AUTOMATIC DIFFERENTIAL SEGMENTATION OF THE PROSTATE IN 3-D MRI USING RANDOM FOREST CLASSIFICATION AND GRAPH-CUTS OPTIMIZATION

Emmanouil Moschidis, Jim Graham

Imaging Science and Biomedical Engineering, School of Cancer and Enabling Sciences, Manchester Academic Health Science Centre, The University of Manchester, Oxford Road, Manchester M13 9PT.
Emmanouil.Moschidis@postgrad.manchester.ac.uk, Jim.Graham@manchester.ac.uk

ABSTRACT

In this paper we address the problem of automated differential segmentation of the prostate in three dimensional (3-D) magnetic resonance images (MRI) of patients with benign prostatic hyperplasia (BPH). We suggest a framework that consists of two stages: in the first stage, a Random Forest classifier localizes the anatomy of interest. In the second stage, Graph-Cuts (GC) optimization is utilized for obtaining the final delineation. GC optimization regularizes the hypotheses produced by the classification scheme by imposing contextual constraints via a Markov Random Field model. Our method obtains comparable or better results in a fully automated fashion compared with a previous semi-automatic technique [6]. It also performs well, when small training sets are used. This is particularly useful in on-line interactive segmentation systems, where prior knowledge is limited, or in automated approaches that generate ground truth used for model-building.

Index Terms — Automatic Segmentation, Prostate Zones, MRI, Random Forests, Graph-Cuts.

1. INTRODUCTION

Biomedical image segmentation is a useful tool for clinical diagnosis and treatment planning. However, it is also a labor intensive task for radiologists, who segment a large number of images as part of their clinical routine. This is due to the fact that this task is often performed interactively, but also in some cases manually, thus involving a high cognitive load and time consumption for the human expert. Therefore, the automation of this process whenever this is possible is highly desired. In this paper we address the problem of automating the differential segmentation of the prostate in three dimensional (3-D) magnetic resonance images (MRI). The study was done in the context of identifying Benign Prostatic Hyperplasia - a non cancerous enlargement of the prostate that affects 50% of men over 60 years old [1].

The prostate is anatomically divided into the Peripheral (PZ), Central (CZ), Transitional (TZ) and Fibromuscular (FZ) zones. In BPH the prostatic enlargement is mainly due to the volumetric increase in the TZ. Therefore the estimation of the TZ volume and the TZ ratio (TZ volume/total prostate volume) is important for monitoring the progress of the disease and the effectiveness of drug treatments [2]. In MRI two regions are identified: the PZ and the Central Gland (CG), which includes the other three anatomical zones (fig. 1). However, in BPH the TZ is the predominant zone in the CG due to its expansion and therefore TZ and CG can be considered as equivalent [6]. Differential segmentation of the prostate is challenging. The appearance of the central and peripheral glands varies significantly among individuals. Furthermore surrounding tissue (seminal vesicles, blood vessels, the urethra and the bladder) present contrast challenges at different locations.

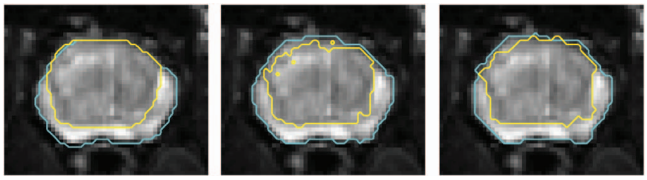


Fig. 1. An example axial mid-section of a prostate, depicting the ground truth (left), the RF hypothesis (middle) and the final GC regularized result (right). The yellow and cyan contours delineate the central gland and the total prostate respectively.

While some studies have addressed the problem of total prostate (ToP) segmentation in MRI [3, 4, 5], to the best of our knowledge it is only the study of Allen *et al.* [6] that addresses the problem of differential prostate segmentation. In their method they combine gray-level voxel classification and a double surface 3-D point distribution model. During classification, three Gaussian intensity models, which correspond to background, PZ and CG, are fit to the data. Subsequently, the 3-D point distribution model regularizes the classification output by providing smooth spatial constraints. The model fitting in this method is performed semi-automatically.

In this paper we present a fully automated method for the differential segmentation of the prostate. Our method operates in two stages: in the first stage a Random Forest (RF) classification scheme is used to localize the anatomy of interest [8]. In the second stage the final delineation is achieved via Graph-Cuts (GC) optimization. The optimization stage regularizes the hypotheses produced by the classification scheme by imposing contextual constraints via a Markov Random Field (MRF) model. In the context of comparing our method with the one of Allen *et al.* [6], we obtain comparable or better results in a fully automated fashion. Our method also performs well, when small training sets are utilized. This is particularly useful in on-line interactive segmentation systems, where prior knowledge is limited but may assist in reducing the user's cognitive load if it is utilized, or in automated approaches that generate ground truth for model-building algorithms.

2. METHODS

2.1. Dataset

We use the same dataset as in the study of Allen *et al.* [6], which consists of 22 3-D T2 fat suppressed MR images of the prostate from individuals with BPH. T2 fat suppressed MRI provides good contrast not only between the prostate and its surrounding tissue, but also between the prostatic anatomical zones. The images were acquired using a 1.5T Phillips Gyroscan ACS MR scanner. After their acquisition, all images were cropped close to the prostate as shown in fig. 1. The ground truth for each image is a binary volumetric mask produced after averaging the manual delineation of two radiologists on the cropped images. Prior to the experiments, the intensities of all images were normalized to lie in the bounded interval [0, 255]. Lastly, all images were resampled to allow for an iso-voxel resolution and volumes of equal sizes to be created.

2.2. Random Forests

RF classification constitutes the first stage of our framework, which aims for the localization of the anatomy of interest, given a set of segmented examples (training set). RF is a machine learning algorithm, which learns the properties of the data in the training set and makes predictions over unseen data using a linear combination of multiple base learners [8]. In classification, these learners are classifiers. In binary RF classification, the classifiers are *decision trees*, which are combined via a *majority voting* scheme. In such a scheme, the voxel in question is assigned the label of the class with more than 50% of the total votes. The winning votes can be seen as likelihood estimates.

Decision trees operate via test functions, similar to if-then rules, on observations of different image properties, termed *features*. In RF, each tree is trained on a *bootstrap* sample of the training set, using a random subset of the available features. A bootstrap sample is created from a given sample via random sampling with replacement. Combining the decision trees results in a classifier that yields outcomes with state-of-the-art accuracy. RF is fast during both training and prediction, it is inherently parallel and makes no assumption about the distribution of the data in the feature space. In this study we used the RF implementation available from [9].

In order to construct a feature space, which can encode the complex appearance of the anatomy of interest, we use voxel descriptors consisting of the gray level intensities, the x, y, z spatial coordinates, the Haralick [11] and the Laws [12] features. The last two sets of features are known to provide energy measures that are useful for the description of textured images. A similar approach has been adopted in [10], where textural measures are utilized to build the feature space of complex 2-D natural images, as part of an interactive segmentation system based on RF learning.

2.3. Graph-Cuts

GC optimization constitutes the second stage of our framework, which aims for the regularization of the RF hypothesis about the anatomy of interest and the provision of its final delineation (fig.1). This is achieved through the imposition of a MRF model, which incorporates contextual constraints in terms of neighboring voxel interactions. The maximum posterior probability (MAP) estimate of a MRF configuration of binary images can be computed by max-flow/min-cut algorithms [13]. The energy function that is usually minimized in segmentation problems is formulated as:

$$E(f)=\lambda\sum_{p\in P}D_p(f_p)+\sum_{\{p,q\}\in N}V_{pq}(f_p,f_q) \quad (1)$$

The first term in eq. 1 is called the *data* or *regional* term and incorporates the hypotheses that a voxel *p is* part of the foreground or the background set (having its label f_p assigned as foreground or background). In this study, these hypotheses are provided by the RF classification stage. The second term is the *smoothness* or *boundary* term, which encompasses the boundary constraints of the MRF model. It is also known as the interaction potential, since it handles the pairwise interactions between two voxels *p* and *q* of a neighborhood *N*. This term acts as the regularizer of the regional hypotheses and is written, following the notation of [14], as:

$$V_{pq}(f_p,f_q)=w_{pq}\cdot\delta(f_p\neq f_q) \quad (2)$$

$$\text{with } w_{pq}=\exp\frac{-(I_p-I_q)^2}{2\sigma^2} \quad (3)$$

where δ is a Kronecker delta function and I_p and I_q the intensities of voxels *p* and *q* respectively. The tunable parameters λ and σ, control the relative importance of the two terms in eq. 1 and the full width at half maximum of the Gaussian function's peak in eq. 3 respectively. In the context of this study we used our own implementation [7] of the GC algorithm described in [15].

2.4. Evaluation Metrics

We assess the performance of our method with a score of classification accuracy (*CA*), the Tanimoto coefficient (*Tc*), the volumetric difference (*VolDiff*) between the ground truth and the segmentation volumes and the maximum and mean point to surface distance between the segmentation and the ground truth surface (*MaxDist* and *MeanDist* respectively).

In order to calculate the *CA* metric, all the voxels are classified into true and false positives (*TP*, *FP*) and true and false negatives (*TN*, *FN*). The *CA* score is then defined as:

$$CA=100\times\frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}\ \% \quad (4)$$

The Tanimoto coefficient *Tc* is computed as:

$$Tc=100\times\frac{|TP|}{|TP|+|FP|+|FN|}\ \% \quad (5)$$

The *VolDiff* metric is defined as:

$$VolDiff=100\times\frac{||FN|-|FP||}{|TP|+|FN|}\ \% \quad (6)$$

The *MaxDist* and *MeanDist* metrics are calculated via a 3-D distance transform. The distance is given in voxels, but since we use images with iso-tropic voxels, this result can be reported in millimeters as well.

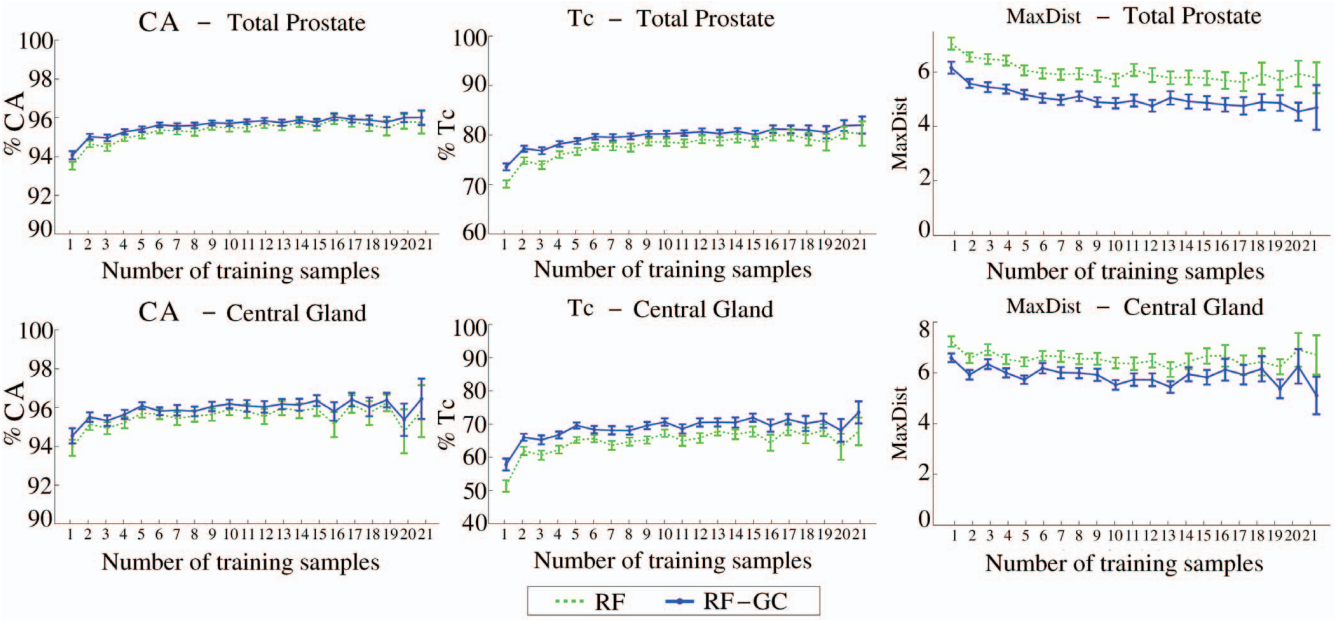


Fig. 2. Graphs depicting the performance of our framework before (RF) and after (RF-GC) regularization with respect to the *CA* (left), *Tc* (middle) and *MaxDist* (right) accuracy metrics for the Total Prostate (top row) and the Central Gland (bottom row) segmentation tasks. The error bars represent the ± 1.96 × standard error of the mean.

The *VolDiff* and *MeanDist* metrics are used to allow for the direct comparison between our results and those reported in [6]. All metrics quantify different aspects of the accuracy of segmentation.

3. EXPERIMENTS AND RESULTS

In order to observe the dependency of our method's accuracy on the number of the training samples, we created 21 training sets with sizes that ranged from 1 to 21. The images in each set were selected randomly. Those that did not take part in the training set, were used to form the test set. The process was repeated 10 times to allow for the production of training sets with considerable variability, leading to a total of 210 training sets, 10 from each size.

Having training sets with 3-D images leads to a large number of voxels that need to be stored in memory during the training phase of RF. This issue intensifies when a large number of features is also utilized. In order to improve the computational efficiency of the method, we performed the following actions: firstly, we sampled a fixed number of voxels from each training set, instead of using all the available voxels. The same number of voxels was sampled from each image of the training set, using a proportionate stratified sampling strategy. Secondly, we performed feature selection before training the final RF classifier, which led to the reduction of the number of features used from 753 to 16.

During RF training about a third of the training data are separated from the original training set and they are used by the algorithm as a validation test. In this stage an error estimate and the importance of the features in the prediction accuracy are provided. We used the *Gini impurity criterion* [8] for selecting the most important features, which assesses how well separated (pure) are the data, once a test function (split) is performed on a tree node. It is noteworthy that during the feature reduction phase, we used a sparse sampling strategy, sampling 5×10^3 voxels from the training set. Once the feature selection was performed, we trained the final RF classifier using denser sampling (5×10^4 sampling voxels per training set). The final RF model was then used for the localization of the ToP and the CG on the unseen images.

Once the RF hypotheses were obtained, we performed GC regularization. The σ and λ parameters were set for all the regularization experiments as follows: $\sigma=10$ and $\lambda=0.1$.

Fig. 2 and table 1 summarize the segmentation results obtained from the first (RF) and second (RF-GC) stage of our framework. Overall, GC regularization improves the RF-based segmentation outcome. The main effect of the regularization is the removal of holes inside the foreground and islands outside it, which exist in the RF hypotheses (fig. 1). Also, in both ToP and CG segmentation, all metrics demonstrate improved segmentation accuracy as the size of the training set increases. The recorded mean values for the *CA*, *Tc* and *MaxDist* score for ToP segmentation improve from 94.1%, 73.6% and 6.2 mm for training sets of size 1 to 96.0%, 82.0%, and 4.7 mm respectively for training sets of size 21. The recorded mean values for the same scores for CG segmentation improve from 94.5%, 57.8% and 6.6 mm for training sets of size 1 to 96.5%, 73.6%, and 5.1 mm for training sets of size 21. It is noteworthy that the method provides accurate results even with small sized training sets (≈10 images).

Table 1 presents the *VolDiff* and *MeanDist* scores, for training sets of size 21, which can be considered equivalent to the leave-one-out experiments used in the study of Allen *et al.* [6]. Our method obtained better *VoldDiff* score for the ToP segmentation task but slightly worse for the CG segmentation task. The *MeanDist* score was better in both segmentation tasks. The CG is a small anatomical region. In such cases, few misclassified voxels

lead to a large decline of the accuracy metrics. We believe that our results are promising, given the fact that our method is fully automated and that no shape priors are imposed to the final delineation.

Table 1: *VolDiff* and *MeanDist* results for the Total Prostate and the Central Gland segmentation tasks.

	Metrics	
Method (ROI)	***VolDiff (%)*** ***μ (σ)***	***MeanDist (mm)*** ***μ (σ)***
RF (ToP)	8.5 (7.2)	1.1 (0.2)
RF-GC (ToP)	**7.3 (4.2)**	**1.0 (0.1)**
Allen *et al.* [6] (ToP)	11.1 (9.5)	4.1 (1.1)
RF (CG)	21.3 (13.1)	1.5 (0.4)
RF-GC (CG)	13.3 (10.9)	**1.2 (0.3)**
Allen *et al.* [6] (CG)	**11.9 (8.9)**	3.1 (2.5)

4. CONCLUDING REMARKS

In this study we presented a fully automated approach for the differential segmentation of the prostate. Our method consists of two stages: in the first stage, a RF classification produces hypotheses about the anatomy of interest, thus tackling the localization problem. In the second stage, GC optimization is used to regularize the previous hypotheses via the MAP-MRF framework. Our empirical performance evaluation shows that the framework can provide accurate segmentation outcomes for a challenging task. Our method demonstrates comparable or better results in a fully automated fashion compared with the semi-automatic method of Allen *et al.* [6].

Furthermore, fig. 2 demonstrates that the method can be used to produce fairly accurate results using small training sets. This is particularly useful for on-line interactive segmentation. Model-based approaches, such as that of [6] require an accurately annotated training set, which is generally obtained by manual segmentation. Fig. 2 shows that automatic classification starts improving after the first training example and classification close to final performance can be achieved after a small number of training examples. This can lead to a dramatic improvement in cognitive load in the training phase of statistical model building.

In our study the ground truth was produced by delineation of the anatomy of interest on a slice by slice basis, thus producing a result that is unlikely to be in complete agreement with the segmentation outcome of a 3-D segmentation method.

It is also noteworthy, that no shape priors were incorporated in the suggested approach. Future work could investigate if such priors could offer any further improvement to the segmentation accuracy.

5. ACKNOWLEDGEMENTS

The project has been partially funded by the United Kingdom Biotechnology and Biological Sciences Research Council (BBSRC).

6. REFERENCES

[1] A. Thorpe and D. Neal, "Benign prostatic hyperplasia," *Lancet*, vol. 361, no. 9366, pp. 1359–67, 2003.

[2] A. Tewari, K. Shinohara and P. Narayan, "Transitional Zone Volume and Transitional Zone Ratio: Predictor of Uroflow Response to Finasteride Therapy in Benign Prostatic Hyperplasia Patients," *Urology*, vol. 45, no. 2, pp. 258–265, 1995.

[3] N. Makni *et al.*, "Automatic 3D segmentation of prostate in MRI combining a priori knowledge, Markov fields and Bayesian framework," in *Proc. EMBS,* 2008, pp. 2992-2995.

[4] S. Klein *et al.*, "Automatic segmentation of the prostate in 3-D MR images by atlas matching using localized mutual information," *Medical Physics*, vol. 35, no. 4, pp. 1407–1417, 2008.

[5] E. Moschidis and J. Graham, "Propagating interactive segmentation of a single 3D example on similar images: an evaluation study using MR images of the prostate," in *Proc. ISBI*, 2011, pp. 1472– 1475.

[6] P. D. Allen *et al.,* "Differential segmentation of the prostate in MR images using combined 3D shape modelling and voxel classification," in *Proc. ISBI*, 2006, pp. 410–413.

[7] E. Moschidis and J. Graham, "A systematic performance evaluation of interactive image segmentation methods based on Simulated User Interaction," in *Proc. ISBI*, 2010, pp. 928– 931.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[9] A. Jaiantilal, "randomforest-matlab," November 2011. [Online]. Available: http://code.google.com/p/randomforest-matlab/

[10] J. Santner *et al.*, "Interactive Texture Segmentation using Random Forests and Total Variation," in *Proc. BMVC*, 2009, pp. 1-12.

[11] R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

[12] M. T. Suzuki, Y. Yaginuma and H. Kodama, "A texture energy measurement technique for 3D volumetric data," in *Proc. SMC*, 2009, pp. 3779-3785.

[13] D. Greig, B. Porteous, and A. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society, Series B*, vol. 51, no. 2, pp. 271-279, 1989.

[14] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *Proc. ICCV,* 2001, vol. 1, pp.105–112.

[15] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE T-PAMI*, vol. 26, no. 9, pp. 1124–1137, 2004.