EPiC Series in Health Sciences

Volume 5, 2022, Pages 116–120

Proceedings of The 20th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

EPiC Health Sciences

CNN based 2D vs. 3D Segmentation of Bone in Ultrasound Images

Benjamin Hohlmann, Peter Brößner and Klaus Radermacher
Rheinisch-Westfälisch Technische Hochschule, Aachen, Germany
hohlmann@hia.rwth-aachen.de

Abstract

Fully-automatic and reliable segmentation of bone surface in volumetric ultrasound images could enable the use of this imaging technique for a variety of tasks, including diagnosis of hip dysplasia, ACL injuries in the knee as well as patient-specific instrumentation and implants in total hip or knee arthroplasty. Interpretation of volumetric data is a hard task, even for humans. In this study, we investigate the benefit of using the spatial information of a third dimension on the task of segmentation of the distal femoral bone. A data set of 52 volumetric image with 12771 image slices is split into a training and test set. We employ 2D and 3D variants of the nnUNet architecture and compare the accuracy in terms of dice coefficient and performance in terms of inference time. Note that processing of 2D data allows for a bigger model due to less memory consumption. Both architectures achieve a Dice of about 82% while the 2D variant shows less false positive segmentation and achieves a surface distance error of 0.44mm, in contrast to 0.81mm for the 3D variant. At the same time, the former infers three times faster at about 10 seconds per volume image. Apparently, model size has a bigger positive effect than the additional spatial information. Thus, we recommend considering 2D segmentation architectures even for volumetric segmentation tasks.

1 Introduction

Ultrasound is widely used in orthopedics for diagnostic purposes. It provides insight into the patient’s body at chair side without causing any radiation exposure. At the same time, the costs of a sonography are low compared to other imaging techniques like computed tomography (CT) or magnetic resonance imaging. However, correct interpretation of ultrasound images require years of experience. Recent advances in fully automatic image processing may enable unskilled clinical personal to make use of this imaging technology, by training a machine-learning model to learn the skills of an expert. Segmentation of the bone surface is one aspect of the automatic diagnosis. Tasks like classification of hip dysplasia in infants [1], screw placement for fixation of bone fractures [2] or manufacturing of patient-specific implants and instruments [3] require an accurate delineation of the bone. Computed tomography, which defines the gold standard for computer-assisted orthopedic surgery related tasks, comes with an in-slice resolution of about half a millimeter, defining the target accuracy of our model. On the other hand, processing time is a crucial factor. During an exam at chair side, several volume images may be acquired to cover the area relevant for diagnostic purposes. For the technology to be accepted, processing should not prolong the exam time noticeably. As such, processing should be in the range of minutes. For intra-operative applications, the acceptable processing time is even lower. We define a target of ten seconds.

In segmentation of volumetric images using convolutional neural networks (CNNs), the two metrics of accuracy and processing time oppose each other: While 2D images can be processed in real-time, incorporation of a third spatial dimensions greatly increases the computational burden. The trade-off between computational expense and bone surface reconstruction accuracy is unclear. Even though various works have been published on one or the other, the literature is missing a direct comparison: There is no common data set and implementations of CNNs vary greatly in network structure, size or hyperparameters.

As such, we investigate the trade-off of incorporating 3D spatial information in segmentation of bone surface in ultrasound images regarding (1) accuracy und (2) computational costs in a fair benchmark environment.

2 Materials and Methods

Fifty-two volume in-vivo images were acquired with a SonixTouch Q+ and a mechanically swiping probe (Ultrasonix, Peabody, USA), totaling in 12771 image slices. The images depict the right femur of three male subjects of age 26-28. While the width of the images is fixed to 418 pixels, the height varies from 289 to 511 pixels depending on the patient's body weight. Only slices that show actual bone surface are included and thus the depth varies from 139 to 378. The images exhibit an isotropic pixel spacing of 0.1mm. The bone surfaces were labeled manually using the open source software 3DSlicer. This data set is processed either as 2D or 3D data. In both cases, the images, labels and splits are the exact same.

The most commonly used CNN in medical image processing is the U-Net. Various extensions were introduced over recent years that altered the training scheme, loss function, convolution operation or model architecture. However, none of these variations proofed to be beneficial over a wide range of tasks. Isensee et al. [4] showed that adaption of the networks hyperparameters to the data set is more important than structural changes. They developed the "no new U-Net" (nnUNet) framework that adjusts the network to the data at hand and established a new state of the art in 49 public medical segmentation challenges. As such, it provides an ideal benchmarking tool that does not require any manual fine-tuning. Furthermore, due to its automatic adaption, no additional validation set for fine-tuning is necessary.

The nnUNet offers a variety of models, including a 2D and 3D one, among others. All hyperparameters were set by the framework, just the architecture type was manually set to the before mentioned options. The networks were trained for three days on a Volta 100 GPU provided by the RWTH Aachen University GPU Cluster. For measuring inference time, just the prediction itself is measured, not the data generation. Model size or capacity was 40.801.632 trainable parameters for the 2D model and 31.167.584 for the 3D model. Note that the framework defines a bigger model for the 2D case as less data needs to be loaded into VRAM, allowing for a more complex model.

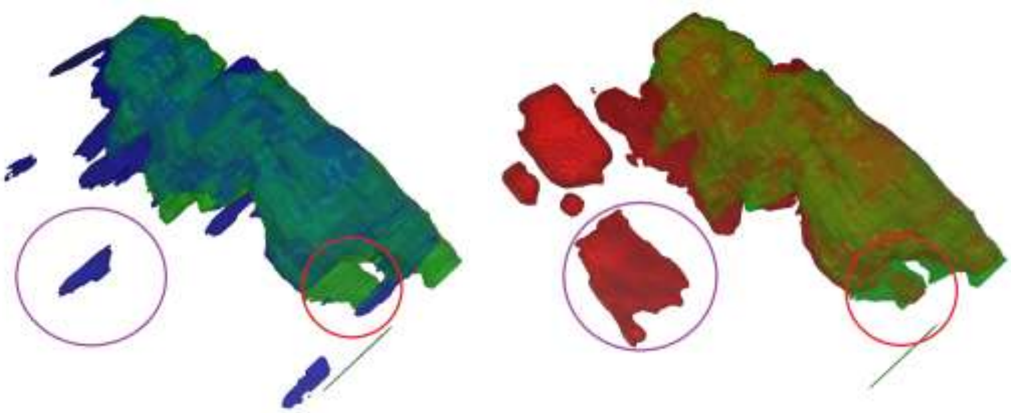

Figure 1: Segmentation of bone in a representative volumetric ultrasound image. 2D (blue, left) and 3D (red, right) predictions are shown along with the manual ground truth labels (green). Most of the surface was correctly detected in both cases. Note the tendency to segment thin slices in the 2D case and the large false positive segmentation in the 3D case (purple circle). In some areas, spatial context may have helped the 3D model to detect the bone surface (red circle).

3 Results

Table 1 shows all metrics for both architectures. The 3D variant has a lower number of trainable parameters and as such, a lower capacity compared to its 2D counterpart. Still, the inference time is higher, on average 90ms compared to 27ms per slice and 34.2s compared to 10.3s per volume. The 3D variant achieves a high dice of 82.26. The 2D variant performs similar with 81.38. We further thin out the segmentation in the 2D image slices to obtain a surface and compute the average symmetric surface distance error (SSDE) in 3D. In this regard, the 3D model performs noticeably worse with an average error of 0.81mm in contrast to 0.44mm for the 2D model. Investigating the individual volume images, we observe substantial false positive segmentation in four of the ten volumes for the 3D model in contrast to just one such case for the 2D model. See Figure 1 for a qualitative evaluation on such a volume image.

Model	Capacity (# parameters)	Inference time		Dice	SSDE
		per slice (ms)	per volume (s)		(mm)
2D	40.801.632	27	10.3	81.38	0.44
3D	31.167.584	90	34.2	82.26	0.81

Table 1: All test results for the 2D and 3D architectures: Model capacity, inference time per slice and volume as well as dice coefficient and average symmetric surface distance error (SSDE).

4 Discussion

Detection of bone surfaces is much easier for a human if information on the surrounding tissue is available. Surprisingly, we could not see any benefit of using 3D spatial information for the task of femur bone segmentation in volumetric ultrasound images with CNNs. Apparently, the increased model size has a stronger positive impact than the additional spatial cues. At the same time, the computational burden increases 3-fold. Processing of a full volume may take up to 34 seconds in case of the 3D model, not meeting the requirements set for intra-operative use. The 2D model on the other hand comes close to the goal of 10 seconds. However, both architectures exhibit very good performance compared to other publications on the segmentation of bone in ultrasound images that report average dice errors of 0.75 and 0.89 [5, 6]. Note that the latter uses a slightly different metric definition. The 2D model even achieves an accuracy comparable to CT resolution. Accordingly, at the current state we recommend to consider 2D segmentation even for volumetric images.

There are several limitations to this study, the most important being that only a single data set and a single architecture is tested. Furthermore, the study cohort is very small and homogeneous and the just a single ultrasound probe was used. The orthopedic ultrasound research community is missing a versatile and public data set, which could strengthen the reliability of studies like this one. Additionally, any evaluation on test images using common metrics needs to be validated in surgical practice. Regarding inference, details of the soft- and hardware may have a strong impact and as such, the reported times should be interpreted with caution. Recently, the vision transformer architecture has been established, which computes global interaction of image patches and as such may leverage 3D spatial information better [7]. The inclusion of these architectures into the analysis is one aspect of our ongoing research.

5 Acknowledgements

Simulations were performed with computing resources granted by RWTH Aachen University, under project rwth0536.

References

1. El-Hariri H, Hodgson AJ, Mulpuri K et al. (2021) Automatically Delineating Key Anatomy in 3-D Ultrasound Volumes for Hip Dysplasia Screening. Ultrasound in Medicine & Biology 47:2713–2722. https://doi.org/10.1016/j.ultrasmedbio.2021.05.011
2. Anas EMA, Seitel A, Rasoulian A et al. (2016) Registration of a statistical model to intraoperative ultrasound for scaphoid screw fixation. Int J Comput Assist Radiol Surg 11:957–965. https://doi.org/10.1007/s11548-016-1370-y
3. Hohlmann B, Radermacher K (2020) Augmented Active Shape Model Search – towards 3D Ultrasound-based Bone Surface Reconstruction. In: The 20th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery. EasyChair, 117-111
4. Isensee F, Jaeger PF, Kohl SAA et al. (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18:203–211. https://doi.org/10.1038/s41592-020-01008-z
5. El-Hariri H, Mulpuri K, Hodgson A et al. (2019) Comparative Evaluation of Hand-Engineered and Deep-Learned Features for Neonatal Hip Bone Segmentation in Ultrasound. In: Shen D (ed) Medical Image Computing and Computer Assisted Intervention - MICCAI 2019: 22nd international conference, Shenzhen, China, October 13-17, 2019, proceedings, vol 11765. Springer International Publishing, Cham, pp 12–20
6. Duong DQ, Nguyen K-CT, Kaipatur NR et al. (2019) Fully Automated Segmentation of Alveolar Bone Using Deep Convolutional Neural Networks from Intraoral Ultrasound Images. Annu Int Conf IEEE Eng Med Biol Soc 2019:6632–6635. https://doi.org/10.1109/EMBC.2019.8857060

7. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov et al. (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations