRI images is done using data provided by two sources:

1. Brain Web - MS Lesion Brain Database, McGill University
2. MICCAI Challenge 2008 Dataset

There exists no common ground truth or gold standard in most medical analysis tasks which results in an increasing need for validation of techniques. The Simulated Brain Database (SBD) acts as a stopgap solution to the validation problem. SBD is based on two anatomical models: (i) Normal brain and (ii) MS Lesion. For both the models, full 3-dimensional data volumes have been simulated using three sequences (T1-, T2-, and proton-density- (PD-) weighted) and a variety of slice thicknesses, noise levels, and levels of intensity non-uniformity. Unlike the vivo acquired data, here the ground truth is known which helps evaluating the various image analysis methods.

The MICCAI challenge dataset is actual MRI image scans provided in 3 sequences - T1, T2 and Flair. The actual MRI image scans are obtained from Boston Children's Hospital (CHB) and University of North Carolina, Chapel Hill (UNC). The labels for the data are provided by two or more experts for each brain which results in a rough estimate for the presence of a lesion. We have 10 labeled brain volumes each from CHB and UNC as training data. Additionally, we have 18 unlabeled volumes from CHB and 10 unlabeled volumes from UNC as test data. The difficulty in using the MICCAI challenge data is that:

1. Less than 0.1% of the brain volume is lesions
2. Different MRI Scanning devices result in different quality of images
3. No uniformly accepted Ground Truth or Gold Standard for validation

Since the datasets are heavily biased, we sample the data from one volume on the brain and test on an entirely different volume of the brain. The sampling is done usually in a ratio 1:2 to 1:7 between positive and negative samples of data, where positive samples are considered to be lesions. For testing the classifier, the features are extracted for every pixel and tested on the entire volume of a brain.

5 Results

Lots of Images! Table with comparative results. Explanation for why we get these results.

Classifier	Dice Score	Accuracy	Sensitivity	Detection
SVM	0	0	0	0
Neural Networks	0	0	0	0
k-Nearest Neighbours	0	0	0	0
Random Forest	0	0	0	0
MRF with RF	0	0	0	0
Logistic Regression	0	0	0	0

5.1 State of the art results

Brief description of the benchmark

5.2 Discussion

Detailed analysis of the results we get from different results

6 Conclusion

7 Acknowledgements

8 References

References

- [1] Olivier Commowick and Simon K Warfield. A continuous staple for scalar, vector, and tensor images: an application to dti analysis. *Medical Imaging, IEEE Transactions on*, 28(6):838–846, 2009.
- [2] Ezequiel Geremia, Olivier Clatz, Bjoern H Menze, Ender Konukoglu, Antonio Criminisi, and Nicholas Ayache. Spatial decision forests for ms lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378–390, 2011.
- [3] Dirk-Jan Kroon, van ESB Oort, and CH Slump. Multiple sclerosis detection in multispectral magnetic resonance images with principal components analysis. 2008.
- [4] Klaus D Toennies. *Guide to medical image analysis: methods and algorithms*. Springer Science & Business Media, 2012.
- [5] Akmal Younis, Ahmed T Soliman, Mansur R Kabuka, Nigel M John, et al. Ms lesions detection in mri using grouping artificial immune networks. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pages 1139–1146. IEEE, 2007.
- [6] Alex P Zijdenbos, Reza Forghani, and Alan C Evans. Automatic” pipeline” analysis of 3-d mri data for clinical trials: application to multiple sclerosis. *Medical Imaging, IEEE Transactions on*, 21(10):1280–1291, 2002.