Segmentation of Head and Neck Organs-At-Risk in Longitudinal CT Scans Combining Deformable Registrations and Convolutional Neural Networks

Liesbeth Vandewinckele[1,3], David Robben[1,3], Wouter Crijns[2,4], and Frederik Maes[1,3](✉)

[1] KU Leuven, Department of ESAT/PSI, Kasteelpark Arenberg 10 bus 2441, 3001 Leuven, Belgium
frederik.maes@kuleuven.be
[2] KU Leuven, Department of Oncology - Laboratory of Experimental Radiotherapy, Herestraat 49 bus 7003 40, 3000 Leuven, Belgium
[3] UZ Leuven, Medical Imaging Research Center, Herestraat 49 bus 7003, 3000 Leuven, Belgium
[4] UZ Leuven, Radiation Oncology, Herestraat 49 bus 7003 40, 3000 Leuven, Belgium

Abstract. Automated segmentation of organs-at-risk (OAR) in follow-up images of the patient acquired during the course of treatment could greatly facilitate adaptive treatment planning in radiotherapy. Instead of segmenting each image separately, the segmentation could be improved by making use of the additional information provided by longitudinal data of previously segmented images of the same patient. We propose a tool for automated segmentation of longitudinal data that combines deformable image registration (DIR) and convolutional neural networks (CNN). The segmentation propagated by DIR from a previous image onto the current image and the segmentation obtained by a separately trained cross-sectional CNN applied to the current image, are given as input to a longitudinal CNN, together with the images itself, that is trained to optimally predict the manual ground truth segmentation using all available information. Despite the fairly limited amount of training data available in this study, a significant improvement of the segmentations of four different OAR in head and neck CT scans was found compared to both the results of DIR and the cross-sectional CNN separately.

1 Introduction

Delineation of Organs-At-Risk (OAR) in a pre-treatment CT scan of the patient is an essential step in radiotherapy (RT) planning to be able to deliver the required dose to the target volume while at the same time minimizing the dose to the surrounding normal tissues in order to reduce the risk of complications. However, since the treatment is fractionated over multiple RT sessions during

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 146–154, 2018.
https://doi.org/10.1007/978-3-030-00889-5_17

the course of several weeks, anatomical changes may occur that invalidate the initial treatment plan. Hence, it can be useful to acquire a new CT scan during the course of treatment and adapt the treatment to the new anatomy if needed, which requires delineation of each of these longitudinal CT scans [1]. Manual segmentation by a clinical expert of OAR in the head and neck (H&N) region is time consuming and takes about 45 min up to two hours in clinical practice, since there are on average 13 3D structures to be delineated. Moreover, the manual delineations are prone to intra- and interobserver variations.

Automatic segmentation of OAR in longitudinal CT scans in the context of RT planning is usually solved by using deformable image registration (DIR) [2]. An already segmented image (a so called atlas) is deformed to fit the new image to be segmented and the delineations in the atlas are deformed in the same way to yield a segmentation of the new image. Several choices for the atlas can be made, but the best results are obtained with a previous CT scan from the same patient, as the similarity between the atlas and the new CT image to be segmented is then likely very high [2]. This strategy was applied by Zhang et al. [3], Veiga et al. [4] and Castadot et al. [5]. Unfortunately, manual adaptation may still be necessary, but the time needed for the adaptation is usually small compared to manual segmentation [6]. The purpose of this work is to replace this manual correction by a neural network that can do the needed corrections.

Convolutional neural networks (CNN) are currently the state-of-the-art neural network architectures for medical image segmentation. The segmentation is formulated as a voxel-wise classification problem, whereby each voxel is individually classified as belonging to a particular organ of interest based on the intensity values within a certain neighborhood (the receptive field). A CNN for segmentation of OAR in the H&N-region is proposed by Ibragimov et al. [7]. The network gives state-of-the-art results for organs with recognizable boundaries in CT-images. However, organs without recognizable boundaries are more difficult to segment, which suggests that additional information is required.

Longitudinal data are not frequently used yet in CNN based segmentation, although such data could provide relevant additional information. Examples of neural networks that incorporate longitudinal data are Birenbaum et al. [8], who used a CNN on longitudinal data for MS lesion segmentation, and Vivanti et al. [9], who proposed an algorithm for liver tumor segmentation in follow-up CT scans. Vivanti et al. [9] did not train their network on longitudinal data, but only used the previous scans to define a region of interest (ROI) to give as input to the neural network for segmenting the tumor. The benefit of defining a ROI is that the amount of false positives will be reduced. In contrast to Vivanti et al. [9], we propose to include the previous segmentations registered to the new CT scan as additional features for CNN-based classification.

2 Methods

2.1 Available Data and Preprocessing

The dataset consists of 17 sets of longitudinal H&N data. Each such set consists of five types of images:

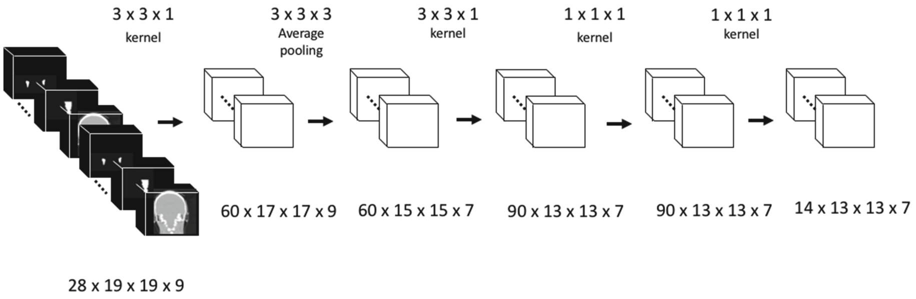

Fig. 1. The proposed CNN architecture for segmentation using longitudinal data. The size of each layer is given by # feature maps × 3D segment size.

- I_0: a previous CT scan acquired before or in week two of the RT treatment;
- I_1: the current CT scan acquired in week two or four of the RT treatment;
- $S_{0,m}$: the clinically approved binary segmentation maps of OAR in I_0;
- $S_{1,m}$: the clinically approved binary segmentation maps of OAR in I_1;
- $S_{1,c}$: the automatically generated binary segmentation maps of OAR in I_1 using the state-of-the-art cross-sectional CNN defined in [10], trained on a separate non-longitudinal dataset (acquired on the same scanner and delineated by the same observer as the longitudinal dataset).

The 17 sets of longitudinal data originate from 9 different patients. All CT scans were acquired in our institute on the same Siemens Sensation Open CT scanner using the same clinical protocol at 120kV. Clinically approved OAR segmentations are available for 13 H&N structures: the brainstem, the cochlea (left and right), the upper esophagus, the glottic area, the mandible, the extended oral cavity, the parotid glands (left and right), the pharyngeal constrictor muscles (PCM inferior, medial and superior), the spinal cord, the submandibular glands (left and right) and the supraglottic larynx. All images are preprocessed to have the same voxel size of $1 \times 1 \times 3$ mm^3 and the intensities of the CT scans are normalized to have zero mean and unit variance over all CT scans together.

2.2 Deformable Image Registration

The first step is to align the previous image I_0 and its segmentation $S_{0,m}$ onto the current image I_1 using DIR, yielding the deformed image $I_{0,r}$ and a DIR-based segmentation $S_{0,r}$ of the OAR in I_1. The registration process consists of two steps: first rigid and then non-rigid B-spline registration. The hyperparameters for each step are optimized in terms of a volume-weighted average of the Dice similarity coefficients (DSC) of $S_{0,r}$ compared to $S_{1,m}$ over all OAR. The most important DIR hyperparameters are the similarity metric used, the number of histogram bins in case mutual information is used, and the final spacing of the B-spline control point grid. Optimal performance was obtained with mutual information as similarity metric, with 64 bins for rigid and 32 bins for non-rigid registration, and a final B-spline grid spacing of 16 mm. All registrations were performed using Elastix [11].

2.3 Neural Network Architecture

The registered longitudinal images $I_{0,r}$ and I_1 and both segmentations $S_{0,r}$ and $S_{1,c}$ are given as input to the neural network, which generates the segmentation $S_{1,l}$ as a prediction of the true segmentation $S_{1,m}$ for image I_1. The longitudinal neural network is built by taking into account two different considerations. Firstly, a certain size of the receptive field is required. Secondly, the amount of parameters must be kept as low as possible to reduce overfitting since the amount of training data is small. The network has four convolutional layers. The first two layers are the feature extraction part with a kernel size of $(3, 3, 1)$ and a stride of 1 and the last two layers are fully connected layers implemented as convolutional layers with a kernel size of $(1, 1, 1)$ to make the network fully convolutional. A scheme of the architecture can be found in Fig. 1. An average pooling layer is inserted since this increases the receptive field without increasing the amount of parameters. It has a pooling size of $(3, 3, 3)$ with a stride of 1. The amount of feature maps cannot be made too low since we expect a lot of interactions between the inputs (and OAR) and the amount of redundant information is not high. The amount of feature maps is set to 60 in the first layer, 90 in the subsequent layers and 14 in the output layer, one for each class (13 OAR and background). The size of the receptive field thus becomes $7 \times 7 \times 3$ voxels or $7 \times 7 \times 9$ mm^3, which is small, implying that the neural network has not much contextual information to base its predictions on. It only makes uses of the intensities of both images and the available segmentations within a small neighborhood around each voxel.

2.4 Neural Network Training

The neural network is trained in a supervised way using the training scheme from Kamnitsas et al. [12], which was implemented by [13]. The training scheme does fully-convolutional predictions on image segments, since the memory requirements for full 3D images and 3D networks are high. In this way, several consecutive segments must be given as input to the network to obtain a segmentation of the complete image. The used evaluation metric is categorical cross-entropy. To prevent class imbalance, the image segments are sampled from the training images with an equal probability to be centered at a voxel of any of the different classes. The Adam optimizer is used with the originally proposed parameters [14]. The initial learning rate is set to 0.008 and is divided by four when a convergence plateau of the cost function is reached. This is done two times. The weights are initialized using He's initialization and PReLU activation functions are used in the hidden layers. Furthermore, batch normalization is applied to all hidden layers. A softmax function is used at the output layer. As the amount of data available to train the network was low, regularization is quite important in this work. Dropout is used in the last layers of the network with a dropout probability of 0.5. The weight against L_2-regularization is equal to 0.001. Data augmentation is done on the samples by flipping them around the sagittal plane.

2.5 Postprocessing

Postprocessing is a standard approach in literature to improve the resulting segmentations of the neural network. Voxels are classified individually to belong to the object of interest or not, without explicitly considering connectivity constraints. Postprocessing can be used to impose such constraints, which causes single pixels or holes to be removed [7,8,12,15]. However, in this work, no postprocessing is used in order to be able to evaluate the intrinsic segmentation performance of the network itself.

3 Results and Discussion

DIR took on average 15 min per dataset on an Intel Xeon E5645. After registration, the segmentation of the OAR by the longitudinal CNN took on average 2–3 min on a Nvidia GTX 1080 Ti.

A 6-fold cross-validation is performed on the longitudinal dataset to obtain segmentations for all patients with the longitudinal CNN. The results of the three segmentation approaches $S_{0,r}$ (DIR), $S_{1,c}$ (cross-sectional CNN), and $S_{1,l}$ (the proposed longitudinal CNN) are summarized in Table 1 by their average DSC compared to the manual ground truth segmentation $S_{1,m}$. Statistical significance between different approaches based on differences in DSC is assessed with a one-sided, paired Wilcoxon signed-rank test with a significance level of 0.05.

DIR ($S_{0,r}$) performs better than the cross-sectional CNN of [10] ($S_{1,c}$) in terms of DSC for five different organs (brainstem, upper esophagus, oral cavity, parotid glands and spinal cord), while the opposite is true for the mandible, which is a bony structure that is clearly defined on a CT scan.

The longitudinal CNN ($S_{1,l}$) performs at least as good as its both input segmentations (except for the spinal cord). It performs better than the cross-sectional CNN for 7 structures and better than DIR for 5 structures, including also the mandible. Moreover, the longitudinal CNN improves the results of both input segmentations for 4 structures: the oral cavity, the parotid glands, the submandibular glands and the supraglottic larynx. Hence, the longitudinal CNN not just selects the best of both segmentations, but succeeds at improving segmentation quality by combining the results of both inputs. An exception is the segmentation of the spinal cord. This can be explained by an inconsistency in the lower border of the spinal cord in the training data for the cross-sectional CNN of [10] and for the longitudinal CNN, which makes it impossible for the longitudinal CNN to learn a consensus.

Table 1. DSC (mean ± SD) for OAR segmentation in image I_1 based on DIR of the previous image I_0 onto the current image I_1 ($S_{0,r}$), the cross-sectional CNN of [10] applied to I_1 ($S_{1,c}$) and the proposed longitudinal CNN ($S_{1,l}$) w.r.t. the manual expert segmentation of I_1 ($S_{1,m}$), averaged over all performed predictions on N datasets. Statistical significant results are indicated by ($> r, c, l$) if the result is better than $S_{0,r}$, $S_{1,c}$ or $S_{1,l}$ respectively. Significance was assessed using a one-sided, paired Wilcoxon signed-rank test ($\alpha = 0.05$).

OAR	N	$S_{0,r}$	$S_{1,c}$	$S_{1,l}$
Brainstem	17	0.88 ± 0.01 ($> c$)	0.84 ± 0.03	0.88 ± 0.02 ($> c$)
Cochlea	5	0.60 ± 0.11	0.55 ± 0.12	0.67 ± 0.09
Upper Esophagus	16	0.64 ± 0.13 ($> c$)	0.58 ± 0.12	0.62 ± 0.12 ($> c$)
Glottic Area	15	0.57 ± 0.18	0.58 ± 0.22	0.56 ± 0.23
Mandible	17	0.87 ± 0.02	0.91 ± 0.02 ($> r$)	0.91 ± 0.02 ($> r$)
Oral Cavity	15	0.88 ± 0.02 ($> c$)	0.87 ± 0.04	0.89 ± 0.02 ($> r, c$)
Parotid Glands	17	0.82 ± 0.03 ($> c$)	0.79 ± 0.06	0.84 ± 0.04 ($> r, c$)
PCM inferior	13	0.59 ± 0.10	0.53 ± 0.18	0.51 ± 0.24
PCM medial	15	0.48 ± 0.16	0.52 ± 0.17	0.50 ± 0.12
PCM superior	10	0.42 ± 0.12	0.38 ± 0.08	0.45 ± 0.11
Spinal Cord	17	0.75 ± 0.10 ($> c, l$)	0.73 ± 0.10	0.73 ± 0.09 ($> c$)
Submandibular Glands	17	0.75 ± 0.05	0.71 ± 0.11	0.78 ± 0.09 ($> r, c$)
Supraglottic Larynx	13	0.71 ± 0.08	0.64 ± 0.13	0.76 ± 0.07 ($> r, c$)

Some example delineations are shown in Fig. 2. We observed that the delineations obtained with the longitudinal CNN mostly lie between the delineations obtained with DIR and the cross-sectional CNN, unless a clear boundary can be perceived in the CT scan. The longitudinal CNN can thus improve the input segmentations if one systematically constitutes an oversegmentation and the other an undersegmentation. This appears to be the case for the parotid glands segmentations. Another possibility is that inaccuracies in both input segmentations occur at different positions in the object. An example are the submandibular glands, for which the cross-sectional CNN performs well for segmenting the upper part, while DIR performs well for the lower part. At the moment, little can be concluded about the other organs, for which segmentation performance is not significantly improved by the longitudinal CNN. Since the receptive field of the proposed longitudinal CNN is limited, it has only limited ability to differentiate between different positions in the structures to be segmented, and therefore has only limited ability to adapt its prediction depending on the position. Improvements can occur if the longitudinal CNN would be able to recognize typical errors of both types of input segmentations at different positions in the organ. Therefore, extra hidden layers or additional pathways should be added to the network. Since this increases the amount of parameters, extra training data would be required.

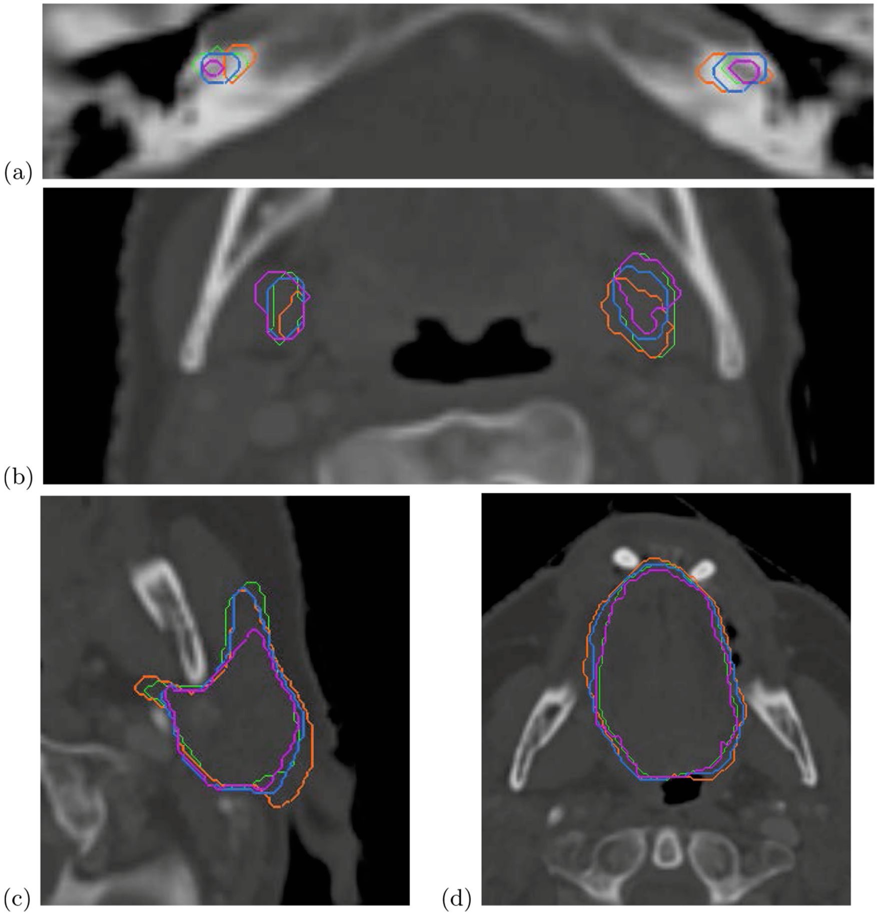

Fig. 2. Examples of OAR segmentations obtained with DIR ($S_{0,r}$, orange), the cross-sectional CNN ($S_{1,c}$, purple), and the longitudinal CNN ($S_{1,l}$, blue), compared to the manual ground truth segmentations ($S_{1,m}$, green), for: (a) cochlea; (b) submandibular glands; (c) right parotid gland; (d) oral cavity. (Color figure online)

4 Conclusion

We propose a manner to combine two different segmentation methods for OAR in H&N CT scans: longitudinal DIR and a CNN trained on cross-sectional data. Both techniques base their predictions on a different type of information: longitudinal data similarity for DIR versus learned intensity features for CNN. Combining both methods using the proposed longitudinal CCN effectively combines both sources of information. This hybrid approach was shown not only to be able to choose the best segmentation obtained with both methods, but also to improve the segmentation performance as achieved with either method separately.

References

1. Castadot, P., Lee, J., Geets, X., Gregoire, V.: Adaptive radiotherapy for head and neck cancer. Semin. Radiat. Oncol. **20**(2), 84–93 (2010)
2. Han, X., et al.: Atlas-based auto-segmentation of head and neck CT images. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008. LNCS, vol. 5242, pp. 434–441. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85990-1_52
3. Zhang, T., Chi, Y., Meldolesi, E., Yan, D.: Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy. Int. J. Radiat. Oncol. Biol. Phys. **68**(2), 522–530 (2007)
4. Veiga, C., McClelland, J., Ricketts, K., D'Souza, D., Royle, G.: Deformable registrations for head and neck cancer adaptive radiotherapy. In: Proceedings of the First MICCAI Workshop on Image-Guidance and Multimodal Dose Planning in Radiation Therapy, pp. 66–73 (2012)
5. Castadot, P., Lee, J., Parraga, A., Geets, X., Macq, B., Gregoire, V.: Comparison of 12 deformable registration strategies in adaptive radiation therapy for the treatment of head and neck tumors. Radiother. Oncol. **89**(1), 1–12 (2008)
6. Daisne, J.F., Blumhofer, A.: Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. Radiat. Oncol. **8**, 154 (2013)
7. Ibragimov, B., Xing, L.: Segmentation of organs-at-risk in head and neck CT images using convolutional neural networks. Med. Phys. **44**(2), 547–557 (2017)
8. Birenbaum, A., Greenspan, H.: Multi-view longitudinal CNN for multiple sclerosis lesion segmentation. Eng. Appl. Artif. Intell. **65**, 111–118 (2017)
9. Vivanti, R., Ephrat, A., Joskowicz, L., Lev-Cohain, N., Karaaslan, O.A., Sosna, J.: Automatic liver tumor segmentation in follow-up CT scans: preliminary method and results. In: Wu, G., Coupé, P., Zhan, Y., Munsell, B., Rueckert, D. (eds.) Patch-MI 2015. LNCS, vol. 9467, pp. 54–61. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-28194-0_7
10. La Greca Saint-Esteven, A.: Deep convolutional neural networks for automated segmentation of organs-at-risk in radiotherapy. Master's thesis, KU Leuven (2018)
11. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W.: elastix: a toolbox for intensity based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2010)
12. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)

13. Robben, D., Bertels, J., Willems, S., Vandermeulen, D., Maes, F., Suetens, P.: DeepVoxNet: voxel-wise prediction for 3D images. Technical report: KUL/ESAT/PSI/1801 (2018)
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR) (2014)
15. Hu, P., Wu, F., Peng, J., Bao, Y., Chen, F., Kong, D.: Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. Int. J. Comput. Assist. Radiol. Surg. **12**(3), 399–411 (2017)