Benjamin Hohlmann*, Jakob Glanz and Klaus Radermacher

Segmentation of the distal femur in ultrasound images

https://doi.org/10.1515/cdbme-2020-0034

Abstract

Objectives: Ultrasound is a widely used imaging technology that allows for fast diagnosis of a broad range of illnesses and injuries of the musculoskeletal system. However, interpreting ultrasound images remains a challenging task that requires expert knowledge and years of training for each exam. One crucial step for the long-term goal of automatic diagnosis is pixel wise semantic segmentation.
Methods: In this work, several state-of-the-art semantic segmentation networks were trained on a new dataset of manually annotated ultrasound images depicting the distal femur.
Results: PSP-Net achieved the best overall performance with an average surface distance error (SDE) of 0.64 mm.
Conclusions: We recommend the PSP-Net architecture for semantic segmentation of bone surfaces.

Keywords: CNN; femur; segmentation; ultrasound.

Introduction

Ultrasound is a widely established imaging technology. It is used in virtually every medical field, ranging from ophthalmology over mammography to orthopedics. However, interpretation of ultrasound images is a challenging task and its usage is limited to experts. Automatic semantic segmentation could enable less experienced personnel to utilize ultrasound. Based on the outline of bone structures in the image for example, standard diagnostic measurements like the alpha and beta angle for DDH or femur to tibia offset for ACL rupture diagnosis can be performed. By combining multiple segmented B-Mode images, three-dimensional models for patient-specific implant planning e.g., in total knee arthroplasty (TKA) or biomechanical analysis can be constructed.

Several limitations complicate this automation: The signal to noise ratio in ultrasound images is very low. Furthermore, the field-of-view of common ultrasound transducer is limited to several centimeters in both, depth and width. The appearance of the bone surface in the image is highly dependent on its local topology and the inclination angle of the ultrasound beam [1]. At the same time, it resembles soft tissue interfaces. Only experienced sonographers are able to cope with these complications. On the other hand, ultrasound is a cheap alternative to other imaging techniques like magnetic resonance imaging (MRI) or computed tomography (CT) and does not harm the patient in any way. Finally, it allows for real-time interaction with the patient. If the limitations in terms of image quality can be overcome, ultrasound has the potential to reduce health care costs strongly while providding fast and reliable diagnosis at the same time. Providing automatic and real-time semantic segmentation could improve the ease of use and reproducibility, therefore enabling unexperienced medical personal to conduct a variety of diagnostic exams.

Since the record-breaking success of AlexNet, a convolutional neuradl network (CNN), on the ImageNet classification challenge [2], CNNs are applied on many other tasks, including semantic segmentation. The U-Net [3], a fully convolutional encoder-decoder variant, got very popular in medical image processing and a high number of publications apply this architecture to this day. However, the evolution of CNNs did not stop in 2015, and several improvements were proposed since then. Furthermore, gathering annotated training data is an especially difficult and expensive task for medical images.

Therefore, this work contributes to the long-term goal of automatic interpretation of ultrasound images two-fold: (1) We compare several state-of-the-art semantic segmentation architectures from the non-medical field, using (2) a new annotated dataset of ultrasound images of the knee joint.

Related work

Many surveys and review papers provide an overview of the state-of-the-art in medical image processing with deep learning. Most of these cover a wide range of different

***Corresponding author: Benjamin Hohlmann**, RWTH, Pauwelsstr 20, Aachen, Germany, E-mail: hohlmann@hia.rwth-aachen.de
Jakob Glanz and Klaus Radermacher, RWTH, Aachen, Germany

imaging modalities, typically including MRI, CT, arthroscopy and Ultrasound, and tasks such as registration, classification and segmentation [4–6]. Several publications focus on a single task (segmentation [7]), architecture (CNN [8]) or imaging technology (Ultrasound [9, 10]). The evaluation metrics differ strongly, typically including the dice coefficient or an intersection-over-union-based computation, as well as the average and maximal surface distance error (SDE). However, the architectures applied in these publications reveal a considerable time lag to the non-medical computer vision community. A high number relies on the famous U-Net and its variants, which dates back to 2015.

Challenges are a common method for an objective comparison of different architectures on a certain task. For semantic segmentation, popular benchmark challenges include the PASCAL Visual Object Classes (VOC), ADE20k, Cityscapes and COCO Stuff [11–14]. The leaderboards of these update every few months with new architectures achieving all-time highs. For medical images, there are only a few challenges, which also come with limited dataset size. For ultrasound segmentation, these include nerve, intra-vascular vessel [15] and cardiac segmentation [16]. To date and the best of our knowledge, there is no published challenge on ultrasound bone segmentation.

Methods

Our dataset consists of 36 volumetric ultrasound images, depicting the distal femora of three study participants. The data was recorded with a SonixTouch Q+ machine (Ultrasonix, Peabody, USA). The images have a width of 381 and height of 465 pixels with 600 slices in each volume image, totaling in 21,600 images. It should be noted, that the isotropic pixel spacing of 0.1 mm results in a high correlation of these neighboring images. We therefore manually segmented visually different slices with two classes, femur and background. We excluded images without visible bone surface, totaling in 3,707 labeled images. Only the bone response is segmented, not the pitch black 'bone shadow' area below. To counter class imbalance issues, we segment a thick line. See Figure 1 for an example. The data is split into two sets, training and validation, with 30 and six vol images, respectively. Each dataset depicts different subjects. The training was performed on two Volta 100 GPUs on the RWTH Aachen University GPU Cluster.

We applied several state-of-the-art semantic segmentation networks from the non-medical domain, which achieved the highest performance in several challenges: Deeplabv3+ [17] on PASCAL VOC 2012, HRNetv2 + object-contextual representation (OCR) [18] on Cityscapes, COCO Stuff and PASCAL Context and PSP-Net [19] on ADE20k.

All of these challenges differ from the task of medical image segmentation. The number of training images available, as well as the number of classes to segment is higher. Furthermore, all datasets consist of natural images with three color channels. The objects to be segmented show a big variety, including the fine-grained 'object' category and 'stuff' segmentation, like grass. Therefore, we apply these networks on our dataset and analyze which architecture choices are beneficial for the task of ultrasound bone segmentation. We also apply the U-Net for reference.

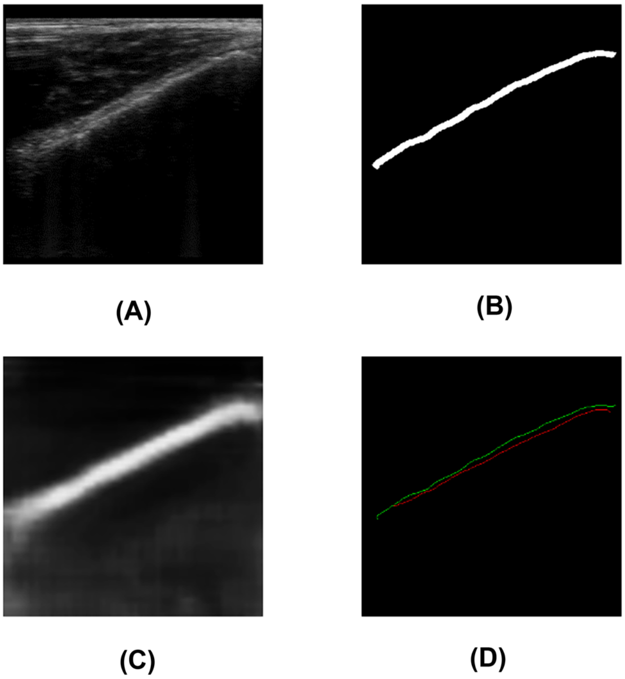

Figure 1: Ultrasound image (A), ground truth (B), prediction (C) and extracted bone surfaces (D) for MobileNetv2. Ground truth surface in green, prediction in red.

As hyper parameter tuning is crucial for performance of deep learning architectures and we do not want it to invalidate the comparison, we try to minimize and even out the effort spent on this. Optimized parameters include learning rate, batch size and data augmentation intensity. Training is aborted when the evaluation metrics on the validation set increase over five epochs. Custom data-augmentation techniques conform to ultrasound image characteristics are applied, including left-right flipping, varying brightness and a combined cropping, shearing and resizing.

Another essential aspect is the model capacity. Given a rather small medical data set, overfitting may likely be an issue. We opted for MobileNetv2 as a Deeplab variant with small capacity to counter this. However, due to non-convergence issues, we did not train from scratch but pre-trained on PASCAL VOC.

Several evaluation metrics are analyzed on the validation set. The ground truth and the predicted bone masks are thresholded and thinned to a single line. Following, the mean SDE and Hausdorff distance, directed as well as symmetric, are determined. These metrics provide an estimate of the accuracy of the segmentation and allow for detection of over-segmentation. We also count the number of empty segmentation masks in order to detect under-segmentation. We did not evaluate any area-based metrics like the dice or Intersection-over-Union, as our ground truth masks are thin lines, which are prone to low surface overlap, even for good segmentations. See Table 1 for an overview of the different architectures.

Table 1: Architecture overview. The number of trainable parameters gives the model size. HR-Net is trained using online hard example mining (OHEM).

Model	Model size	Loss function	Pre-training
U-Net	31.031.688	Focal loss [20]	No
MobileNet	2.141.762	Focal loss	Yes
PSP-Net	49.066.948	Cross entropy	No
HR-Net	65.847.122	OHEM-CE	No

Results

As expected, U-Net achieved low errors. As can be seen in Figure 2, the average SDE of 0.87 mm is promising. The small difference between the directed (4.2 mm) and symmetric (4.9 mm) Hausdorff distance error indicates a good trade-off between over- and under-segmentation. Still, in five out of 384 images the bone was not segmented.

HR-Net, combined with the OCR module, achieved similar results with an average SDE of 0.88 mm, a lower Hausdorff of 3 mm directed and 4.6 mm in the symmetric case. The number of not segmented images was slightly higher with 16.

The stand-alone HR-Net performed better in terms of the average SDE of 0.63 mm and Hausdorff distance of 2.4 mm with a similar symmetric Hausdorff distance of 4.8 mm. In addition, all bone surfaces were segmented.

MobileNetv2 was able to further reduce the average SDE down to 0.56 mm. Again, no bone surface was missed and the Hausdorff distance of 2.5 mm shows a high robustness. The high symmetric Hausdorff of 6.4 mm on the other hand revealed over-segmentation issues.

Finally, PSP-Net combined a low average SDE of 0.64 mm with a good directed (2.6 mm) and symmetric (4.2 mm) Hausdorff distance. Only a single image was falsely segmented to contain no bone surface.

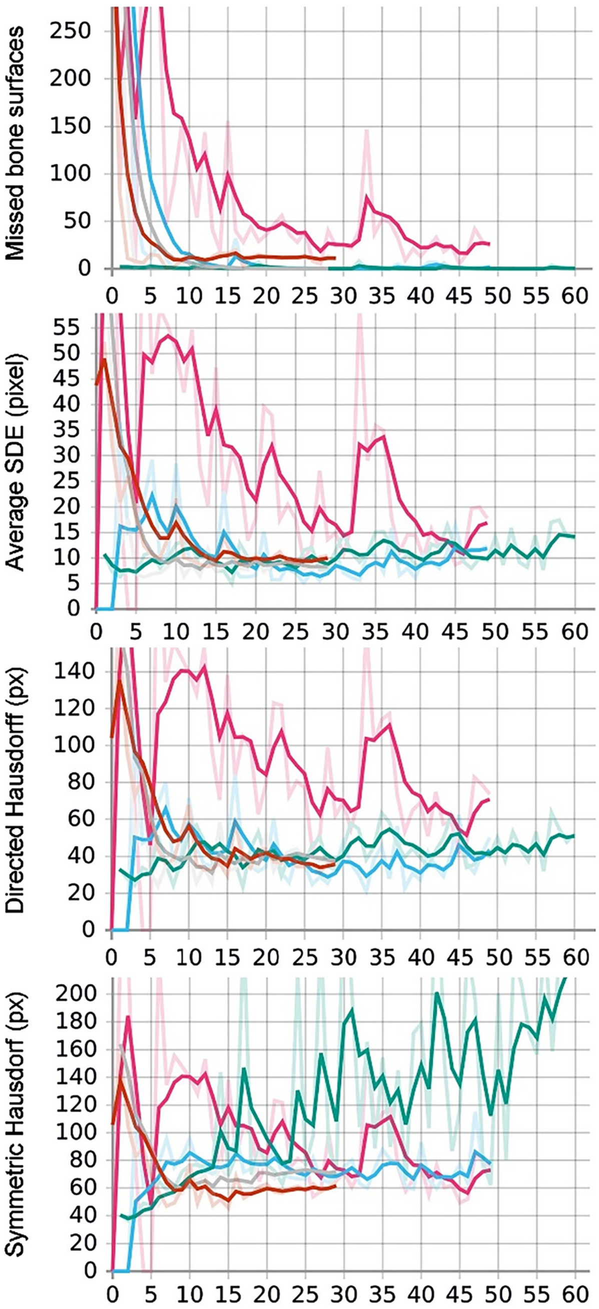

Figure 2: Different metrics over epochs on the validation set. U-Net in magenta, HR-Net in gray, HR + OCR in red, MobileNetv2 in teal and PSP-Net in petrol. Non-smoothed values in transparent. See colored figure online.

Discussion

U-Net showed a reasonable overall performance with a good trade-off in over- and under-segmentation. The other architectures' average SDE approaches the in-plane resolution of about 0.5 mm per pixel in the gold standard for bone imaging, CT. Similarly, to the Hausdorff distance, slice thickness in CTs is in the range of several millimeter. This is sufficient for several tasks: In pre-operative implant size selection in TKA for example, a rule-of-thumb for the maximal overhang of the implant over the bone is 3 mm [21].

We could not observe an advantage of using the OCR module. We hypothesize that the low number of classes combined with strong texture differences within the background class limit the benefit of the region context. In a setup, where the background class is split into the individual tissues this may change.

MobileNetv2 shows promising results while maintaining a lightweight architecture. However, it remains unclear whether the low capacity or rather pre-training enabled its success.

For PSP-Net, the training parameters are most likely not yet optimal. After only three epochs, the model starts to overfit. A different learning rate policy may help to improve its segmentation. Still, it performed best.

To further objectify this architecture comparison, extensive testing under versatile conditions is required. These include varying dataset sizes, as deep learning models depend on a big data basis. Pre-training, as well as synthetic images can help to increase the dataset size drastically. Assumingly, high capacity architectures like the HR-Net will profit stronger from this. Additionally, the segmentation mask definition may vary: Defined as is, strong class imbalance is a task specific challenge, which may be eased by labeling the whole bone volume instead of the bone response only. Furthermore, an intense hyper parameter tuning could once again alter the results of this evaluation. Keeping in mind these limitations, we recommend using the state-of-the-art PSP-Net for semantic segmentation of ultrasound images in orthopedics.

Future research will focus on an alternative loss function using distance transform images. To confirm the generalization of our analysis, a third test set will be acquired and evaluated. Additionally, the manual segmentation quality will be examined to determine a lower error bound for this specific dataset and to further lower it.

Research funding: The author state no funding involved.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study.
Informed consent: Informed consent was obtained from all individuals included in this study.
Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' Institutional Review Board or equivalent committee.

References

1. Jain AK, Taylor RH. Understanding bone responses in B-mode ultrasound images and automatic bone surface extraction using a Bayesian probabilistic framework. In: Medical imaging 2004: ultrasonic imaging and signal processing; 2004, vol. 5373:131–42 pp.
2. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60: 84–90.
3. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2015, vol. 9351:234–41 pp.
4. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
5. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–48.
6. Kumar ES, Bindu CS. Medical image analysis using deep learning: a systematic literature review. In: Somani A, Ramakrishna S, Chaudhary A, Choudhary C, Agarwal B, editors. Emerging Technologies in Computer Engineering: Microservices in Big Data Analytics 2019. https://doi.org/10.1007/978-981-13-8300-7_8.
7. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Martinez-Gonzalez P, Garcia-Rodriguez J. A survey on deep learning techniques for image and video semantic segmentation. Appl Soft Comput 2010;70:41–65.
8. Qayyum A, Anwar SM, Awais M, Majid M. Medical image retrieval using deep convolutional neural network. Neurocomputing 2017; 266:8–20.
9. Liu S, Wang Y, Yang X, Lei B, Liu L, Li SX, et al. Deep learning in medical ultrasound analysis: a review. Engineering 2017;5: 261–75.
10. Hacihaliloglu I. Ultrasound imaging and segmentation of bone surfaces: a review. Technology 2017;5:74–80.
11. Caesar H, Uijlings J, Ferrari V. COCO-stuff: thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018:1209–18 pp.
12. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016:213–23 pp.
13. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. Int J Comput Vis 2010;88:303–38.
14. Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A. Scene parsing through ADE20K dataset. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. https://doi.org/10.1109/CVPR.2017.544.
15. Balocco S, Gatta C, Ciompi F, Wahle A, Radeva P, Carlier S, et al. Standardized evaluation methodology and reference database for evaluating IVUS image segmentation. Comput Med Imag Graph: Off J Comput Med Imag Soc 2014;38:70–90.
16. Tobon-Gomez C, Craene M, McLeod K, Tautz L, Shi W, Hennemuth A, et al. Benchmarking framework for myocardial tracking and deformation algorithms: an open access database. Med Image Anal 2013;17:632–48.
17. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV); 2018:801–18 pp.
18. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, et al. TPAMI: deep high-resolution representation learning for visual recognition. In: IEEE transactions on pattern analysis and machine intelligence; 2020.

19. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. https://doi.org/10.1109/CVPR.2017.660.
20. LinT-Y, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: The IEEE International Conference on Computer Vision (ICCV); 2017. https://doi.org/10.1109/ICCV.2017.324.
21. Mahoney OM, Kinsey T. Overhang of the femoral component in total knee arthroplasty: risk factors and clinical consequences. J Bone Joint Surg 2010;92:1115–21. American volume.