Discriminative Random Field Modeling of Lung Tumors in CT Scans

Brian Liu

December 17, 2010

1 Abstract

The ability to conduct high-quality automatic 3D segmentation of tumors in CT scans is of high value to busy radiologists. Discriminative random fields (DRFs) were used to segment 3D volumes of lung tumors in CT scan data. Optimal parameters for the DRF inference were first calculated using gradient ascent. These parameters were then used to solve the inference problem using the graph cuts algorithm. Results of the segmentation were varied, with DRFs performing better on isolated tumors, but exhibiting bleed-through to adjacent tissues with similar intensities. Improvements can be made in the selection of features, discriminative models, and parameter optimization algorithm.

2 Introduction

Markov random fields have been used in the area of computer vision for various tasks, such as image segmentation. The problem is often formulated as a classification problem: Given a finite set of labels corresponding to objects, label each pixel in the image with the object it belongs to. There are many variants of this basic problem, and their difficulties vary greatly between them. For example, using segmentation to correct for noise in an OCR context tends to be easier than identifying objects in a real world scene.

One of the more popular interpretations of Markov random fields has been as an energy minimization problem. We use the pixel grid as a graph, in which each pixel is a vertex and neighboring pixels share an edge between them. We can then define an energy cost for any given labelling as a function of various features of the MRF. In the traditional MRF definition, the energy potential can be expressed as an association potential function of each node and an interaction potential function of pairs of neighbors. The goal is then to find an optimal labelling which minimizes the total energy.

MRFs have been applied to a wide variety of vision problems. While 2D images have generally been the most popular segmentation problem in computer vision, the framework can be easily generalized to three dimensions, where we can consider a neighborhood of 6 voxels: up, down, left, right, front, back. This has applications in medical imaging, where it is often of interest to segment out volumes such as organs or tumors [9]. Most commercial software for segmentation tasks today use some form of human-assisted region growing algorithm.

In some aspects, MRFs possess an advantage in the segmentation problem. Theoretically, with a correct problem formulation (i.e. a convext potential function), MRFs give a good estimate backed

by strong math. Solving the inference problem afterwards can be done quickly and optimally (for binary labels, for multiple labels, within an approximation factor) using a variety of optimization methods, such as graph cuts [2] [5] [1] [3]. There is a caveat, however: while MRFs give a good solution to the problem, it does not guarantee anything about how well the problem formulation itself works. Often, picking the right potential functions can be a matter of trial and error.

There are several variants of MRFs out in the literature. In particular, conditional random fields (CRFs) generalize the MRF formulation by allowing data to factor into the traditional MRF interaction potential formulation, with a discriminative model instead of a generative model. Kumar and Herbert's discriminative random fields (DRFs) [6] extends the usual work of conditional random fields to multiple dimensions. In particular, Kumar and Herbert's construction allows for the use of a variety of discriminative models, like SVMs [7].

Our goal in this paper is to apply DRF methodology to the segmentation of lung tumors in CT scans. A previous study by Lee et al. have attempted tumor segmentation in the brain with MRI scans [7]. CT lung tumor segmentation has some important differences. Lee et al.'s study used specific information about each patient's brain's spacial location to define better features. We would like to avoid using patient specific information when possible. In addition, the MRI scans used information from a variety of different modalities to aid in feature selection. CT data is limited to one modality, and in our particular case we do not even have access to pre-contrast agent injection scans. CT scans do have the advantage that a reported intensity for a given voxel corresponds to an actual physical unit, so relative intensities are not an issue.

One of the advantages of using DRFs is the ability to use any discriminative model for learning. Lee et al., for example, used an SVM classifier as the discriminative model for their brain tumor segmentation. Unfortunately, a preliminary study using a variety of features did not get good results using SVM classification.

3 Methods

The data set consisted of 6 abdominal CT scans of a patient in a time series study. There were around seven tumors detected in each scan, for a total of 41 different tumors. The total subvolume for all 41 tumors consisted of about 930000 negative voxels and 3300 positive voxels. All data was provided by Dr. Krishna Juluru at Weill Cornell Medical College. Ground truth segmentations were labelled by hand.

We will use supervised learning to learn a discriminative random field model of a lung tumor. After we have learned the parameters, we can solve the inference problem using graph cuts.

We construct a DRF model of the CT volume as follows:

Let $G = (S, E)$ be the graph that represents the 3D volume, where each node in S represents a voxel and an edge in E connects adjacent voxels in a 6-neighborhood. Let n_i be the observed intensity at voxel $s_i \in S$, p_i be the 3-vector of the relative coordinates of voxel s_i in the volume, and let label $x_i \in \{-1, 1\}$ be the label associated with s_i. We define an observation $y_i = (n_i, p_i)$. The random variables x_i obey the Markov property that $Pr(x_i|y, x_{Si}) = Pr(x_i|y, x_{N_i})$, where N_i is the set of neighbors of s_i and S
i is everything in S except s_i.

Following Kumar and Herbert's notation, we use the Hammersley-Clifford theorem and assume only pairwise clique potentials to be nonzero, and thus write:

$$Pr(x|y) = \frac{1}{Z} exp(\sum_{i \in S} A_i(x_i, y) + \sum_{i \in S} \sum_{j \in N_i} I_{ij}(x_i, x_j, y))$$

Where Z is the partition function, A_i is an association potential and I_{ij} is an interaction potential.

3.1 Association potential

We model the association potential discriminatively using a logistic model, since the labels are binary. We will define a feature vector f_i at site s_i as a function of the observations y. We first define some constants calculated from an outside source. The distribution of lung tumor voxel intensities was modeled as a Gaussian, with constants μ_{int} and σ_{int} calculated from the training data. The location of the lung tumor voxels was also modeled as a Gaussian deviating from a prior known location, with constants $l = (l_x, l_y, l_z)$ and $\sigma_{loc} = (\sigma_x, \sigma_y, \sigma_z)$. These constants are pre-computed prior to the DRF model construction.

We then define our feature vector to be:

$$f_i(y) = [\frac{(n_i - \mu)^2}{\sigma^2}, \frac{(p_i - l)^2}{\sigma_{loc}^2}]$$

Our two features capture the distance between the voxel intensity and the average voxel intensity of a lung tumor and the distance between its spacial location and a previously known location. The aforementioned constants specify some prior knowledge about the amount tumors can move and what intensities the tumors should be.

We then have the option of transforming our feature vector via some non-linear transformation to $h_i(y) = [1, \phi_1(f_i(y)), ...\phi_2(f_i(y))]^T$, which is a kernal mapping of our original feature vector with the introduction of a bias element. As an initial test, we simply used a linear kernel - that is, $\phi(f_i(y)) = f_i(y)$.

Our probability that we're trying to maximize is then:

$$Pr(x_i = 1|y) = \frac{1}{1 + e^{-w^T h_i(y)}}$$

Since $Pr(x_i = -1|y) = 1 - Pr(x_i = 1|y)$, we can express this probability more compactly as:

$$Pr(x_i|y) = \frac{1}{1 + e^{-x_i w^T h_i(y)}}$$

Finally, we model the association potential as the log of this probability:

$$A(x_i, y) = log(\frac{1}{1 + e^{-x_i w^T h_i(y)}})$$

The parameter to learn in the association potential is then w.

3.2 Interaction potential

We model the interaction potential using the pairwise smoothing of the Ising model, modulated by the difference in intensities of the two sites. We will define a new feature vector $\delta_{ij}(y)$ that captures this difference:

$$\delta_{ij}(y) = [1, |n_i - n_j|/1000]^T$$

Where the first element is there to accomodate a bias parameter. We then define the interaction potential informally as a modification of the Ising model potential used in a typical Markov random field, using a simplified form following that in Kumar and Herbert:

$$I(x_i, x_j, y) = \beta(x_i x_j v^T \delta_{ij}(y))$$

The β term is a constant term controlling the degree to which the smoothing cost affects the potential. The parameter to optimize, then, is v.

3.3 Learning

The learning problem can then be formulated as finding the maximum likelihood of $Pr(x|y; w, v)$. To evaluate this, however, we would need to evaluate the partition function Z, which is NP-hard because we would need to sum over $2^{|S|}$ number of possible labellings. Instead, for simplicity we maximize the pseudo-likelihood: $Pr(x|y; w, v) \approx \prod_{i \in S} Pr(x_i|x_{N_i}, y; w, v)$. That is:

$$Pr(x|y; w, v) \approx \prod_{i \in S} \frac{1}{z_i} exp(A(x_i, y) + \sum_{j \in N_i} I(x_i, x_j, y))$$
$$z_i = \sum_{x_i \in \{-1,1\}} exp(A(x_i, y) + \sum_{j \in N_i} I(x_i, x_j, y))$$

Defining $\theta = (w, v)$, with M training examples, the θ that maximizes the log of this pseudolikelihood is:

$$\hat{\theta} = \arg\max_{\theta} \sum_{m=1}^{M} \sum_{i \in S} (A(x_i, y) + \sum_{j \in N_i} I(x_i, x_j, y) - log(z_i)) \tag{1}$$

We use gradient ascent to calculate a θ that maximizes this expression. The maxima is global since the function is convex.

3.4 Inference

An exact maximum a posteriori solution can be obtained for the pairwise Ising model by a graph cuts algorithm. We cut off the interaction potentials to a minimum of 0, since graph cuts cannot deal with negative interaction potentials. Graph cuts was performed using Olga Veksler's gco-3.0 library in C++ with a Matlab wrapper [2] [3].

4 Results

4.1 Gaussian model of lung tumor voxel intensities

We chose to model the intensity values of positive voxel examples as a Gaussian. As one can see from Figure 1, this is an acceptable model as an approximation, though it suffers from certain

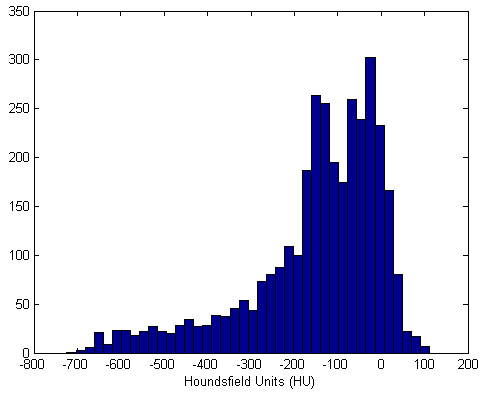


Figure 1: Histogram of positive tumor intensities. There may be two separate peaks close together, but a Gaussian fits well enough as an approximation. There also appears to be a heavy tail, which is likely due to perceptual differences in visual labelling of ground truth.

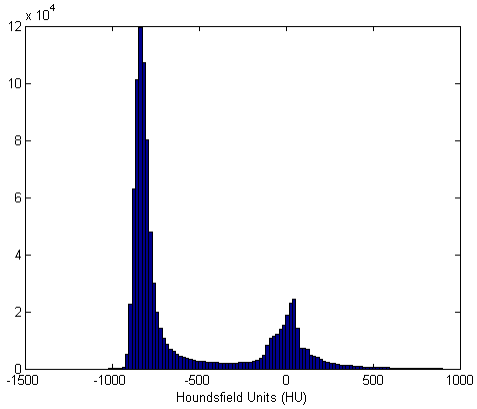
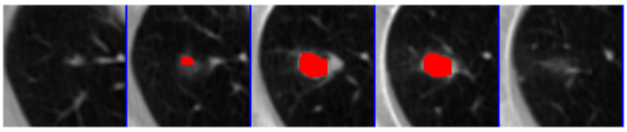


Figure 2: Histogram of negative tumor intensities. There is a large peak at around -1000 due to the fact that the lungs are mostly empty air. There is however a significant peak at around the same intensities as tumor intensities, which makes feature selection more difficult.

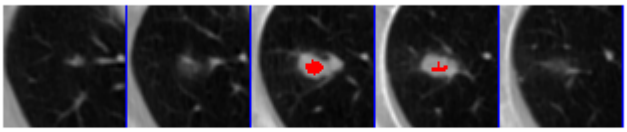
characteristics such as a heavy tail which skews the mean and variance. Compared to the histogram of negative values seen in Figure 2, this is negligible. Of particular note is the fact that negative examples also exhibit a peak at around the same intensities. In fact, due to the sheer number of negative training examples compared to positive ones, the negative peak overwhelms the positive peak. This is not surprising, as other tissue besides tumors can very well have similar intensities. Nevertheless, this is an issue we must overcome.

4.2 Gradient Ascent

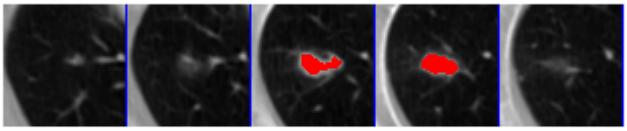
We used gradient ascent to calculate the optimal parameters w and v. However, because we have more than 930000 negative voxel examples and only 3300 positive examples, the convex function to optimize exhibited a sharp plateau that made gradient ascent difficult. As a result, good initial parameter values were required for gradient ascent to work properly.



(a) Inferred segmentation.



(b) Custom graph cuts segmentation.



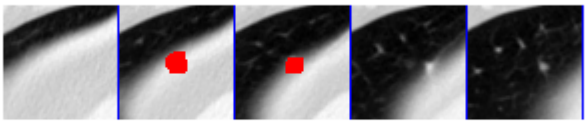
(c) Ground truth segmentation.

Figure 3: Comparison of an inferred segmentation versus a custom graph cuts energy function and the ground truth labelling. Inferred segmentation gets fairly close.

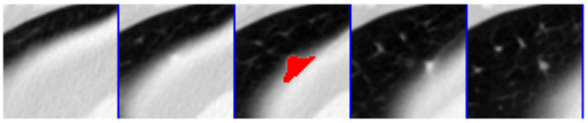
4.3 Segmentation

Parameter learning was done through 4-fold cross validation on a total of 40 data sets. Segmentation was done using graph cuts. An approximation to the potential function had to be made in order to keep costs positive. The average precision was 0.48 and the average recall was 0.82. This performs better than a custom graph cuts energy function in recall (0.64) and worse in precision (0.77). A slice of an example segmentation, a custom graph cuts segmentation, and the ground truth can be seen in Figure 3.

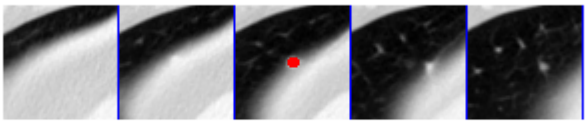
For tumors attached to adjacent tissue of similar intensities, the segmentation exhibited some bleed through to the surrounding tissue. An example can be seen in Figure 4. This is because the features selected were insufficient to differentiate the adjacent tissue areas from the tumor.



(a) DRF segmentation.



(b) Custom graph cuts segmentation.



(c) Ground truth segmentation.

Figure 4: Example of segmentation bleeding through to adjacent tissue. The custom graph cuts energy function bleeds through on the same slice, but the DRF segmentation bleeds through to a new slice.

5 Discussion

Segmentation results using the DRF framework were generally pretty sharp, with some exceptions. The algorithm worked particularly well on tumors that were well isolated in space - in other words, not next to larger tissue areas of similar intensity. When tumors were attached to larger tissue areas, however, the segmentation tended to bleed through to adjacent areas, causing a large error in the precision of the segmentation.

Overall, the recall rate of 0.82 is better than 0.64, the recall rate of the custom graph cuts energy function segmentation, where the energy function was tweaked by hand. The previous algorithm also makes stronger assumptions about the location of a tumor. The precision of 0.48 is far worse than the previous precision of 0.77. The biggest contributing factor tho this poor precision are the bleed through areas in the segmentation. There were also some cases in which the segmentation was slightly overestimated, such as in Figure 5, where the false negatives are a small ring around the tumor. Since the previous graph cuts algorithm tended to consistently underestimate the segmentation, it had a much higher precision.

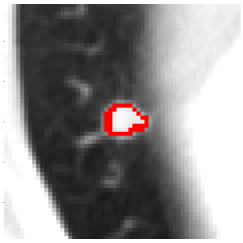


Figure 5: Areas of a segmentation that the algorithm included which was not in the original ground truth segmentation.

It is difficult to compare the performance of our current problem formulation with the performance of other studies done in the subject. For example, Huang et al. did work on the diagnosis of hepatic (liver) tumors using SVMs and texture analysis [4], but their problem was classifying tumors as benign or malignant versus actually comparing the size of the tumors. Furthermore, hepatic tumors are different from lung tumors in their characteristics. Earlier studies using Markov random fields also only addressed the problem of detection [8]. Zhang et al. did work on brain tumor classification using one-class SVMs on MRI data and managed to achieve a mean recall of around 0.9 and precision between 0.6 and 0.99, but they used both pre-contrast and contrast images in order to find a good margin [11]. Trials with two-class SVMs using our CT data did not return good results, but one-class SVMs may be something to try, and can even fit into the DRF framework. Lee et al.'s work similarly involved MRI data on brain tumors and reported results which implied their precision and recall was around 0.8 [7]. Non-learning approaches at lung tumor segmentation had mixed results as well [10]. Opfer et al. also mentioned that variation between manual segmentations by different radiologists was also significant.

The current formulation of the problem suffers from several factors, most notably the dearth of good features and issues with parameter optimization. Future work should be aimed at fix these issues. Possibilities for features include texture features and intensities of neighbors. Different optimization methods may produce more stable solutions for parameter optimization. With better features also comes the possibility of trying different discriminative models, such as large-margin classifiers, or linear models with various kernels.

The main advantage of the DRF learning framework is the automatic learning of energy function parameters for segmentation. The previous graph cuts segmentation used a custom energy function and parameters that were tweaked repeatedly by hand, and often through trial and error, until something reasonable was achieved. A possible approach to improving the DRF framework is to formulate the energy function more like our previous energy function, and learn the parameters for that automatically. Unfortunately, our previous energy function was not convex, and thus is difficult to optimize.

References

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE TPAMI*, 26(9):1124 – 1137, 2004.

[2] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *IEEE TPAMI*, 20(12):1222 – 1239, 2001.

[3] A. Delong, A. Osokin, H. Isack, and Y Boykov. Fast approximate energy minimization with label costs. In *CVPR*, 2010.

[4] Yu-Len Huang, Jeon-Hor Chen, and Wu-Chung Shen. Diagnosis of hepatic tumors with texture analysis in nonenhanced computed tomography images. *Academic Radiology*, 13(6):713 – 720, 2006.

[5] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE TPAMI*, 26(2):147 – 159, 2004.

[6] Sanjiv Kumar and Martial Hebert. Discriminative fields for modeling spatial dependencies in natural images. In *In NIPS*. MIT Press, 2003.

[7] Chi-Hoon Lee, Mark Schmidt, Albert Murtha, Aalo Bistritz, Jerg Sander, and Russell Greiner. Segmenting brain tumors with conditional random fields and support vector machines. In *Computer Vision for Biomedical Image Applications*, volume 3765 of *Lecture Notes in Computer Science*, pages 469–478. Springer Berlin / Heidelberg, 2005.

[8] H.D. Li, M. Kallergi, L.P. Clarke, V.K. Jain, and R.A. Clark. Markov random field for tumor detection in digital mammography. *Medical Imaging, IEEE Transactions on*, 14(3):565 –576, September 1995.

[9] Zhengrong Liang, J.R. MacFall, and D.P. Harrington. Parameter estimation and tissue segmentation from multispectral mr images. *Medical Imaging, IEEE Transactions on*, 13(3):441 –449, September 1994.

[10] Roland Opfer and Rafael Wiemker. A new general tumor segmentation framework based on radial basis function energy minimization with a validation study on lidc lung nodules. In *Medical Imaging 2007: Image Processing*, volume 6512, page 651217. SPIE, 2007.

[11] Jianguo Zhang, Kai kuang Ma, and Meng Hwa Er. V.: Tumor segmentation from magnetic resonance imaging by learning via one-class support vector machine. In *International Workshop on Advanced Image Technology*, pages 207–211, 2004.