Published in final edited form as:

Conf Proc IEEE Int Conf Signal Image Process Appl. 2017 September ; 2017: 503–506. doi:10.1109/

ICSIIPA.2017.8120664.

Fully automated esophagus segmentation with a hierarchical deep learning approach

Roger Trullo^{†,*}, Caroline Petitjean[†], Dong Nie^{*}, Dinggang Shen^{*}, and Su Ruan[†]

[†]Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France

^{*}Department of Radiology and BRIC, UNC-Chapel Hill, USA

Abstract

Segmentation of organs at risk in CT volumes is a prerequisite for radiotherapy treatment planning. In this paper we focus on esophagus segmentation, a challenging problem since the walls of the esophagus have a very low contrast in CT images. Making use of Fully Convolutional Networks (FCN), we present several extensions that improve the performance, including a new architecture that allows to use low level features with high level information, effectively combining local and global information for improving the localization accuracy. Experiments demonstrate competitive performance on a dataset of 30 CT scans.

I. Introduction

Radiation therapy, alone or in combination with surgery or chemotherapy, is an essential element of treatment for tumors such as in lymphoma, or lung and esophageal cancer. Before radiotherapy, organs at risk located near the target tumor, such as the heart, the lungs, the esophagus, etc. must be outlined, in order to minimize the quantity of irradiation they receive. Today, segmentation of the esophagus is performed manually by clinicians despite some partial software support, such as thresholding for lung and bone structures. Manual delineation is a tedious task, prone to intra and inter-observer variabilities. Furthermore the time required to manually segment the image is the cause of delays potentially detrimental to the patient care (tumor progression) and prevents a fine adaptation to patient morphological changes (gain or loss of weight) and its tumor (volume or functional changes).

In particular, the esophagus is very challenging to delineate because of the low contrast of its boundaries in CT images, its complex and varying shape, and changing position, between patients, especially in case of mediastinal invasion by the tumor. (Fig. 1).

Hence the literature on (semi-) automatic esophagus segmentation is not abundant, with respect to other organs. Given the difficulty of the problem, semi-automated approaches have mainly been proposed. In [1], Rousson et al. proposed a two step segmentation method in cardiac CT: first the esophagus centerline is extracted using probabilistic spatial and appearance modeling, then the outer surface of the esophagus is extracted using multiple and coupled ellipse fittings. The method requires as input two points on the centerline of the esophagus and manual segmentation of the aorta and the left atrium. In [2], several organs are segmented using deformable models but the method fails to accurately contour the

esophagus. In [3], the manual drawing of one esophagus contour in an axial slice is propagated to other slices via registration, but no quantitative evaluation is proposed. In a similar way, the authors of [4] propose to interpolate missing contours in the frequency domain, from several contours in axial slices provided by the user. Fully automatic methods in [5], [6] consist in first finding the approximate shape using a detect and connect approach, and then a classifier is trained to find short segments of the esophagus which are approximated by an elliptical model. In [7] the approach makes use of a training set of manually contoured images to build a shape prior, which is incorporated into a graph cut framework. However the shape model was unidimensional and not fully leveraging the training set. Another strategy is to segment multiple organs at the same time. For example a recent work was presented in [8] where the authors proposed to combine multi-atlas deformable registration with a level set-based local search to segment several organs, i.e., aorta, esophagus, trachea, heart. The results for the esophagus were not good enough for clinical usage with Dice ratio (DR) as low as 0.01. Note also that in a recent survey presenting deep learning in medical imaging [9], no contribution on esophagus segmentation is reported.

Most of these methods require a significant amount of user input, which can hamper the results reproducibility. They also underline the difficulty of obtaining clinically viable results. As shown in Figure 1, the esophagus is a deformable organ whose contours may barely be visible, and whose shape might be hardly modeled with an ellipse. To overcome these issues, we propose a two step approach to segment the esophagus: a first rough segmentation is obtained by a deep learning architecture. As we have shown in a previous work [10], the results contain several false negatives, but the true positives can provide reliable information about the position of the organ and neighboring structures. Based on this information, we train a new model which is focused on a smaller region, still large enough to provide context information for the segmentation of the esophagus. We show experimentally that this fully automatic framework is able to outperform a state-of-the-art atlas based method and more complex deep learning architectures based on full slices.

II. Method

A. Overview

As mentioned above, the segmentation of the esophagus is a very challenging task and thus we believe that a two-stage approach is necessary to improve the performance. First, we propose to use a Sharpmask architecture [11] to perform multiorgan segmentation. These organs are the esophagus, heart, trachea and aorta. Then, in the second stage, we utilize the information learnt by the first network about the position and the probabilities of voxels of belonging to a specific organ to focus the attention of a second network on the esophagus.

B. FCN and SharpMask feature fusion architecture

Several works have shown that while in classification tasks deep networks that use pooling operations can have good results, the performance in semantic segmentation tasks can be affected due to loss of resolution [12]. A common strategy is to use features from early layers along with high level features from deep layers, which has two big advantages: first it

can improve the localization accuracy in the segmentation task, and second, it can help to alleviate the vanishing gradient problem since the errors from deep layers will be inserted in early layers. This has been exploited in the original FCN work [13], the U-Net [14], and more recently in SharpMask's (SM) Facebook work [11]. In our work we used an SM architecture (Fig. 2) with the difference that instead of using bilinear upsampling in the refinement module, we used transposed convolutions which gives more parameters and hence more capacity to the network. Different from [14], instead of copying the whole feature maps, convolution+ReLU operations are used to have the same number of channels as the deep level, avoiding the network to be biased due to big difference in the number of channels from early and deep layers.

C. Focus on esophagus

In the first network, all organs are segmented; however, as we have noticed from previous experiments [10], the esophagus results still includes several false negatives. We propose to use an additional SharpMask architecture which is dedicated to refine the segmentation of the esophagus. To do this, we use the fact that even if the first network has several misclassifications, it still can provide the position of the esophagus in the images reliably. Using this information we can crop a region of interest (ROI) which will include the esophagus, and some neighboring structures to provide context information. In particular, we find the centroid of the esophagus in each slice found by the first network, and compute the mean of these points. We only take into account slices where the first network outputted a connected segmentation with more than 10 voxels. Then we cropped the image around the mean center with a size of 128×128 , a power-of-2 size (because the max pooling layers of the network) integrating enough context surrounding the esophagus. For the z-axis, we find the minimum and maximum slice that hold the same restriction in term of number of connected voxels, and use all these slices in this range. We also use the probability map of the other organs (cropped in the same way) as this provides additional prior information about neighboring structures. This process is similar to the auto-context model (ACM) [15] which has been successfully applied on several segmentation tasks including brain segmentation on MRI [16]. Our proposed method is different in the sense that we use a deep learning architecture where the features are learnt by backpropagation instead of hand-crafted features as used in classical ACM. The framework is shown in Fig 3. Since classes are highly imbalanced (a lot more background voxels than esophagus voxels), typically a weighted cross entropy is used. However, the weight must be defined and this adds another hyper-parameter to the framework. We propose to use the Dice Loss function which has been introduced in the V-net paper [17]. This loss is more intuitive since it is directly related to the metric used in semantic segmentation problems, and it has the advantage of not needing any weight. It is defined as follows:

$$D_L = \frac{2 \sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \quad (1)$$

III. Experiments

We perform experiments on the standard SharpMask architecture and compare the results with our proposed architecture. We use rather large kernel size (7×7) for the filters, in accordance with other works in CT images [18]. Stochastic gradient descent (SGD) is used with a learning rate of 0.1 which was decreased by a factor of ten every 20 epochs and initialized by Xavier initialization [19] for all the weights in the networks.

A. Dataset and pre-processing

We evaluate our method on a dataset of 30 thoracic CT scans where the patients have either lung cancer or Hodgkin lymphoma, along with the manual delineations of the esophagus, heart, trachea, aorta and body contour. The latter is used in the network to avoid using background information during training. Scans have a resolution of $0.98 \times 0.98 \times 2.5$ mm for a size of $512 \times 512 \times (150 \sim 284)$ voxels. As pre-processing, each scan is normalized to have zero mean and unit variance. We perform 6-fold cross validation, resulting in 25 subjects for training and 5 for testing. The dataset is augmented applying a random affine transform and a randomly deformed version of each scan by using a deformation field obtained through a $2 \times 2 \times 2$ grid of control-points and B-spline interpolation [17].

B. Results

In Fig. 4 we show the segmentation result of the esophagus for a slice of a testing CT scan. Fig. 5 shows the volume rendering. We can see how Step 2 helps the segmentation reach accurate topology and smooths the surface. We compare our method with our own implementation of a patch based label fusion method called Optimized PatchMatch for Near Real Time and Accurate Label Fusion (OPAL) presented in [20]. In addition, we compare our method with the results in [10] where an SM network is used, in addition to Conditional Random Fields by using the CRFasRNN architecture [21]. We can see that our proposed framework outperformed the classical atlas based approach, and a more complex deep learning architecture that includes CRFasRNN and requires the fine-tuning of additional hyper-parameters like the standard deviations of the Gaussian filters used.

IV. Conclusions

We have presented a framework for the segmentation of the esophagus in CT images. The method uses data augmentation and a SharpMask architecture allowing an effective combination of low-level features with high-level semantic features. Dice scores obtained on a dataset of 30 patients are very encouraging and show valuable results for the clinic can be obtained, for this highly challenging applications. We are currently engaged in investigating new ways to incorporate spatial context, such as in auto-context models.

Acknowledgments

This work is co-financed by the European Union with the European regional development fund (ERDF, HN0002137) and by the Normandie Regional Council via the M2NUM project.

References

1. Rousson M, Bai Y, Xu C, Sauer F. Probabilistic minimal path for automated esophagus segmentation. *Proc. of SPIE Vol.* 2006; 6144:614449–1.
2. Ragan D, Starkschall G, McNutt T, Kaus M, Guerrero T, Stevens C. Semiautomated four-dimensional computed tomography segmentation using deformable models. *Medical physics.* 2005; 32:2254.
3. Huang T, Zhang G, Guerrero T, Starkschall G, Lin K, Forster K. Semi-automated ct segmentation using optic flow and fourier interpolation techniques. *Computer methods and programs in biomedicine.* 2006; 84(2–3):124–134. [PubMed: 17027116]
4. Fieselmann A, Lautenschläger S, Deinzer F, John M, Poppe B. Esophagus segmentation by spatially-constrained shape interpolation. *Bildverarbeitung für die Medizin 2008.* 2008:247–251.
5. Feulner J, Zhou S, Cavallaro A, Seifert S, Hornegger J, Comaniciu D. Fast automatic segmentation of the esophagus from 3d ct data using a probabilistic model. *Medical Image Computing and Computer- Assisted Intervention–MICCAI 2009.* 2009:255–262.
6. Feulner J, Zhou SK, Hammon M, Seifert S, Huber M, Comaniciu D, Hornegger J, Cavallaro A. A probabilistic model for automatic segmentation of the esophagus in 3-d ct scans. *IEEE transactions on medical imaging.* 2011; 30(6):1252–1264. [PubMed: 21303741]
7. Grosgeorge D, Petitjean C, Ruan S. Multilabel statistical shape prior for image segmentation. *IET Image Processing.* 2016; 10(10):710–716.
8. Schreiber E, Marcus D, Fox T. Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. *J App Clin Med Phys.* 2014; 15(4)
9. Litjens GJS, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. *CoRR.* 2017 vol. abs/1702.05747.
10. Trullo R, Petitjean C, Ruan S, Dubray B, Nie D, Shen D. Segmentation of organs at risk in thoracic CT images using a sharpmask architecture and conditional random fields. *ISBI.* 2017
11. Pinheiro PHO, Lin T, Collobert R, Dollár P. Learning to refine object segments. *CoRR.* 2016 vil. abs/1603.08695.
12. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Rodríguez JG. A review on deep learning techniques applied to semantic segmentation. *CoRR.* 2017 vol. abs/1704.06857.
13. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *CVPR.* 2015
14. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *MICCAI.* 2015:234–241. vol. 9351 of *LNCS*.
15. Tu Z, Bai X. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Trans on Pattern Anal Mach Intell.* 2010; 32(10):1744–1757.
16. Wang L, Gao Y, Shi F, Li G, Gilmore JH, Lin W, Shen D. Links: Learning-based multi-source integration framework for segmentation of infant brain images. *NeuroImage.* 2015; 108:160–172. [PubMed: 25541188]
17. Milletari F, Navab N, Ahmadi S. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR.* 2016 vol. abs/1606.04797.
18. Dou Q, Chen H, Jin Y, Yu L, Qin J, Heng P. 3d deeply supervised network for automatic liver segmentation from CT volumes. *CoRR.* 2016 vol. abs/1607.00582.
19. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *AISTATS.* 2010
20. Ta V-T, Giraud R, Collins DL, Coupé P. Optimized patchmatch for near real time and accurate label fusion. *MICCAI.* 2014:105–112.
21. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C, Torr P. Conditional random fields as recurrent neural networks. *ICCV.* 2015

Fig. 1. Two CT scans with manual delineation of the esophagus. Note how the esophagus is hardly distinguishable.

Fig. 2. Details of the SharpMask architecture. The numbers indicate the number of channels at each layer.

Fig. 3. Proposed architecture for the segmentation of the esophagus. Details of the SharpMask network are given in Fig. 2.

(a) After step 1

(b) After step 2

Fig. 4.

Example of segmentation results after step 1 (a) and step 2 (b). The green represents the ground truth, and the red is the output of the system.

Fig. 5. Volume rendering of the esophagus for two patients. Segmentation results after step 1 (a) and step 2 (b). The green represents the ground truth, and the red is the output of the system.

TABLE I

Comparison of mean Dice score $\pm$ stdev for the esophagus segmentation

Method	Dice score
PatchMatch [20]	0.39 ± 0.05
Sharpmask (Step 1)	0.66 ± 0.08
Sharpmask+CRF [10]	0.67 ± 0.04
Proposed fine-tuned model	0.72 ± 0.07

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript