Clinical Target-Volume Delineation in Prostate Brachytherapy Using Residual Neural Networks

Emran Mohammad Abu Anas[1](✉), Saman Nouranian[1], S. Sara Mahdavi[1], Ingrid Spadinger[2], William J. Morris[2], Septimu E. Salcudean[1], Parvin Mousavi[3], and Purang Abolmaesumi[1]

[1] Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada
emrana@ece.ubc.ca
[2] Vancouver Cancer Centre, Vancouver, BC, Canada
[3] School of Computing, Queen's University, Kingston, ON, Canada

Abstract. Low dose-rate prostate brachytherapy is commonly used to treat early stage prostate cancer. This intervention involves implanting radioactive seeds inside a volume containing the prostate. Planning the intervention requires obtaining a series of ultrasound images from the prostate. This is followed by delineation of a clinical target volume, which mostly traces the prostate boundary in the ultrasound data, but can be modified based on institution-specific clinical guidelines. Here, we aim to automate the delineation of clinical target volume by using a new deep learning network based on residual neural nets and dilated convolution at deeper layers. In addition, we propose to include an exponential weight map in the optimization to improve local prediction. We train the network on 4,284 expert-labeled transrectal ultrasound images and test it on an independent set of 1,081 ultrasound images. With respect to the gold-standard delineation, we achieve a mean Dice similarity coefficient of 94%, a mean surface distance error of 1.05 mm and a mean Hausdorff distance error of 3.0 mm. The obtained results are statistically significantly better than two previous state-of-the-art techniques.

Keywords: Object segmentation · Deep convolutional neural networks · Residual networks · Dilated convolution · Clinical target volume · Prostate

1 Introduction

Low dose-rate brachytherapy is a widely used procedure to treat localized prostate cancer [6]. The initial step involves planning the implantation of radioactive seeds in a target anatomy. For this purpose, axial transrectal ultrasound (TRUS) images are acquired from the base to the apex of the prostate followed by clinical target volume (CTV) delineation. The International Commission of Radiation Units and Measurements, defines CTV as a patient-specific contour that does not necessarily follow the exact prostate boundary and may contain

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 365–373, 2017.
DOI: 10.1007/978-3-319-66179-7_42

neighboring tissues (Fig. 1) [10]. The objective of enclosing adjacent tissue to the prostate, in CTV, is to ensure the inclusion of subclinical malignant disease. The extent of such inclusion is subject to the experience of the radiation oncologist and institution-specific clinical guidelines. Based on the CTV-boundary, further modifications are carried out to determine the volume for implantation.

Current clinical approaches for CTV delineation are mostly based on manual or semi-automatic techniques; they usually consist of prostate segmentation (e.g. [5]) followed by manual correction from an expert to rectify the boundary and modifications to include adjacent tissue. An automatic CTV delineation is highly desirable to reduce user interaction and inter-observer variability. In addition, an automatic technique ensures that certain institution-specific clinical guidelines (e.g., smooth and symmetric/nearly-symmetric contour) are met, which are often difficult to maintain in manual delineation approaches.

Recently, Nouranian *et al.* proposed automatic approaches for CTV generation by learning these boundaries from a set of clinician-generated, gold-standard contours [7,8]. In [7], a multi-atlas fusion framework was presented that transformed multiple training datasets to the coordinates of a test image. Subsequently, the CTV was obtained by fusing the training image labels. They extended their work in [8] by learning a sparse dictionary from a training set of TRUS images and their label maps using K-singular value decomposition.

In this paper, we present an alternative learning technique based on deep convolutional neural networks (CNN) for automatic delineation of CTV in TRUS images. In contrast to the dictionary learning technique in [8] that uses a linear combination of dictionary elements, CNN can produce a non-linear function to capture the relationship between input images and CTV. The use of deep CNN in visual recognition tasks has gained popularity due to their superior performance compared to the state-of-the-art techniques. Recently, deep CNN were successfully applied for medical image segmentation [9], where feed-forward nets were put together to develop a ‘U’-shape architecture. We propose an extended CNN architecture, hereafter referred to as ResU-net, that is built upon U-net architecture. In addition, we introduce three major modifications to improve optimization and to enable accurate local prediction (Fig. 2(a)):

1. We use residual neural networks (res-net) [3] as the constituent net in the proposed architecture instead of plain feed forward networks. A deep res-net is easier to optimize compared to its plain counterpart [3]. In addition, a built-in property of res-net is that features from each layer are combined with the obtained features in the next layer. Therefore, higher level features (Fig. 2(b), right branch) and lower level features (Fig. 2(b), left branch) are combined to make the final prediction [9].
2. We use dilated convolution [11] to incorporate the learnt features at deeper layers in the final outcome, without loss of resolution.
3. Rather than allowing for equal weights for all pixels of the prostate in a training label map, our loss function assigns higher values to pixels close to the CTV boundary for improved local prediction.

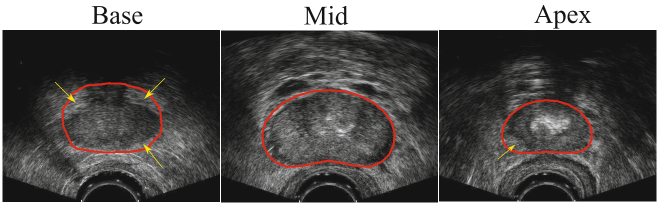

Fig. 1. Examples of TRUS images in different prostate regions with superimposed gold-standard CTV delineation. The gold-standard (red colored contour) is obtained by using a semi-automatic prostate segmentation technique [5] followed by clinical expert correction based on institutional guidelines. The inclusion of surrounding tissue at the base and apex slices is shown by arrows.

Our proposed architecture is able to accurately determine CTV with respect to gold-standard delineation, and performs statistically significantly better than two state-of-the-art techniques.

2 Methods

2.1 Materials

A total of 5,365 clinical TRUS images from 598 subjects were used in this study. Transverse TRUS images were acquired from the base to the apex of the prostate at 5 mm axial spacing using a side-firing endorectal ultrasound probe. For each patient, 7–16 TRUS images of size 415 × 490 pixels were collected with a pixel spacing of 0.15 mm × 0.15 mm. These 2D images were initially segmented using a semi-automatic technique [5]. Subsequently, each segmented contour was modified by one of seven radiation oncologists at our institution to generate the gold-standard CTV contours.

We resize all of the TRUS images to 224 × 224 pixels to fit them to the input of our proposed network. For training, validation and test, we divided the TRUS images into three mutually exclusive groups with no overlapping patients (60% vs 20% vs 20%), and they consisted of a total of 3,212 images from 358 subjects, 1,072 images from 120 subjects, and 1,081 images from 120 subjects, respectively.

2.2 Network Architecture

The proposed network architecture is demonstrated in Fig. 2(a). It takes an input image of size of 224 × 224 pixels and generates a corresponding label map of the same size. The network consists of a contracting and an expanding branch. Each branch comprises six layers and each layer includes two sublayers. We follow the res-net architecture [3] to develop our sublayer unit as shown in Fig. 2(b).

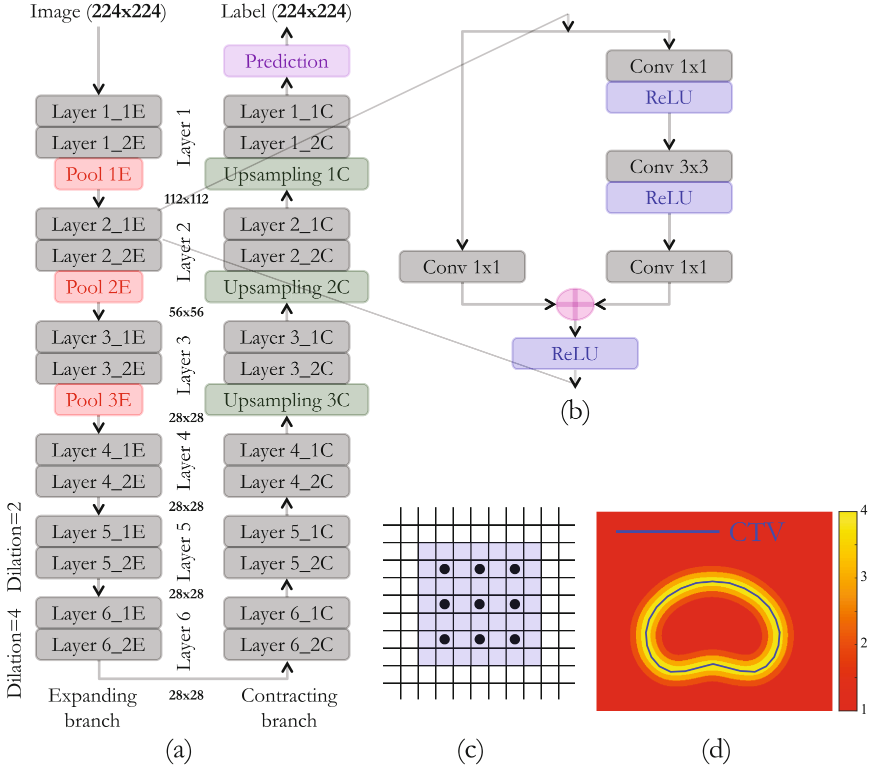

Fig. 2. The proposed deep CNN-based approach. (a) The network architecture. (b) Architecture of a sublayer unit. (c) Increased size of the receptive field of a 3×3 convolutional kernel with a dilation factor of 2. (d) The proposed weight map, which assigns highest weight to the CTV boundary and decreases exponentially.

Two kinds (1×1 and $3{\times}3$) of convolutions are used in the network. Zero padding is applied in each convolutional layer to keep the size of feature maps unchanged after the convolution operation. The number of feature maps successively increases from lower to higher layer. We set the number of feature maps to 64, 128, 256, 512, 512, and 512 for layers 1–6, respectively. Note that within each layer, the number of feature maps across multiple sublayers remains the same. In addition, dilated convolutions are applied at layers 5 and 6 in both branches (with factor = 2 and 4 at layers 5 and 6, respectively). The advantage of using a dilated convolution is to increase the size of the receptive field without loss of resolution and using the same number of weights; e.g., a 3×3 convolution with a dilation factor of 2 represents an effective receptive field size of 7×7 (Fig. 2(c)).

Rectified linear unit (ReLU) is used to introduce non-linearity in the network. At layers 1–3 in the contracting path, we use max-pooling with stride of 2 pixels. In the expanding path, we have upsampling units that successively increase the size of the feature maps through interpolation with a factor of 2. Finally, we use a softmax layer to predict the label map for any given input image. The prediction layer simply consists of a 1×1 convolutional layer with the number of feature maps equal to the number of possible labels (i.e., background and CTV).

2.3 Training and Loss Function

We implemented our proposed network in Caffe [4]. The weights of the convolutional layers are initialized as small random numbers. In addition, the upsampling unit is initialized by setting the weight matrix to a bilinear interpolation kernel. We consider the negative log-likelihood of the softmax probability (cross-entropy) as the loss function. Note that individual cross-entropy loss from each pixel contributes to the total loss, and we propose to provide more weights to the pixels close to the labeled CTV boundary (Fig. 2(d)). The weights are maximum on the CTV boundary and decrease exponentially. The idea of non-uniform weights is also used in other works (e.g., in 3D U-net [1]); however, they assign higher weights to all the pixels enclosed by the labeled boundary. In our approach, we assign more weights only to the adjacent pixels of the boundary, which leads to learning the CTV boundary more effectively. The validation set is used to choose the optimum value of the maximum weight and the decay rate of the exponential function. In addition, given a GPU memory size, one has to balance between the number of layers and the number of feature maps to achieve optimal performance. For this purpose, we train the network for a different number of feature maps as well as a different number of layers. The results, based on the validation set, suggest that it is preferred to increase the number of feature maps rather than the number of layers given a GPU memory size. Therefore, we choose only two sublayers in each main layer with more number of feature maps over more layers with less number of feature maps. The training time for this architecture was 14 h 23 min.

3 Evaluation

To evaluate our proposed approach, a set of 1,081 test TRUS images were used. We computed the Dice similarity coefficient (DSC), mean surface distance error (mSDE), and Hausdorff distance error (HDE) to quantify the estimated CTV contour with respect to the gold-standard contour. The DSC measures the similarity between two samples A and B, and it is defined as the percentage ratio of twice the number of elements in the intersection of A and B to the summation of the number of elements in A and B, i.e., $\mathrm{DSC} = \frac{2n(A \cap B)}{n(A)+n(B)} \times 100$, where $n(.)$ indicates the total number of elements in the set. The SDE at each point in the estimated contour is defined as its Euclidean distance to the closest neighboring point in the reference contour. mSDE and HDE are subsequently defined as the average and maximum of SDEs across all points in the contour, respectively.

For comparison, we chose two state-of-the-art techniques: U-net [9], and dictionary-based CTV delineation (DBCTV) [8]. To determine the statistically significant difference between the accuracies of the competing methods, we performed paired rank-sum test. In addition, we broke down our results across different prostate regions to investigate the dependencies of the accuracies on the prostate regions. Furthermore, we recorded the run-time of our trained network to generate a label map from an input image.

We also analyzed the effect of dilated convolution and exponential weight map (in the loss function) on the evaluation indices. To eliminate the effect of dilation, we modified the proposed network by introducing max-pooling operations at layers 4–5 in the expanding branch, removing the dilation effect in convolution at layers 5–6 in both expanding and contracting branches, and introducing upsampling at layers 4–5 in the contracting branch. To investigate the effect of the weight map, we nullified this effect by considering uniform weights for all pixels in the loss function.

4 Results

Comparison with state-of-the-art: A comparison of the evaluation indices among the proposed network (ResU-net), U-net and DBCTV is provided in Table 1. Our method demonstrates a statistically significant improvement over U-net and DBCTV in all three evaluation indices based on paired rank-sum test ($p << 0.05$). Some example comparisons are shown in Figs. 3(a-h) for different regions of the prostate (base, mid and apex), and we can observe the improvements in CTV prediction using our network.

Comparison across prostate lobes: Figures 3(i-k) show comparisons among accuracies at the base, mid and apex regions for our method. We observed drops in the accuracies at base and apex regions compared to those at mid lobe.

Effect of dilation: We notice a decrease in the accuracy (Table 1) when dilated convolutions are removed at layers 5–6. It is interesting to note that the results without using dilated convolution are still better than those for the U-net, and it indicates the superiority of res-net over the plain feed forward nets.

Effect of the weight map: With uniform weights, we observed a drop in the accuracies (Table 1), indicating that effective learning of the CTV boundary

Table 1. Comparative results of the proposed network, and effect of dilated convolution and weights.

Method	DSC (%)	mSDE (mm)	HDE (mm)
U-net	89.62 ± 4.98	1.52 ± 0.97	4.22 ± 2.26
DBCTV	89.92 ± 5.18	1.45 ± 0.83	3.93 ± 2.17
ResU-net	**93.67 ± 3.71**	**1.05 ± 0.71**	**2.96 ± 2.05**
ResU-net w/o dilated convolution	91.12 ± 4.16	1.26 ± 0.86	3.96 ± 2.09
ResU-net w/o exponential weights	93.48 ± 4.06	1.13 ± 0.81	3.41 ± 2.18

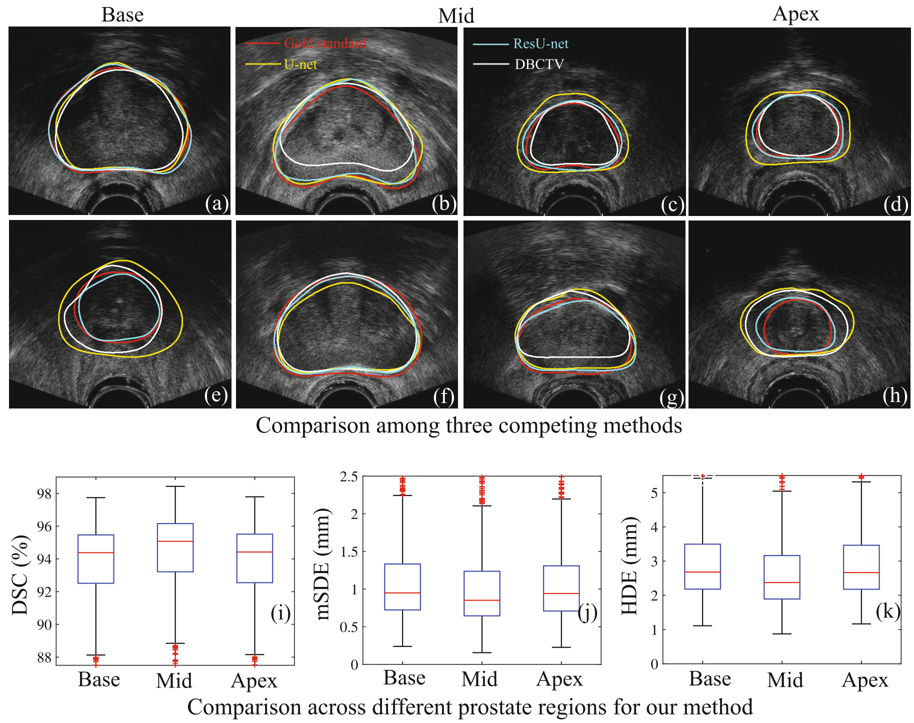

Fig. 3. CTV delineation results. (a-h) An example comparison of our technique with U-net and DBCTV indicates substantial improvement. (i-k) Comparison among delineation accuracies across different prostate regions for our proposed method.

was due to the proposed weight map. Again note that even the errors without exponential weights are better than those for both of the competing methods.

Computation time: The computation time obtained using our GPU (NVIDIA GeForce GTX 980 Ti) is 0.15 s for a 2D 415 × 490 pixels TRUS image. This time includes resizing the input TRUS image to 224 × 224 pixels, generating the label map through CNN, and resizing the label back to 415 × 490 pixels.

5 Discussion and Conclusion

We presented ResU-net, a new learning architecture for fast, automatic delineation of CTV in TRUS images of the prostate. ResU-net is based on res-net, which leads to an improved optimization of the network. The superiority of res-net has inspired a number of groups (e.g., [2] for endoscopic video segmentation). In this work, in addition to res-net, we proposed to use dilated convolution to take advantage of increased size of receptive field without loss of resolution. Moreover, we introduced an exponential weight map in the loss function that allowed us to obtain further improvement in CTV delineation. We achieved a mean DSC of 94%, a mean mSDE of 1.05 mm, and a mean HDE of 2.96 mm from a set of 1,081 test images. We demonstrated statistically significant improvement in CTV delineation compared to two state-of-the-art methods.

The drop of accuracy at the base and apex regions can be attributed to the expert's bias in the gold-standard delineation. As the prostate boundary is poorly visible in those regions, we observed experts to be biased to the prior semi-automatically generated delineation results. Such bias in gold-standard reduces the effectiveness in training, subsequently affecting the delineation accuracy.

Introducing dilated convolution at deeper layers also helps with significant improvement in our result (compare fourth and fifth rows in Table 1). Including exponential weight map also resulted in statistical significant improvement, however, we noticed more improvement for the SDE and HDE indices (less p-values). The probable reason is that the proposed exponential weight map emphasizes learning features related to the CTV contour. Therefore, the above mentioned error indices based on contour discrepancies were improved more.

A limitation of ResU-net is the absence of a batch normalization layer that makes the training easier and less dependent on the initialization. In order to include a batch normalization layer, we need to maintain a certain mini-batch size, and a GPU with more memory solves such problem. Another limitation is the loss of information due to the resizing operation between images of size 415×490 pixels and 224×224 pixels. One possible solution is to include the resizing (interpolation) operation in the deep network, subsequently learning the interpolation weights through training.

Future work includes the extension of ResU-net to delineate 3D volumes. Furthermore, adoption of the network for multi-object segmentation is one of our future targets. To improve robustness, we also aim to increase the size of training set by incorporating data from other institutions. In conclusion, we have demonstrated promising results of the proposed method to delineate the CTV in a large set of TRUS images, and the proposed ResU-net has the potential to find applications in other clinical tasks.

Acknowledgements. We would like to thank the Natural Sciences and Engineering Research Council of Canada, and the Canadian Institutes of Health Research for funding this project. The authors also would like to thank the physicians and staff at the Vancouver Cancer Centre who have contributed to this project.

References

1. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_49
2. Gibson, E., et al.: Deep residual networks for automatic segmentation of laparoscopic videos of the liver. In: SPIE Medical Imaging (2017)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR, pp. 770–778 (2016)

4. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceeding International Conference Multimedia, pp. 675–678. ACM (2014)
5. Mahdavi, S.S., Chng, N., Spadinger, I., Morris, W.J., Salcudean, S.E.: Semi-automatic segmentation for prostate interventions. MIA **15**(2), 226–237 (2011)
6. Morris, W., et al.: Population-based 10-year oncologic outcomes after low-dose-rate brachytherapy for low-risk and intermediate-risk prostate cancer. Cancer **119**(8), 1537–1546 (2013)
7. Nouranian, S., et al.: A multi-atlas-based segmentation framework for prostate brachytherapy. IEEE TMI **34**(4), 950–961 (2015)
8. Nouranian, S., et al.: Learning-based multi-label segmentation of transrectal ultrasound images for prostate brachytherapy. IEEE TMI **35**(3), 921–932 (2016)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). doi:10.1007/978-3-319-24574-4_28
10. Salembier, C., et al.: Tumour and target volumes in permanent prostate brachytherapy: a supplement to the ESTRO/EAU/EORTC recommendations on prostate brachytherapy. Radiother. Oncol. **83**(1), 3–10 (2007)
11. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)