Graph cut based Automatic Prostate Segmentation

Dwarikanath Mahapatra

Department of Computer Science, ETH Zurich, Switzerland.

Abstract. We propose a graph cut based automatic method for prostate segmentation using image features, context information and semantic knowledge. A volume of interest (VOI) is first identified using supervoxel oversegmentation and their subsequent classification. All voxels within the VOI are labeled prostate or background using graph cuts. Semantic information obtained from Random forest (RF) classifiers is used to formulate the smoothness cost. Use of context and semantic information contributes to higher segmentation accuracy than conventional methods using only image features.

1 Introduction

Prostate cancer is one of the leading cause of male cancer death in the USA [1]. Automatic prostate segmentation from magnetic resonance (MR) images is important for radiotherapy, prostate volumetry and calculation of prostate specific antigen (PSA) density. However it faces challenges due to: 1) variability of prostate size and shape between subjects; 2) different MR scanning protocols resulting in different image appearances and intensity ranges; 3) the lack of clear prostate boundaries due to similar intensity profiles of surrounding tissues.

Prostate segmentation from computed tomography (CT) and Ultrasound (US) images have been widely investigated [2,3]. MRI has gained popularity because of lack of ionising radiations as in the case of CT images and better contrast between soft tissue organs than US scans. However, accurate delineation of the prostate’s border remains difficult even for the human observer.

Klein et al. [4] proposed a multi-atlas matching for prostate segmentation using localized mutual information. Toth et al. [5] used shape prior deformable models while Li et al. [6] used auto context and level-sets to achieve binary segmentation of the prostate. In the recent MICCAI 2012 prostate segmentation challenge, different approaches used learning techniques [7], multi-atlases [8], active shape models (ASM) [9] and active appearance models (AAM) [10,11] to name a few. Graph cut based techniques on other prostate segmentation datasets have employed level-set based shape priors [12], and modified boundary terms [13].

In this paper we propose an automatic graph cut method for prostate segmenting using supervoxel oversegmentation and semantic information learned from training data. The primary novelty of our work is in the use of semantic information obtained from learned random forest (RF) classifiers to define a
novel smoothness cost for graph cut based prostate segmentation. We describe our method in Section 2, present experimental results in Section 3 and conclude with Section 4.

2 Methods

2.1 Volume of interest identification

In our segmentation pipeline we first identify a VOI for the prostate using supervoxels. First the test volume is normalized using the method in [14] such that the intensity values are in the range [0, 1]. The normalized volume is oversegmented into supervoxels using the method in [15]. Each supervoxel is analyzed for presence of part of the prostate using trained RF classifiers. For training, features described in Section 2.2 are extracted from ‘prostate’ and ‘background’ supervoxels which can be identified from the manually segmented training images. The set of supervoxels make up the VOI.

Figures 1 (a)-(c) show three consecutive slices from a volume. The outlines of the supervoxels for the whole image is shown in green, the manual segmentation of the prostate is shown in red, and outline of ‘prostate’ supervoxels are shown in yellow. For all voxels within the VOI, probability maps are generated (using a different set of RF classifiers) which give their likelihood of being prostate or background voxels. Semantic information is obtained from the trained RF classifiers to design a novel smoothness cost, and the final labels are obtained using graph cut segmentation.

2.2 Probability maps using RF classifiers

A probabilistic classifier like RF [16] is used to obtain class probability values for each voxel in the VOI. RFs can handle large datasets, multi-class classification and interpret learned knowledge. This section describes different image features and extraction of semantic information from the learned RF classifier.

**Image Features:** The mean and variance of intensity, texture and 3D mean curvature values are used as features. Texture maps are obtained using oriented Gabor filters at angles of \{0°, 45°, 90°, 135°\} and two scales (1, 0.5). This gives an initial feature vector of 20 values. Note that these features are calculated for every supervoxel in VOI identification. To generate probability maps of each pixel these features are calculated over a pixel’s neighborhood.

**Relative Context:** As the relative arrangement of organs is constant (except for missing organs) one organ can provide contextual information about others through relative distance and orientation. Context information has been used for medical image segmentation and registration [6,17,18,19]. Figure 1 (d) shows an illustration of the sampling scheme where the circle center is the pixel in question and the relative sample points for context features are denoted by a red ‘X’. At each ‘X’ we extract a 3 × 3 × 3 region and calculate the mean intensity, texture and curvature values. The texture values were derived from the texture maps obtained at 90° orientation and scale 1.
Fig. 1. (a)-(c) original image with superimposed supervoxels (green), manual segmentation (red) and ‘prostate’ supervoxels (yellow). Three consecutive slices for patient 41 are shown. (d) sampling diagram for context information.

The ‘X’s are at distances of 3, 8, 15, 22 and the angle between consecutive rays is 45°. The values from the 32 regions are concatenated into a 96 dimensional feature vector. The final feature vector has 116 values. Equal number of samples from prostate and background voxels (from training images) are extracted and used to train a RF classifier (different from the one trained on supervoxel features). The trained classifier is used to generate probability maps (from the number of votes for different classes) for every voxel within the identified VOI of the test volume. Each voxel has 2 probability values corresponding to prostate and background. For supervoxel feature extraction intensity and texture features remain the same. In the case of context features, the first two supervoxels along rays (in Fig 1 (d)) at 0°, 90°, 180°, 270° are chosen.

2.3 Graph Cut Segmentation

The final segmentation is obtained by optimizing a second order Markov random field (MRF) energy function which is written as

$$E(L) = \sum_{s \in P} D(L_s) + \lambda \sum_{(s,t) \in N} V(L_s, L_t),$$

where $P$ denotes the set of pixels and $N$ is the set of neighboring pixels for pixel $s$. The cost function is optimized using graph cuts [20]. $\lambda$ is a weight that determines the relative contribution of penalty cost ($D$) and smoothness cost ($V$). $D(L_s)$ is given by

$$D(L_s) = -\log (Pr(L_s) + \epsilon),$$

where $Pr$ is the likelihood (or probabilities) previously obtained using RF classifiers and $\epsilon = 0.00001$ is a very small value to ensure that the cost is a real number. Figure 2 shows the probability maps of a test image for prostate and background regions. Higher the probability for a class lower is the corresponding data penalty for that class.

**Semantic Information for Smoothness Cost** $V$ ensures a smooth solution by penalizing spatial discontinuities. We formulate the smoothness cost by using
Fig. 2. Probability maps for (a) Prostate; (b) background. Higher values indicate greater likelihood of belonging to that region. Red indicates maximum probability while blue indicates zero probability. (c) final segmentation output in green and manual segmentation in red.

semantic information from the trained RF classifier. The RF classifier returns a measure of the importance of each dimension in the feature vector to the classification task. Inspite of the multiple dimensional feature vector, the features can be classified into three types - intensity, texture and curvature. The context information is a combination of the three. By aggregating the importance values of each feature category and normalizing them we obtain the relative importance of each feature in the classification task. This provides the necessary semantic information by quantifying the importance of each feature in classifying a voxel into different categories. Let the weight of the different features be $w_I$ (intensity), $w_T$ (texture) and $w_C$ (curvature), where $w_I + w_T + w_C = 1$. The smoothness cost $V$ is given by

$$V(L_s, L_t) = \begin{cases} w_I V_I + w_T V_T + w_C V_C, & L_s \neq L_t, \\ 0, & L_s = L_t. \end{cases}$$

(3)

where $V_I, V_T, V_C$ are the individual contributions to the smoothness by intensity, texture and curvature. $V_I$ is defined as

$$V_I(L_s, L_t) = e^{-\frac{(I_s - I_t)^2}{2\sigma^2}}, \quad \frac{1}{\|s - t\|}.$$  

(4)

$I$ is the intensity. $V_T$ and $V_C$ are similarly defined using texture and curvature. To choose the value of $\lambda$ we adopt the following steps. We choose a small subset of the training data consisting of 10 patient volumes, and perform segmentation using our method but with $\lambda$ varying from 0 to 1 in steps of 0.001. Based on the segmentation accuracy using Dice Metric (DM) we set $\lambda = 0.02$. After training of the classifiers we obtain $w_I = 0.22, w_T = 0.31$ and $w_C = 0.47$.

3 Experiments and Results

As part of the MICCAI 2012 prostate segmentation challenge (http://promise12.grand-challenge.org/) the training dataset consists of 50 transversal T2-weighted MR datasets of the prostate. The dataset is a representative set of the types of
MR images acquired in a clinical setting. The data is multi-center and multi-vendor and has different acquisition protocols (e.g. differences in slice thickness, with/without endorectal coil). The set is selected such that there is a spread in prostate sizes and appearance. Reference segmentations are available for each dataset. We employ a 10 fold cross validation where training is done on 45 datasets and tested on 5 datasets. We report average test results over all patients. N3 intensity non-uniformity correction was applied to reduce intensity inhomogeneity and the image intensities are normalized using the method in [14]. Table 2 gives details about different aspects of our algorithm.

3.1 Evaluation Metrics

The quality of our segmentations was evaluated using two measures: 1) Dice Metric (DM) and 2) Hausdorff Distance (HD). DM measures the overlap between the segmented diseased region obtained by our algorithm and reference manual annotations. It is given by

\[ DM = \frac{2 |A \cap M|}{|A \cup M|} \]  \hspace{1cm} (5)

where \( A \) - segmentation from our algorithm and \( M \) - manual annotations. The DM measure yields values between 0 and 1 where high DM corresponds to a good segmentation.

**Hausdorff Distance (HD):** The DM gives a measure of how much the actual manual segmentation was recovered by the automatic segmentation. But the boundaries of the segmented regions may be far apart. The HD aims to measure the distance between the contours corresponding to different segmentations. If two curves are represented as sets of points \( A = \{a_1, a_2, \cdots \} \) and \( M = \{m_1, m_2, \cdots \} \), where each \( a_i \) and \( m_j \) is an ordered pair of the \( x \) and \( y \) coordinates of a point on the curve, the distance to the closest point (DCP) for \( a_i \) to the curve \( M \) is calculated. The HD, defined as the maximum of the DCP’s between the two curves, is:

\[ HD(A, M) = \max \left( \max_i \{DCP(a_i, M)\}, \max_j \{DCP(m_j, A)\} \right) \]  \hspace{1cm} (6)

3.2 Segmentation Results

We present segmentation results for Table 1 summarizes the performance of Our (our proposed method) under different conditions. Our\(_{nC} \) - Our without context information from images for training the RF classifier; Our\(_{nV} \) - Our without semantic information in \( V \). \( w_I = w_T = w_C = 0.33 \); Our\(_{nV_T} \) - Our without intensity in \( V \); Our\(_{nV_C} \) - Our without texture in \( V \); Our\(_{nV_C} \) - Our without using curvature in \( V \);

The results show a significant reduction in segmentation accuracy without using context information. A \( t \)–test between the values of Our and Our\(_{nV} \) gives
Table 1. Quantitative measures for prostate segmentation. DM- Dice Metric in % and HD is Hausdorff distance mm

<table>
<thead>
<tr>
<th></th>
<th>Our</th>
<th>Our$_{nC}$</th>
<th>Our$_{nV_T}$</th>
<th>Our$_{nV_C}$</th>
<th>Our$_{nV}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>90.2±4.1</td>
<td>81.2±4.7</td>
<td>89.0±3.8</td>
<td>87.2±4.1</td>
<td>86.5±3.9</td>
</tr>
<tr>
<td>HD</td>
<td>1.4±2.2</td>
<td>13.2±4.2</td>
<td>5.6±2.8</td>
<td>6.8±3.2</td>
<td>7.9±3.3</td>
</tr>
</tbody>
</table>

$p_{Our-Our_{nV}} < 0.002$ indicating a large degree of difference between the results. Other similar tests show that $Our$ performs significantly better (with $p < 0.01$ in all cases) than all other methods. $Our$’s better performance can be attributed to two factors: 1) use of machine learning techniques to identify most discriminant features; and 2) incorporating semantic information into the smoothness cost.

Figure 3 shows segmentation results for training Patient 38 obtained by $Our$, $Our_{nC}$, $Our_{nV_C}$ and $Our_{nV_T}$. The results reflect the values in Table 1 with $Our_{nC}$ giving the worst results. The low DM values for $Our_{nC}$ highlights the fact that context plays a very important role in our method. Although all the low level features are used, without context information it is very difficult to discriminate between prostate and background. For other methods although one low level feature is excluded, context information provides greater discrimination.

4 Conclusion

We have presented a method for prostate segmentation that is fully automated and does not require any manual intervention. As part of our method we use supervoxel segmentation and subsequent classification of supervoxels to identify a VOI. We have developed novel context features to train a Random forest classifier and classify each voxel as prostate or background. Additionally we incorporate semantic information from trained RF classifiers into the smoothness cost to achieve higher segmentation accuracy. Experimental results on the MICCAI 2012 prostate segmentation database show our method is highly accurate, robust and can efficiently segment images from different vendors.

References

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>MATLAB</td>
</tr>
<tr>
<td>Libraries/Packages</td>
<td>Image Processing toolbox, Graph cut optimization, Random Forest Classifiers, SLIC Supervoxel Segmentation</td>
</tr>
<tr>
<td>Multi Threaded</td>
<td>No</td>
</tr>
<tr>
<td>User Interaction</td>
<td>None</td>
</tr>
<tr>
<td>CPU Clock Speed</td>
<td>2.66 GHz</td>
</tr>
<tr>
<td>Machine CPU Count</td>
<td>2</td>
</tr>
<tr>
<td>Machine Memory</td>
<td>4 GB</td>
</tr>
<tr>
<td>Memory Used during Segmentations</td>
<td>&lt; 1 GB</td>
</tr>
<tr>
<td>Training Time (Supervoxel Classifier)</td>
<td>3 hours</td>
</tr>
<tr>
<td>Training Time (Voxel Classifier)</td>
<td>5 hours</td>
</tr>
<tr>
<td>Image Preprocessing time</td>
<td>2-3 mins</td>
</tr>
<tr>
<td>VOI Detection Time</td>
<td>2-4 mins</td>
</tr>
<tr>
<td>Voxel Classification + Segmentation Time</td>
<td>2-5 mins</td>
</tr>
<tr>
<td>Total Segmentation Time</td>
<td>7-12 mins</td>
</tr>
</tbody>
</table>

Table 2. Parameter values used in segmentation