Iterative convolutional neural networks for automatic vertebra identification and segmentation in CT images

Nikolas Lessmann[a], Bram van Ginneken[b] and Ivana Išgum[a]

[a]Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands
[b]Diagnostic Image Analysis Group, Radboud University Medical Center Nijmegen, The Netherlands

ABSTRACT

Segmentation and identification of the vertebrae in CT images are important steps for automatic analysis of the spine. This paper presents an automatic method based on iterative convolutional neural networks. These utilize the inherent order of the vertebral column to simplify the detection problem, so that the network can be trained with as little as ten manual reference segmentations. Vertebrae are segmented and identified one-by-one in sequential order, using an iterative procedure. Vertebrae are first roughly localized and identified in low-resolution images that enable the analysis of context information, and afterwards reanalyzed in the original high-resolution images to obtain a fine segmentation.

The method was trained and evaluated with 15 spine CT scans from the MICCAI CSI 2014 workshop challenge. These scans cover the whole thoracic and lumbar part of the spine of healthy young adults. In contrast to a non-iterative convolutional neural network, which made labeling mistakes, the proposed iterative method correctly identified all vertebrae. Our method achieved a mean Dice coefficient of 0.948 and a mean surface distance of 0.29 mm and thus outperforms the best method that participated in the original challenge.

Keywords: spine segmentation, vertebra identification, deep learning, convolutional neural networks

1. PURPOSE

Precise segmentation of the spine and identification of the individual vertebrae are prerequisites for computerized analysis of the spine in CT images. Recent methods in the literature have been primarily model-based, with some of these methods relying on machine learning to guide the model.[1–6] However, machine learning, especially convolutional neural networks (CNNs), has recently been successful in many segmentation tasks without explicitly modeling the shape or appearance of the structure of interest.[7,8] CNNs have been also used for automatic identification of vertebrae in CT. Sekuboyina et al.[6] used a multi-label 2D U-net[9] to label and segment the five lumbar vertebrae. While their CNN was able to distinguish the lumbar vertebrae, the high similarity in appearance of the thoracic vertebrae motivated other authors to combine CNNs with prior structural knowledge. Chen et al.[2] classified vertebra candidates with a CNN that incorporates a loss term for neighboring vertebrae to enforce a structural prior. Yang et al.[10] used a modified 3D U-net[11] to obtain probability maps for the individual vertebrae and applied a message passing scheme to correct mistakes and ensure a plausible order of the vertebrae. In this paper, we propose an alternative strategy based on a 3D CNN that iteratively localizes (*i.e.* finds in the image), identifies (*i.e.* assigns an anatomical label) and segments (*i.e.* marks all voxels in the image that are part of the respective vertebra) the lumbar and thoracic vertebrae in CT scans. To ensure consistent labeling of the vertebrae, the CNN is recursively informed about already detected vertebrae.

2. DATA DESCRIPTION

This study analyzed the publicly available data from the spine segmentation challenge originally held at the Computational Spine Imaging (CSI) workshop at MICCAI 2014.[12] This data set consists of 15 spine CT scans of healthy young adults, aged 20–34 years, who were scanned with either a Philips iCT 256 slice CT scanner or a Siemens Sensation 64 slice CT scanner (120 kVp, with IV-contrast). The in-plane resolution ranges from 0.31 mm to 0.36 mm and the slice thickness is either 0.7 mm or 1.0 mm. Each scan covers the entire thoracic and

Send correspondence to N. Lessmann (email: n.lessmann@umcutrecht.nl)

Medical Imaging 2018: Image Processing, edited by Elsa D. Angelini, Bennett A. Landman, Proc. of SPIE
Vol. 10574, 1057408 · © 2018 SPIE · CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2292731

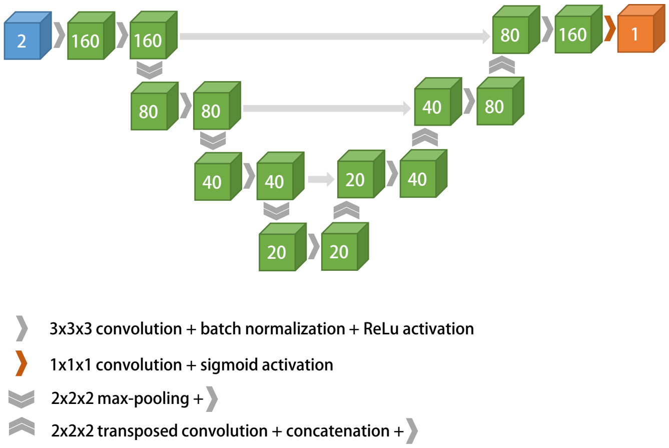
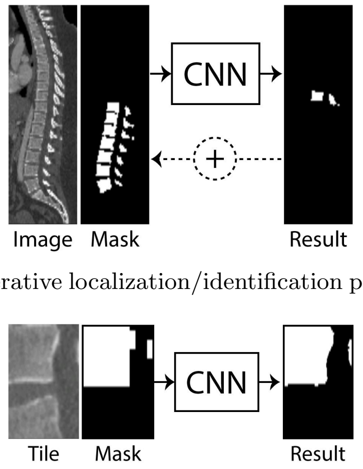

2
160
160
80
160
1
80
80
40
80
40
40
20
40
20
20
3x3x3 convolution + batch normalization + ReLu activation
1x1x1 convolution + sigmoid activation
2x2x2 max-pooling +
2x2x2 transposed convolution + concatenation +

Figure 1: Architecture of the CNNs. A 3D U-net[11] design with fewer filters per layer, which are indicated by the numbers in the green cubes. The network receives a 2-channel 3D input volume (image and mask) and produces a binary 1-channel 3D output volume.



(a) Iterative localization/identification process

(b) Segmentation refinement step

Figure 2: Illustration of the two-step segmentation process.

lumbar section of the spine. The challenge organizers provided reference segmentations for all scans. During development, we used eight of the ten training scans for training and the remaining two scans to evaluate and optimize the method. After finalizing the method, we retrained the two networks using all ten training scans and evaluated the performance on the five scans that were used as the unseen test set in the original challenge.

3. METHOD

Localization and identification of the individual vertebrae requires information about the spatial context while precise segmentation of each vertebra requires detailed local image information. The proposed method therefore consists of two steps. In the first step, a CNN repeatedly analyzes a downsampled image to coarsely localize and identify the vertebrae (in the following referred to as the localization and identification network). In each iteration, all voxels of a single vertebra are identified, and the step is repeated until all vertebrae are identified. The iterative vertebra-by-vertebra identification allows the network to learn the more general task of detecting a single vertebra rather than learning to recognize and distinguish all vertebrae simultaneously. With this strategy, the CNN can be trained using a low number of manual reference segmentations. In the second step, another CNN with the same architecture performs a fine segmentation of the vertebra at the original resolution (in the following referred to as segmentation network). This CNN classifies all voxels in a region of interest that was inferred by the analysis of the low-resolution image in the first step.

The architecture of the CNNs is based on the architecture of 3D U-net,[11] but uses fewer filters per layer to avoid overfitting as little training data was available (Figure 1). For localization and identification of the vertebrae, the image is downsampled by a factor eight in-plane and a factor four along the z-axis, and cropped or padded to standard dimensions of $64 \times 64 \times 128$ voxels. These volumes are fed to the localization and identification network to detect voxels that are part of the N-th vertebra relative to a defined reference vertebra. Given that only a single vertebra is localized and identified at a time, the network needs to keep track of already detected vertebrae. Therefore, as additional input, the network receives a binary mask containing the $N-1$ previously detected vertebrae. We used the last lumbar vertebra L5 as the reference vertebra for these scans because the sacrum is at least partially visible and L5 is therefore relatively easy to recognize. Given a CT image volume and an empty mask, the network is trained to identify L5. Given the same image and a binary mask containing L5, the network is trained to identify L4 in the image, which is added to the mask and fed again into the network to identify L3. This process is repeated until all lumbar and thoracic vertebrae are identified (Figure 2a). The network output in each iteration is binary because the iterative process starts with a known

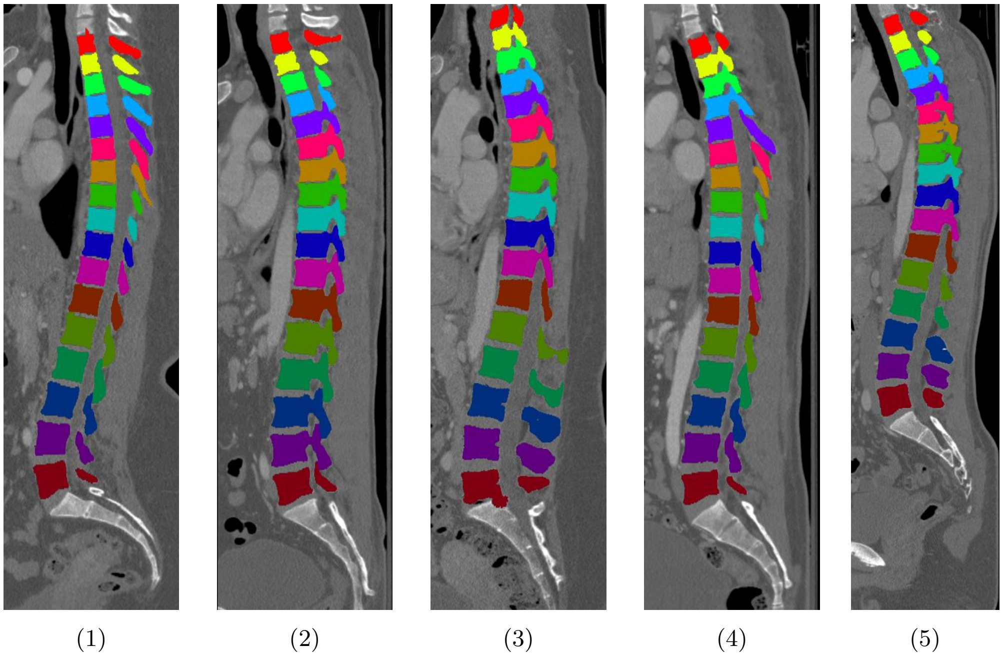


Figure 3: Examples of segmentations obtained with the proposed method. These correspond to test cases 1–5 of the challenge data set.[12] Despite an overall accurate segmentation, minor inaccuracies can be observed, e.g., in case 3 part of the sacrum is segmented.

reference vertebra and the vertebral label in each following iteration can then be deduced from the order of the vertebral column.

Downsampled images are sufficient for localization and identification of the vertebrae and have the benefit of allowing the whole scan to be analyzed by the CNN. However, precise segmentation of each detected vertebra requires the original high resolution image. To segment each vertebra, a bounding box is extracted from the low-resolution mask and transferred to the original image. Additionally, the bounding box is expanded by 24 voxels in each direction to ensure that the entire vertebra is covered. These volumes of interest (VOI) need to be analyzed in tiles of 96 × 96 × 48 voxels due to hardware limitations. These tiles are fed into the segmentation CNN, which determines which voxels in the VOI are part of the vertebra. Because this network therefore performs essentially the same task as the localization network, we use the same architecture and a similar segmentation strategy: In the iterative localization and identification step, the network received a binary mask with all already detected vertebrae as input. In the segmentation step, the network receives the upsampled mask containing all voxels that were marked as part of the vertebra in the low-resolution image (Figure 2b). This is necessary because the tiles may contain parts of two neighboring vertebrae, so that the networks needs to be able to infer from the input which part of vertebral bone to segment. Since the vertebrae are segmented one-by-one, the vertebral label is known in advance for each tile and the output of the segmentation network is therefore only binary. The segmentation step is followed by minimal post-processing: For each vertebral label, only the largest connected component is retained and morphological closing is performed in-plane to close potentially appearing small holes in the segmentation.

The CNNs were implemented using the Theano framework[13] and trained on Nvidia Titan X Pascal GPUs with 12 GB memory. The Dice coefficient was used as cost function[14] and Adam[15] was used for optimization with a fixed learning rate of 0.001. Batches of single image volumes were used and therefore a momentum of 0.99 to average gradients over multiple image volumes. During training, vertebrae and tiles were chosen randomly and the binary masks of the vertebrae for the localization and identification network were derived from the reference segmentation. The segmentation network was trained using the upsampled low-resolution masks generated by first training the localization and identification network and afterwards applying it to the training data. Random elastic transformations were used to augment the data during training of both networks.

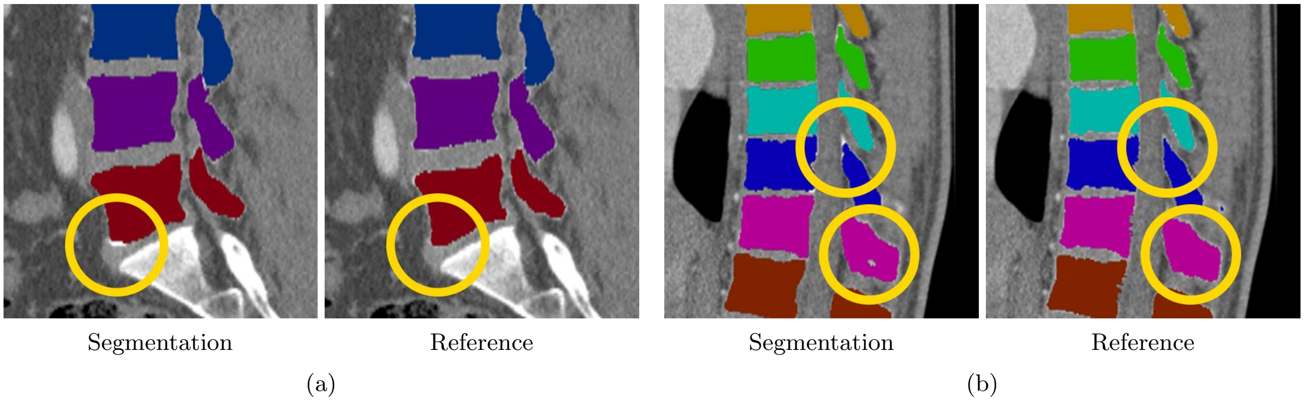


Figure 4: Examples of inaccuracies in the segmentation obtained with the proposed automatic segmentation method, compared to the manual reference standard. In (a), a small part of the last lumbar vertebra L5 is missed. In (b), the automatic method misclassified few voxels on the surface of Th10 as well as some voxels in the spinous process of Th11. Note, however, the high overall accuracy of the segmentations.

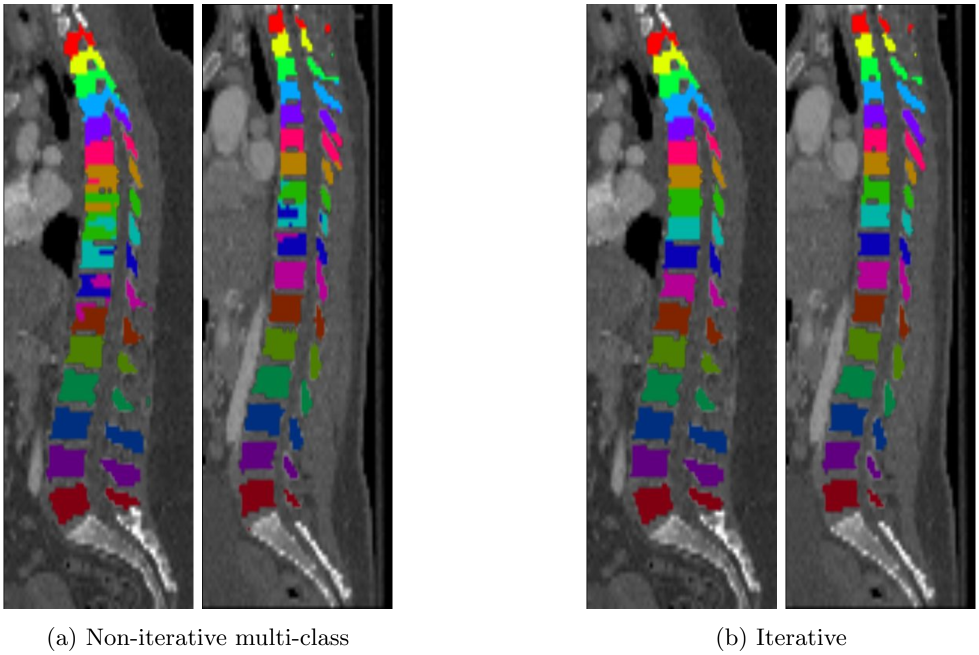
4. RESULTS AND DISCUSSION

To allow for a comparison of the results with other methods that participated in the challenge, we used the evaluation metrics that were also used in the challenge,[12] Dice coefficient and mean absolute surface distance (ASD). Overall, the method achieved a mean Dice coefficient of 0.948 and a mean ASD of 0.29 mm. Our method outperforms the best method that participated in the challenge, which achieved a comparable mean Dice of 0.947, but a slightly higher mean ASD of 0.37 mm. All five methods that participated in the challenge were model-based methods and all of them had lower performance on the smaller upper thoracic vertebrae. Conversely, our method performs better on the upper vertebrae and makes slightly more mistakes on the lumbar vertebrae. We achieved mean Dice coefficients of 0.942 for T1–T6, of 0.954 for T7–T12 and of 0.948 for L1–L5. The mean ASDs in these categories were 0.23 mm for T1–T6, 0.26 mm for T7–T12 and 0.44 mm for L1–L5. Examples of segmentations obtained with the proposed method are shown in Figure 3. While the segmentations achieve high volume overlap scores, there are some minor inaccuracies. Examples of these are shown in Figure 4.

The localization and identification CNN applied iteratively correctly identified all thoracic and lumbar vertebrae in the five test images. When training the same network as a multi-class network that identifies all vertebrae in one pass, without the iterative process, we observed labeling mistakes especially in the last thoracic vertebrae T8–T12 (Figure 5). Iterative identification, i.e., enforcing an ordered localization and simplification of the detection task, therefore substantially improved the performance when training with only ten images. However, a CNN might be able to implicitly learn the spatial relationship between the individual vertebrae when trained on a larger number of scans. Manual segmentation of the spine, however, is a very labor intensive process that can take around 40 hours per scan, so that a large number of reference annotations is often not available.

The iterative use of a CNN mimics a recurrent neural network, which analyze sequences of inputs and retain information over multiple sequential samples. While recurrent networks retain information in dedicated variables, our iterative approach implicitly retains information by receiving the accumulated output of the previous iterations as input in the next iteration. This iterative strategy can therefore be understood as a simplified recurrent network. Compared to a recurrent neural network, the iterative approach is more memory efficient, which facilitates the use of large 3D input volumes, and is easier to train. It also uses the available training data more efficiently, since each vertebra is treated as an individual sample.

A reference vertebra is required as starting point for the iterative process. Here, we used the last lumbar vertebra as it was well recognizable in all scans since all scans covered at least part of the sacrum. However, other vertebra that can be reliably identified can be used as reference point as well, e.g., the first thoracic vertebra T1 or the last thoracic vertebra T12, which both can be identified from the rib cage. With more training data, the network might also be able to recognize other vertebrae both from their anatomical context and their shape.


(a) Non-iterative multi-class

(b) Iterative

Figure 5: Results of vertebra identification using (a) a network with 17 output channels compared to (b) the proposed iterative approach. With only ten training images, the non-iterative multi-class segmentation approach does not generalize as well as the binary iterative approach.

A limitation of this work is that only scans of subjects with healthy spines were used. After the challenge, the organizers provided five test scans of diseased subjects. However, such examples were not available for training and we observed that our approach did not perform well for most vertebra with severe compression fractures. Since the proposed method is fully supervised, representative training data is required for a fair evaluation. In future work, we are therefore aiming to evaluate our method on a larger set of subjects with various pathologies. Furthermore, the proposed CNN-based iterative segmentation strategy has applications in other sequential segmentation tasks, such as vessel segmentation. Additionally, future generations of GPUs with a larger internal memory will potentially enable the localization and identification network to analyze the scans at the original resolution, which would allow unifying the two networks into a single network.

5. CONCLUSIONS

We presented a method based on iteratively applied convolutional neural networks for identification and segmentation of the thoracic and lumbar vertebrae in spine CT images. The iterative segmentation strategy introduces structural priors for iterative detection of sequences of objects and enables training of deep convolutional neural networks with a limited number of manual reference segmentations. Evaluated on publicly available data from a segmentation challenge, this method outperforms the methods that participated in the original challenge.

ACKNOWLEDGMENTS

We would like to thank the organizers of the spine segmentation challenge at MICCAI 2014 for their efforts in creating the challenge dataset and for making the dataset publicly available.

REFERENCES

[1] Glocker, B., Feulner, J., Criminisi, A., Haynor, D., and Konukoglu, E., "Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], *LNCS* **7512**, 590–598, Springer (2012).

[2] Chen, H., Shen, C., Qin, J., Ni, D., Shi, L., Cheng, J. C. Y., and Heng, P.-A., "Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], *LNCS* **9349**, 515–522, Springer (2015).

[3] Chu, C., Belavỳ, D. L., Armbrecht, G., Bansmann, M., Felsenberg, D., and Zheng, G., "Fully automatic localization and segmentation of 3D vertebral bodies from CT/MR images via a learning-based method," *PLoS ONE* **10**(11), e0143327 (2015).

[4] Korez, R., Likar, B., Pernuš, F., and Vrtovec, T., "Model-based segmentation of vertebral bodies from MR images with 3D CNNs," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], *LNCS* **9901**, 433–441, Springer (2016).

[5] Bromiley, P. A., Kariki, E. P., Adams, J. E., and Cootes, T. F., "Fully automatic localisation of vertebrae in CT images using random forest regression voting," in [*International Workshop on Computational Methods and Clinical Applications for Spine Imaging*], *LNCS* **10182**, 51–63, Springer (2016).

[6] Sekuboyina, A., Valentinitsch, A., Kirschke, J. S., and Menze, B. H., "A localisation-segmentation approach for multi-label annotation of lumbar vertebrae using deep nets." arXiv:1703.04347 (2017).

[7] Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J., and Išgum, I., "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Transactions on Medical Imaging* **35**, 1252–1261 (2016).

[8] Litjens, G. J. S., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I., "A survey on deep learning in medical image analysis," *Medical Image Analysis* **42**, 60–88 (2017).

[9] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], *LNCS* **9351**, 234–241, Springer (2015).

[10] Yang, D., Xiong, T., Xu, D., Huang, Q., Liu, D., Zhou, S. K., Xu, Z., Park, J., Chen, M., Tran, T. D., et al., "Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization," in [*IPMI*], *LNCS* **10265**, 633–644, Springer (2017).

[11] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O., "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], *LNCS* **9901**, 234–241, Springer (2016).

[12] Yao, J., Burns, J. E., Forsberg, D., Seitel, A., Rasoulian, A., Abolmaesumi, P., Hammernik, K., Urschler, M., Ibragimov, B., Korez, R., et al., "A multi-center milestone study of clinical vertebral CT segmentation," *Computerized Medical Imaging and Graphics* **49**, 16–28 (2016).

[13] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions." arXiv:1605.02688 (2016).

[14] Milletari, F., Navab, N., and Ahmadi, S.-A., "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in [*International Conference on 3D Vision*], 565–571 (2016).

[15] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization." arXiv:1412.6980 (2014).