EPiC Series in Health Sciences

Volume 5, 2022, Pages 148–153

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

Standardized Evaluation of Current Ultrasound Bone Segmentation Algorithms on Multiple Datasets

Prashant U. Pandey[1], Benjamin Hohlmann[2], Peter Broessner[2], Ilker Hacihaliloglu[1], Keiran Barr[3], Tamas Ungi[3], Oliver Zettinig[4], Raphael Prevost[4], Guillaume Dardenne[5], Zian Fanti[6], Wolfgang Wein[4], Eric Stindel[5], Fernando Arambula Cosio[6], Pierre Guy[1], Gabor Fichtinger[3], Klaus Radermacher[2], and Antony J. Hodgson[1]

[1] University of British Columbia, Vancouver, BC, Canada
prashant@ece.ubc.ca
[2] RWTH Aachen University, Aachen, Germany
[3] Queen's University, Kingston, ON, Canada
[4] ImFusion GmbH, Munich, Germany
[5] University Hospital of Brest, Brest, France
[6] Universidad Nacional Autónoma de México, CDMX, Mexico

Abstract

Ultrasound (US) bone segmentation is an important component of US-guided orthopaedic procedures. While there are many published segmentation techniques, there is no direct way to compare their performance. We present a solution to this, by curating a multi-institutional set of US images and corresponding segmentations, and systematically evaluating six previously-published bone segmentation algorithms using consistent metric definitions. We find that learning-based segmentation methods outperform traditional algorithms that rely on hand-crafted image features, as measured by their Dice scores, RMS distance errors and segmentation success rates. However, there is no single best performing algorithm across the datasets, emphasizing the need for carefully evaluating techniques on large, heterogenous datasets. The datasets and evaluation framework described can be used to accelerate development of new segmentation algorithms.

1 Introduction

Ultrasound (US) is increasingly being used in orthopaedic procedures, as a radiation-free alternative to fluoroscopic (X-ray) imaging. In particular, identifying bone surfaces in US images is a core task in enabling several computer-assisted orthopaedic surgeries (CAOS) [8], whilst minimizing harmful radiation exposure [4]. However, interpreting US images remains a challenging task for physicians and algorithms alike, due to significant speckle and noise, limited field-of-view, and sensitivity to machine and patient characteristics.

Several algorithms for segmenting bone in US images have been proposed, in efforts to automate and improve the repeatability of clinical care [8, 2]. However, most of these algorithms

are developed in isolated 'silos', and are therefore evaluated on datasets which are unique to each institution and using non-standardized definitions and implementations of evaluation metrics [8]. This makes it difficult to directly compare segmentation techniques with each other, and therefore evaluating the contribution of new techniques over previous ones.

In this paper we attempt to resolve these issues by making three contributions: 1) Curating a multi-institutional dataset of 20,810 unique US images containing labeled bone surfaces; 2) Defining and implementing a set of standardized evaluation metrics; 3) Performing a systematic evaluation of six previously-published bone segmentation algorithms on the curated datasets using the proposed evaluation metrics. We believe these contributions will help standardize and democratize US bone imaging, and facilitate developing US bone segmentation techniques in the future.

2 Methods

We organized US bone imaging data from three institutions into datasets, and created automated evaluation environments for each of these. Six previously published automatic bone segmentation algorithms were benchmarked using this system. Example images and reference segmentations are presented in Figure 1A.

2.1 Datasets

The University of British Columbia (UBC) dataset consists of 16,995 2D US images collected from 10 healthy subjects at UBC, using a Telemed MicrUS system with a L12-5L40S-3 (Telemed, Italy). We split the dataset into a training set of 13,687 images from 8 subjects and a private test set of 3,308 images from 2 subjects.

The RWTH Aachen dataset consists of 2,550 slice images depicting 11 healthy subjects, imaged at RWTH Aachen University with a SonixTouch Q+ machine (Ultrasonix, USA). We split the dataset into a training set of 2,030 images from 9 subjects and a private test set of 520 images from 2 subjects.

The Rutgers University dataset consists of 1,265 2D US images collected from 14 healthy subjects at Rutgers University. Subjects were scanned with either a SonixTouch system or a Clarius convex C3 probe (Clarius Mobile Health Corporation, Canada). We split this dataset into a training set of 962 2D images from 11 subjects, and a private test of 303 2D images from 3 subjects.

2.2 Automated Evaluation

Each dataset was organized into its own evaluation task, in the style of typical medical image segmentation 'challenge'. When relevant, we followed the best practice guidelines proposed in [6], as our intention is to make this dataset and evaluation framework publicly available in the future. The private test set segmentation labels from each dataset were not used or available for training or guiding the development of segmentation algorithms.

Each evaluation was deployed as a Docker container, which included a program to process each segmentation prediction, as generated by each algorithm, against the reference segmentation using the evaluation metrics described below.

2.2.1 Evaluation Metrics

We implemented three evaluation metrics. The first was the Dice score (equivalent to F1 score), defined as the overlap between the dilated predicted and reference segmentations. The reference R and predicted P segmentations were dilated by a 1 mm disk structuring element, and the Dice score was calculated as $Dice = \frac{2|R \cap P|}{|R|+|P|}$. Here $|R|$ represents the count of all bone surface pixels in the reference segmentation.

We also calculated the root-mean-squared (RMS) Euclidean distance error from the predicted segmentation to the reference segmentation after the segmentations had been skeletonized. Finally, we reported the success rate, where a successful segmentation is defined as where the predicted and reference bone surface share at least one scan-line in the US image.

2.3 Bone Segmentation Algorithms

We evaluated a total of six bone segmentation algorithms, two 'classical' non-learning methods and four learning methods based on deep convolutional neural networks (CNNs). The non-learning methods were Confidence-weighted Structured Phase Symmetry (CSPS) [9], and Shadow Peak (SP) [7]. The learning algorithms included three U-Net bone segmentation methods: U-Net (Salehi 2017) [10], U-Net (El-Hariri 2019) [1], U-Net (Ungi 2020) [11], and a DeepLabV3+ based model [3]. Further details of each method can be found in the original cited publications.

2.4 Statistical Analyses

We used a Friedman test to compare segmentation algorithms within each dataset evaluation, as it is a non-parametric, repeated measures analysis. This was followed by multiple post-hoc Wilcoxon signed-rank tests with Bonferroni's correction to find statistically significant differences ($p < 0.05$) between each algorithm.

3 Results

The performance of each algorithm on the datasets is summarized in Table 1, and box plots illustrating the Dice score and RMS distance errors are presented in Figure 1.

The DeeplabV3+ algorithm achieved the highest success rate across all three evaluations (range 89% - 100%), and all four learning algorithms achieved a 100% success rate on the Rutgers dataset (Table 1). DeeplabV3+ also achieved the highest mean Dice scores of 0.72 and 0.80 on the UBC and RWTH Aachen evaluations respectively, compared to the other methods ($p < 0.05$). The U-Net (El-Hariri 2019) algorithm achieved the lowest RMS distance errors on the UBC evaluation (mean: 0.69 mm, stdev: 1.76 mm; $p < 0.05$), and along with DeeplabV3+ achieved the lowest RMS distance error on the Aachen dataset (means: 0.64 mm and 0.70 mm, stdevs: 1.66 mm and 1.55 mm respectively). The U-Net (El-Hariri 2019), U-Net (Ungi 2020), and DeeplabV3+ algorithms achieved the lowest, statistically equivalent, RMS distance errors on the Rutgers dataset. The learning methods outperformed the non-learning methods across all metrics over the three evaluations.

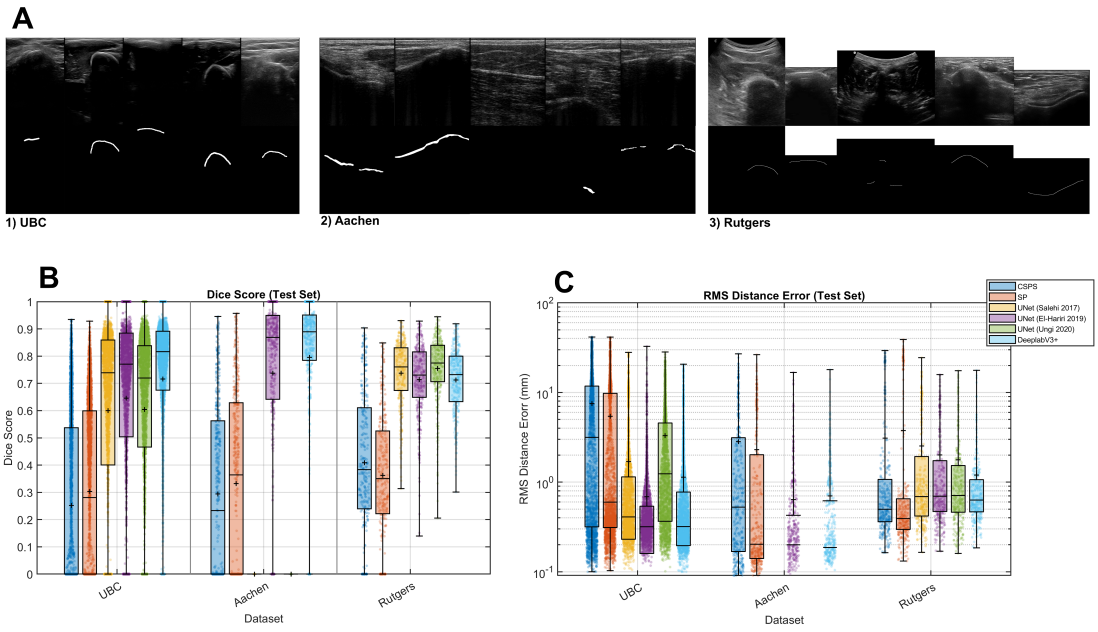

Figure 1: **A:** Examples of ultrasound images and corresponding reference segmentations from the three datasets. **B:** Dice score and **C:** Root-mean-squared (RMS) distance error box plots of each segmentation algorithm on each dataset. Central horizontal line represents the median and black cross represents the mean of each algorithm.

4 Discussion And Conclusion

We curated three US bone segmentation image datasets, on which we benchmarked six bone segmentation algorithms. Our results demonstrated that learning-based methods are the best performing over all datasets. However, there was no single algorithm that performed best across all datasets and metrics (Table 1). Furthermore, the performance of the same algorithms across the datasets varied significantly, emphasizing the need to validate algorithms on large, heterogeneous sets of data [6]. The results also indicated that differences in hyperparameters, such as learning rate and loss function, can lead to significant performance differences even when the network architecture is similar or identical, as demonstrated by the three U-Net based methods.

Our future work aims to benchmark these algorithms across additional institutional datasets. Furthermore, we plan to define new evaluation metrics specifically suited to bone surfaces, as Dice overlap and RMS distance error provide two different measurements of performance which are not always in agreement [5].

To the best of our knowledge, this work is the first to provide a framework for systematically comparing bone segmentation algorithms to each other across multiple datasets in a consistent manner [8]. We plan to release the datasets and evaluation frameworks publicly in order to facilitate the development of new bone segmentation algorithms and to accelerate US-guided CAOS.

Table 1: Mean (and standard deviations) of Dice score, RMS distance error, and success rate for each algorithm and dataset evaluation. Bold values indicate best performers per dataset, tested statistically.

	UBC	**RWTH Aachen**	**Rutgers**
CSPS	0.25 (0.30) 7.50 mm (9.50 mm) 56%	0.29 (0.31) 2.83 mm (4.74 mm) 65%	0.41 (0.26) 3.10 mm (6.54 mm) 92%
SP	0.30 (0.31) 5.43 mm (7.94 mm) 59%	0.33 (0.33) 2.31 mm (4.18 mm) 61%	0.36 (0.23) 3.76 mm (8.43 mm) 86%
U-Net (Salehi 2017)	0.60 (0.35) 1.69 mm (3.53 mm) 78%	-	0.74 (0.12) 2.53 mm (3.92 mm) 100%
U-Net (El-Hariri 2019)	0.65 (0.35) **0.69 mm (1.76 mm)** 81%	0.74 (0.32) **0.64 mm (1.66 mm)** 87%	0.71 (0.13) **2.01 mm (2.88 mm)** **100%**
U-Net (Ungi 2020)	0.60 (0.32) 3.30 mm (4.44 mm) 83%	-	**0.75 (0.12)** **1.77 mm (2.81 mm)** **100%**
DeeplabV3+	**0.72 (0.29)** 1.14 mm (2.46) **89%**	**0.80 (0.28)** **0.70 mm (1.55 mm)** **91%**	0.71 (0.12) **1.20 mm (2.01 mm)** **100%**

References

[1] Houssam El-Hariri, Kishore Mulpuri, Antony Hodgson, and Rafeef Garbi. Comparative Evaluation of Hand-Engineered and Deep-Learned Features for Neonatal Hip Bone Segmentation in Ultrasound. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11765 LNCS, pages 12–20. Springer, oct 2019.

[2] Ilker Hacihaliloglu. Enhancement of bone shadow region using local phase-based ultrasound transmission maps. *Int. J. Comput. Assist. Radiol. Surg.*, 12(6):951–960, mar 2017.

[3] Benjamin Hohlmann, Jakob Glanz, and Klaus Radermacher. Segmentation of the distal femur in ultrasound images. *Current Directions in Biomedical Engineering*, 6(1), may 2020.

[4] Jae Young Hong, Kyungdo Han, Jin Hyung Jung, and Jung Sun Kim. Association of Exposure to Diagnostic Low-Dose Ionizing Radiation With Risk of Cancer Among Youths in South Korea. *JAMA Network Open*, 2(9):e1910584–e1910584, sep 2019.

[5] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. *Medical Image Analysis*, 67:285–296, 2021.

[6] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching Wei Wang, Marc André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.*, 9(1), 2018.

[7] Prashant Pandey, Pierre Guy, Antony J. Hodgson, and Rafeef Abugharbieh. Fast and automatic bone segmentation and registration of 3D ultrasound to CT for the full pelvic anatomy: a comparative study. *Int. J. Comput. Assist. Radiol. Surg.*, 13(10):1515–1524, may 2018.

[8] Prashant U. Pandey, Niamul Quader, Pierre Guy, Rafeef Garbi, and Antony J. Hodgson. Ultrasound Bone Segmentation: A Scoping Review of Techniques and Validation Practices. *Ultrasound Med. Biol.*, 46(4):921–935, apr 2020.

[9] Niamul Quader, Antony Hodgson, and Rafeef Abugharbieh. Confidence weighted local phase features for robust bone surface segmentation in ultrasound. *MICCAI Work. Clin. Image-Based Proced.*, 1:76–83, 2014.

[10] Mehrdad Salehi, Raphael Prevost, José Luis Moctezuma, Nassir Navab, and Wolfgang Wein. Precise ultrasound bone registration with learning-based segmentation and speed of sound calibration. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, volume 10434 LNCS of *Lecture Notes in Computer Science*, pages 682–690. Springer International Publishing, Cham, 2017.

[11] Tamas Ungi, Gabor Fichtinger, Hastings Greer, Kyle Sunderland, Victoria Wu, Zachary M. C. Baum, Christopher Schlenger, Matthew Oetgen, Kevin Cleary, and Stephen Aylward. Automatic spine ultrasound segmentation for scoliosis visualization and measurement. *IEEE Trans. Biomed. Eng.*, pages 1–1, mar 2020.