Vertebrae Localization and Segmentation with SpatialConfiguration-Net and U-Net

Christian Payer[1,2][0000−0002−5558−9495], Darko Štern[1,2][0000−0003−3449−5497], Horst Bischof[1][0000−0002−9096−6671], and Martin Urschler[2,3][0000−0001−5792−3971]

[1]Institute of Computer Graphics and Vision, Graz University of Technology, Austria
[2]Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria
[3]University of Auckland, New Zealand

1 Introduction

This technical report introduces our proposed pipeline for fully automatic vertebrae localization and segmentation in CT volumes for the VerSe 2019 Large Scale Vertebrae Segmentation Challenge. The challenge consists of two tasks, where the first one is to localize and label the centers of the individual vertebrae, and the second one is vertebrae segmentation. For more details of the dataset and creation of the annotations, visit the homepage of the challenge and see the respective publications [1, 4].

2 Method

We perform vertebrae localization and segmentation in a three-step approach. Firstly, due to the large variation of the field of view of the input CT volumes, a CNN with a coarse input resolution predicts the approximate location of the spine. Secondly, another CNN in higher resolution performs multiple landmark localization and identification of the individual vertebra centroids. Lastly, the segmentation CNN performs a binary segmentation of each located vertebra. The results of the individually segmented vertebrae are merged into the final multi-label segmentation.

2.1 Spine Localization

For localizing the approximate position of the spine, we use a variant of the U-Net [3] to regress a heatmap of the spinal centerline, i.e. the line passing through vertebral centroids, with an L2-loss [2]. The heatmap of the spinal centerline is generated by combining Gaussian heatmaps of all individual landmarks (see Fig. 1). The input image is resampled to a uniform voxel spacing of 8 mm and centered at the network input. Since the network input resolution is $[64 \times 64 \times 128]$, every volume of the dataset fits into the network.

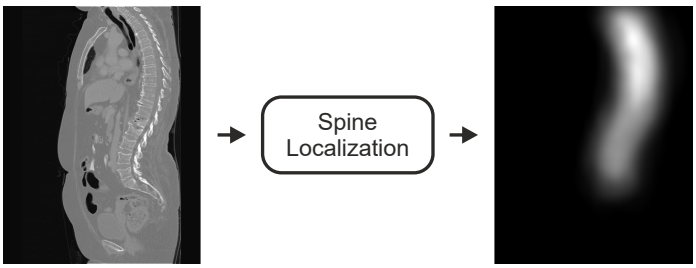

Fig. 1: Input volume and spine heatmap prediction of the spine localization network.

2.2 Vertebrae Localization

To localizes centers of the vertebral bodies, we use the SpatialConfiguration-Net proposed in [2]. The network effectively combines the *local appearance* of landmarks with their *spatial configuration*. The *local appearance* part of the network is based on the U-Net, while the *spatial configuration* part consists of four convolutions with $[7 \times 7 \times 7]$ kernels in a row and is processed in 1/4 of the resolution of the *local appearance* part. For more details of the network architecture and loss function, we refer the reader to [2].

A schematic representation of how the input volumes are processed to predict the final heatmaps is shown in Fig. 2. Every input volume is resampled to have a uniform voxel spacing of 2 mm, while the network is set up for inputs of size $[96 \times 96 \times 128]$. With these volume size and spacing, many images of the dataset do not fit into the network and cannot be processed at once. To narrow the region of interest in the vertebral localization step, we used the predicted of the spine localization network, see Sec. 2.1. Furthermore, as some volumes have a larger extent in the z-axis (i.e., the axis perpendicular to the axial plane) that would not fit into the network, we process such volumes as follows: During training, we

Fig. 2: Input volume and individual heatmap predictions of the vertebrae localization network. The yellow rectangle indicates that not the whole input volume is processed at once, but overlapping cropped sub-volumes. For each possible landmark, an separate heatmap volume is predicted, which is visualized with different colors.

crop a subvolume at a random position at the z-axis. During inference, we split the volumes at the z-axis into multiple subvolumes that overlap for 96 pixels, and process them one after another. Then, we merge the network predictions of the overlapping subvolumes by taking the maximum response over all predictions, for more details please check [2].

We detect the final landmark positions as follows: For each predicted heatmap volume, we detect multiple local heatmap maxima that are above a certain threshold. Then, we determine the first and last vertebrae that are visible on the volume by taking the heatmap with the largest value that is closest to the volume top or bottom, respectively. We identify the final predicted landmark sequence by taking the sequence that does not violate following conditions: consecutive vertebrae may not be closer than 12.5 mm and farther away than 50 mm, as well as a following landmark may not be above a previous one.

2.3 Vertebrae Segmentation

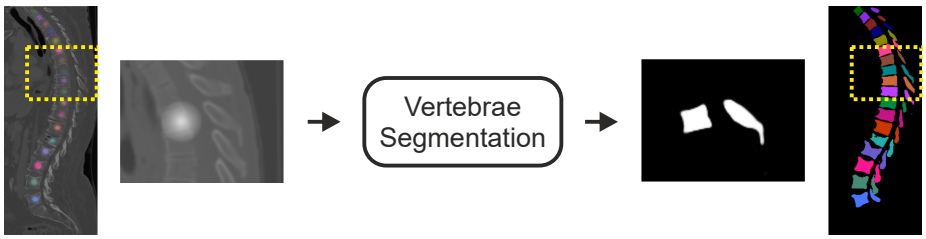

Fig. 3: Input volume and segmented vertebrae of the spine segmentation network. The yellow rectangle shows the cropped region around a single vertebrae and indicates that each localized vertebrae is processed individually. Each individually detected vertebra is then merged back to the final mulit-label segmentation.

For creating the final vertebrae segmentation, we use a U-Net [3] to segment each localized vertebra (see Fig. 3). The U-Net is set up with a sigmoid cross-entropy loss for binary segmentation to separate individual vertebrae from the background. Since each vertebra is segmented independently, the network needs to know, which vertebra it should segment. Thus, from the whole spine image we crop the region around the localized centroid (see Sec. 2.2), such that the vertebra is in the center of the image. Furthermore, in the same way as the vertebral image, we also cropped a heatmap image of vertebral centroid from the heatmap prediction of the vertebral localization network. Both cropped vertebral image and heatmap image of vertebral centroid are used as an input for the segmentation network. Both input volumes are resampled to have a uniform voxel spacing of 1 mm, while the network is set up for inputs of size $[128 \times 128 \times 96]$.

To create the final multi-label segmentation result, the individual predictions of the cropped inputs are resampled back to the original input resolution and translated back to the original position.

3 Implementation Details

Training and testing of the network was done in Tensorflow[1], while we perform on-the-fly data augmentation using SimpleITK[2]. As data augmentations we use intensity shift and scale, as well as spatial translation, scaling, rotation and elastic deformation. We evaluate training and network hyperparameters with a three-fold cross validation. The results submitted to the challenge were generated with networks that were trained with all 80 annotated training volumes from the VerSe 2019 challenge.

4 Conclusion

In this technical report we have proposed a three step fully automatic approach for vertebrae localization and segmentation. The predicted localizations and segmentations submitted to the VerSe 2019 await comparison to other participating methods.

References

1. Glocker, B., Feulner, J., Criminisi, A., Haynor, D.R., Konukoglu, E.: Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. In: Proc. Med. Image Comput. Comput. Interv. vol. 15, pp. 590–8 (2012)
2. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating Spatial Configuration into Heatmap Regression Based CNNs for Landmark Localization. Med. Image Anal. 54, 207–219 (may 2019)
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Proc. Med. Image Comput. Comput. Interv., pp. 234–241. Springer (2015)
4. Sekuboyina, A., Rempfler, M., Kukačka, J., Tetteh, G., Valentinitsch, A., Kirschke, J.S., Menze, B.H.: Btrfly Net: Vertebrae Labelling with Energy-Based Adversarial Learning of Local Spine Prior. Proc. Med. Image Comput. Comput. Interv. pp. 649–657 (2018)

[1] https://www.tensorflow.org/
[2] http://www.simpleitk.org/