Automatic Delineation of the Clinical Target Volume in Rectal Cancer for Radiation Therapy using Three-dimensional Fully Convolutional Neural Networks

Rasmus Larsson, Jun-Feng Xiong, Ying Song, Ling-Fu, Yi-Zhi Chen, Xu Xiaowei, Puming Zhang, Jun Zhao* *Member, IEEE*

***Abstract*— Accurate, robust, and fast delineation of the clinical target volume (CTV) for the use in radiotherapy of rectal cancer (RC) is highly sought-after. Convolutional neural networks (CNNs) have proven themselves very effective in various segmentation tasks on medical images. Despite this, their application in CTV delineation is not yet fully explored. This study uses the three-dimensional fully convolutional neural network architecture called V-net for CTV delineation. The West China Hospital (Chengdu, China) provided this study with 120 annotated CT scans. For improved performance and to battle data scarcity, the available scans were augmented. Trained on 100 CT-scans for 20 hours and tested on 20 previously unseen CT-scans the network achieved a mean dice similarity coefficient (DSC) of 0.90 and a mean delineation time per CTV of 0.60 seconds. The proposed method is compared with two other state-of-the-art CNNs and is shown to be superior.**

I. Introduction

An indispensable part of radiotherapy of rectal cancer (RC) is the delineation of the clinical target volume (CTV). Improper delineation of the CTV may lead to underdosage of areas that can contain cancer cells or pose risk of irradiation of neighboring organs [1]. In response, several guidelines have been released that contain boundaries, atlases, and recommendations for the CTV delineation [1]−[3]. However, high intra-observer variability between radiation oncologists is still an issue [4]. To address this problem and further improve precision and accuracy in the delineation of the CTV, automatic systems for delineation can be used. Automatic systems specifically for the delineation of the CTV for RC have been evaluated in [6]−[8]. The systems evaluated were the atlas-based auto segmentation system (ABAS, Elekta CMS Software) and SmartSegmentation-Knowledge-Based-Contouring software (SS-KBC, Varian Medical Systems). They are both based on an atlas containing many CT images used for registration to estimate the CTV [6]−[8]. While they are accurate and clinically approved, they are still time consuming to use. In addition, these systems both require manual adjustment, so the term automatic is a misnomer.

Recently, deep neural networks have demonstrated their strength in medical segmentation tasks [9]. Specifically, convolutional neural networks (CNNs) have so far been used in two principally different ways for segmentation. Firstly, the CNN can be trained on patches that are either part of the segmentation class or not. Each pixel of the image is later fed into the CNN and classified in a sliding window fashion. A problem with this approach is that patches may overlap, and the same convolutions are calculated many times. A variant of this approach was applied in [10] where organs-at-risk (OAR) in head and neck CT images were delineated for radiation therapy planning. The second approach is the use of fully convolutional neural networks (fCNNs) where pixel- (or voxel) wise segmentation is achieved in one forward pass of the full image through the network.

In [11] very good results were reached in the CTV delineation for prostate brachytherapy using residual neural networks. In addition to its being fully convolutional it introduced residual functions. The residual functions allow the gradient to flow more freely and combats the vanishing gradient problem many deep neural networks experience [12]. Also using fCNNs but with the use of a deep dilated convolutional neural network (DDCNN), [13] delineated the CTV and OARs for RC radiotherapy. A highly effective structure of the fCNN is the U-net [9], 14]. The U-net is based on two-dimensional convolutions on slices of volumetric medical data. It consists of two stages, one contracting (reduction in image size) and one expanding (restoration of the original image size). These two stages give the architecture a characteristic U-shape. The previously mentioned deep learning methods all utilizes two-dimensional convolutions which may miss three-dimensional context. In [15] the aptly called V-net extends the U-net with the use of three-dimensional convolutions, to capture this context.

We believe that the established automatic delineation systems for CTVs in RC can be substantially improved, in accuracy and speed, using CNNs. Moreover, harnessing 3D context can offer crucial data in the delineation of CTVs. In this work the V-net is therefore adopted. The novelty lies in the application of a 3D CNN for CTV delineation in RC. The result of our proposed method (the V-net based method) is also compared with both the U-net and the DDCNN.

This study is partly supported by National Key R&D Program of China (2016YFC0104608), the National Natural Science Foundation of China (No. 81371634), and Shanghai Jiao Tong University Medical Engineering Cross Research Funds (YG2017ZD10, and YG2014ZD05)

Rasmus Larsson, Jun-Feng Xiong, Ling Fu, Yi-Zhi Chen, Xu Xiaowei and Puming Zhang are with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, 200240

Ying Song is with the Division of Radiation Physics, State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, China.

*Jun Zhao is with the School of Biomedical Engineering, SJTU-UIH Institute for Medical Imaging Technology, and MED-X Research Institute, Shanghai Jiao Tong University, Shanghai, 200240, China (email: junzhao@sjtu.edu.cn).

II. Materials

Data used in this work were obtained from the West China Hospital (Chengdu, China). A total of 120 CT scans were obtained. The scans were accompanied by CTV delineations performed by an experienced oncologist. These are the ground truth delineations used for evaluation. Two sets were formed for the processing by the network: one training set consisting of 100 randomly chosen scans, and one test set with the remaining 20 scans. The institutional review board of the West China Hospital approved the protocol of this study.

III. Methods

A. Pre-processing

The scans were initially of different resolution and size. For consistent processing by the network, the scans were resampled to a common size of 144x144x48 with a resolution of 1x1x3 mm. Data scarcity is a big problem in many medical image segmentation tasks. Manually labelling medical data is a time consuming and laborious task hence there are usually not much data available. The V-net accentuates this problem as it treats each scan as one sample. This is opposed to the models in [10], [11], [13] that uses slices, where more samples can be obtained from each scan. To address this issue, data augmentation is used. The data have been augmented in three ways: flipping of the axes, translation, and scaling. Before each training epoch the training set is randomly augmented in the three previously mentioned ways. In this way, more samples are generated which reduces overfitting of the network and increases performance.

B. Network

In this work the original convolutional neural network structure of the V-net was adopted [15]. It is a fully-convolutional network capable of producing voxel-wise segmentation of medical data. The network can be divided into a contracting and expanding stage as can be seen in Fig. 1 below. For a more detailed view of the dimensions of the input see Table I. Firstly, the entire volume of size 144x144x48 is fed to the network. In the subsequent contracting stage, the volume size is reduced in four down sampling layers. The layers convolve the volume with filters of size 5x5x5 one to three times, each filter producing a unique feature map. These maps are then joined through element-wise summation and convolved with a filter of size 2x2x2 with stride 2 which halves the size in every dimension.

Traditionally, pooling operations are used for input size reduction but, inspired by [16], the V-net uses strided convolutions. The reason is that no switch mapping is required in the back propagation for deconvolutions compared to the un-pooling operations used for max pooling. As such, it requires less parameters in memory. Additionally, a residual function is learned. The involved residual skips the 5x5x5 kernel convolutions and is directly added to the element-wise summation. This lets the gradient flow more freely and alleviates the vanishing gradient problem mentioned in the introduction [12]. Afterwards, a PRelu function is used to add non-linearity.

Starting at 16 filters this number is doubled for every following layer. At the bottom layer the data consists of 256 feature maps with size 9x9x3. The reduction in size effectively increases the receptive field of each output neuron allowing them to "see" a larger part of the input. The next stage is the expanding one. It is nearly identical to the contracting one except that the 2x2x2 filters are used to de-convolve the volumes, projecting each voxel to a larger region. This results in a size doubling, conversely, the number of 5x5x5 filters at each expanding layer is halved. Another part of the network is that the feature maps from the contracting stage are forwarded to the expanding one via concatenation. At the very end, where the data once again have the size 144x144x48, a 1x1x1 filter is applied which produces two feature maps, for the fore- and background. These are then passed to a soft-max layer to generate the final probabilistic predictions.

An issue with feeding entire volumes to the network is class imbalance. Most of a volume consists of background voxels that are not part of the CTV. In [14], for the 2-dimensional case, the solution was to add weights to the cross-entropy loss function to suppress the negative classes. Another approach is to use an overlap metric which is done in this work using the V-net [15]. The network is trained by maximizing the dice similarity coefficient (DSC). The DSC for two binary volumes, prediction P and ground truth G, is defined as

$$DSC = \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2}, \quad (1)$$

where $p_i \in P$ and $g_i \in G$ for N-voxels. The DSC ranges from 0 (no overlap) to 1 (total overlap).

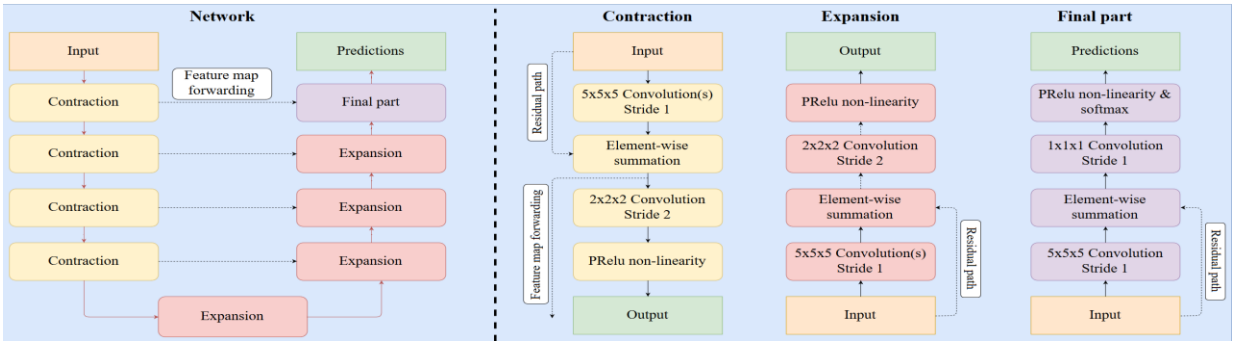

Figure 1. The network of the V-net in its entirety. The exact size of each feature map can be seen in Table I. The forwarded feature maps are concatenated on the input of the destination layer.

TABLE I. DETAILED STRUCTURE OF THE V-NET LAYERS

Layers	*Applied 5x5x5 convolutions*	*Input size*	*Output size*	*Feature maps*
Contracting part	1	144x144x48	72x72x24	16
Contracting part	2	72x72x24	36x36x12	32
Contracting part	3	36x36x12	18x18x6	64
Contracting part	3	18x18x6	9x9x3	128
Expanding part	3	9x9x3	18x18x6	256
Expanding part	3	18x18x9	36x36x12	256
Expanding part	3	36x36x12	72x72x24	128
Expanding part	2	72x72x24	144x144x48	64
Final part	1	144x144x48	144x144x48	32

IV. RESULTS

The network was implemented in python using TensorFlow [17]. It was trained using the gradient descent optimizer. Maximization of the DSC can be done by minimizing $1 - DSC$. The model had an initial learning rate of 0.01 that decayed exponentially every 3000 iterations with a base decay of 0.95. The filter weights were initialized using He initialization [18]. The network was trained for 700 epochs or 110K iterations during 20 hours on a workstation equipped with a NVIDIA GTX 1080 TI GPU. Due to memory limitations the network was trained with mini-batches containing only two volumes.

It took the model 0.60 seconds to automatically delineate the CTV for a previously unseen volume. For accuracy evaluation the DSC was used as defined in (1). The final binary segmentation is calculated by a threshold filter on the predicted foreground feature map. Voxels with probabilities > 0.5 are considered part of the CTV and the DSC is calculated on these voxels. The mean DSC of the 20 test volumes was 0.90. Qualitative results can be seen in Fig. 2

The proposed model is also compared with two other deep learning methods: The DDCNN [13] and the U-net [14]. To the best of our knowledge the DDCNN is the only other deep learning method applied to CTV delineation in RC for radiotherapy apart from ours. The U-net is a highly successful architecture in medical image analysis [9] that the V-net builds upon and is used as a baseline. The networks were implemented in Tensorflow according to their original papers [13], [14]. The tests were run on the same dataset with the same workstation. The U-net and DDCNN were fed 2D slices instead of whole volumes as per their design. The results can be seen in Table II.

V. DISCUSSION

Compared to the U-net and DDCNN, the V-net is both faster and more accurate. There are two main differences between the other methods and the V-net that may have influenced this. Firstly, the U-net and DDCNN are processing two-dimensional slices of the data and so may miss three-dimensional context that the V-net can pick up with its volumetric filters. Secondly, the V-net utilizes residual paths which have been described in [12] to alleviate the vanishing gradient problem and with that also increase performance. The U-net does contain these as opposed to the DDCNN, but only the longer kind that stretches over several layers. In this paper they are referred to as feature map forwards. The conclusion of [12] was that a combination of both short and long residual paths achieves the best performance. The combination is used in our proposed network which may be a reason for the superior result.

The mean DSC is calculated based on the final threshold filter of 0.5, however, we noticed the mean DSC can be improved by 0.01 if this filter is adaptively chosen for each prediction. This would, however, require manual intervention and make the network not fully automatic.

In literature, several studies have evaluated CTV delineation for RC radiotherapy used in clinics, mainly comparing manual and atlas-based automatic delineation. The median DSC for manual delineation by expert physicians was 0.84 in [6] and 0.75 for [7]. For automatic atlas-based systems it was 0.75 [6], 0.70 [7] and 0.85 [8]. The time costs for these methods ranged between 12 to 23 minutes [6]−[8]. The automatic systems were invariably faster than the expert physicians for delineation but still above 10 minutes. How our results compare to the manual experts and atlas-based systems is difficult to tell. Medical segmentation is heavily reliant on the dataset used and in this case the datasets differ. In future work we therefore want to arrange for both a second expert to manually delineate our test data as well as utilize an atlas-based system and see how they perform.

TABLE II. CTV DELINEATION METHOD COMPARISONS

Method	*Mean DSC*	*Mean delineation time*
U-net	0.84	7.0 seconds
DDCNN	0.87	16.0 seconds
Our proposed method	0.90	0.60 seconds

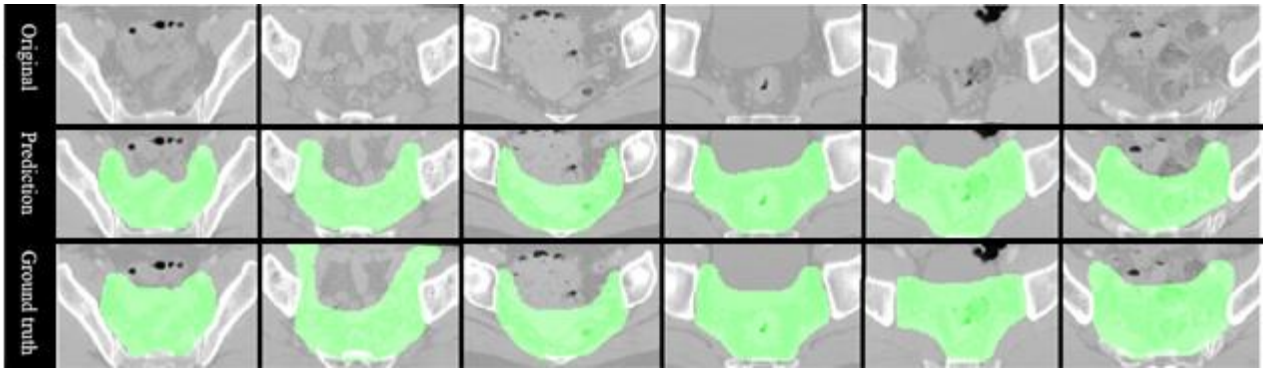

Figure 2. Qualitative results from the V-net: the top row shows axial slices of original volumes, the second row the V-net predictions and the last row the ground truth labels for the corresponding slices.

An important part of radiation therapy of RC is the delineation of OARs. These were largely ignored in this study focused solely on the CTV. An extension to this work would be to also delineate the OARs.

VI. Conclusion

So far, the delineation of the CTV in rectal cancer for radiation therapy has either been done manually by an expert or semi-automatically by atlas-based systems. In this paper deep learning-based methods are studied for the same application. The three-dimensional CNN based on the V-net outperformed the two-dimensional slice-based CNNs in both accuracy and speed.

References

[1] S. Roels, W. Duthoy, K. Haustermans, F. Penninckx, V. Vandecaveye, T. Boterberg and W. De Neve, "Definition and delineation of the clinical target volume for rectal cancer", *International Journal of Radiation Oncology*Biology*Physics*, vol. 65, no. 4, pp. 1129-1142, 2006.

[2] M. Ng, T. Leong, S. Chander, J. Chu, A. Kneebone, S. Carroll, K. Wiltshire, S. Ngan and L. Kachnic, "Australasian Gastrointestinal Trials Group (AGITG) Contouring Atlas and Planning Guidelines for Intensity-Modulated Radiotherapy in Anal Cancer", *International Journal of Radiation Oncology*Biology*Physics*, vol. 83, no. 5, pp. 1455-1462, 2012.

[3] R. Myerson, M. Garofalo, I. El Naqa, R. Abrams, A. Apte, W. Bosch, P. Das, L. Gunderson, T. Hong, J. Kim, C. Willett and L. Kachnic, "Elective Clinical Target Volumes for Conformal Therapy in Anorectal Cancer: A Radiation Therapy Oncology Group Consensus Panel Contouring Atlas", *International Journal of Radiation Oncology*Biology*Physics*, vol. 74, no. 3, pp. 824-830, 2009.

[4] V. Valentini, M. Gambacorta, B. Barbaro, G. Chiloiro, C. Coco, P. Das, F. Fanfani, I. Joye, L. Kachnic, P. Maingon, C. Marijnen, S. Ngan and K. Haustermans, "International consensus guidelines on Clinical Target Volume delineation in rectal cancer", *Radiotherapy and Oncology*, vol. 120, no. 2, pp. 195-201, 2016.

[5] M. La Macchia, F. Fellin, M. Amichetti, M. Cianchetti, S. Gianolini, V. Paola, A. Lomax and L. Widesott, "Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer", *Radiation Oncology*, vol. 7, no. 1, p. 160, 2012.

[6] M. Gambacorta, L. Boldrini, C. Valentini, N. Dinapoli, G. Mattiucci, G. Chiloiro, D. Pasini, S. Manfrida, N. Caria, B. Minsky and V. Valentini, "Automatic segmentation software in locally advanced rectal cancer: READY (REsearch program in Auto Delineation sYstem)-RECTAL 02: prospective study", Oncotarget, vol. 7, no. 27, 2016.

[7] M. Gambacorta, C. Valentini, N. Dinapoli, G. Mattiucci, D. Pasini, M. Barba, S. Manfrida, L. Boldrini, N. Caria and V. Valentini, "PO-0851 CLINICAL VALIDATION OF ATLAS-BASED AUTO-SEGMENTATION OF PELVIC VOLUMES AND NORMAL TISSUE IN RECTAL TUMORS", Radiotherapy and Oncology, vol. 103, pp. S332-S333, 2012.

[8] L. Anders, F. Stieler, K. Siebenlist, J. Schäfer, F. Lohr and F. Wenz, "Performance of an atlas-based autosegmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer", *Radiotherapy and Oncology*, vol. 102, no. 1, pp. 68-73, 2012.

[9] G. Litjens, T. Kooi, B. Bejnordi, A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken and C. Sánchez, "A survey on deep learning in medical image analysis", *Medical Image Analysis*, vol. 42, pp. 60-88, 2017.

[10] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks", *Medical Physics*, vol. 44, no. 2, pp. 547-557, 2017.

[11] E. Anas, S. Nouranian, S. Mahdavi, I. Spadinger, W. Morris, S. Salcudean, P. Mousavi and P. Abolmaesumi, "Clinical Target-Volume Delineation in Prostate Brachytherapy Using Residual Neural Networks", *Lecture Notes in Computer Science*, pp. 365-373, 2017.

[12] Drozdzal, Michal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury and Christopher Joseph Pal, *"The Importance of Skip Connections in Biomedical Image Segmentation."* LABELS/DLMIA@MICCAI, 2016.

[13] K. Men, J. Dai and Y. Li, "Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks", *Medical Physics*, vol. 44, no. 12, pp. 6377-6389, 2017.

[14] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", *Lecture Notes in Computer Science*, pp. 234-241, 2015.

[15] Milletari, F., Navab, N., Ahmadi, S.-A., "V-Net: fully convolutional neural net- works for volumetric medical image segmentation", arxiv:1606.04797, 2016b.

[16] Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M. "Striving for simplicity: The all convolutional net" *arXiv preprint arXiv:1412.6806.* 2014.

[17] Martín Abadi et al. "TensorFlow: Large-scale machine learning on heterogeneous systems", Software available from tensorflow.org, 2015.

[18] K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification", *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.