Feature-aggregated spatiotemporal spine surface estimation for wearable patch ultrasound volumetric imaging

Baichuan Jiang\textsuperscript{a}, Keshuai Xu\textsuperscript{a}, Abhay Moghekar\textsuperscript{b}, Peter Kazanzides\textsuperscript{a}, Emad Doctor\textsuperscript{a}

\textsuperscript{a}Department of Computer Science, Johns Hopkins University, Baltimore, USA; \textsuperscript{b}Department of Neurology, Johns Hopkins Medical Institute, Baltimore, USA

ABSTRACT

Clear identification of bone structures is crucial for ultrasound-guided lumbar interventions, but it can be challenging due to the complex shapes of the self-shadowing vertebra anatomy and the extensive background speckle noise from the surrounding soft tissue structures. Therefore, in this work, we will present our method for estimating the vertebra bone surfaces by using a spatiotemporal U-Net architecture learning from the B-Mode image and aggregated feature maps of hand-crafted filters. Additionally, we are integrating this solution with our patch-like wearable ultrasound system to capture the repeating anatomical patterns and image the bone surfaces from multiple insonification angles. 3D bone representations can then be created for interventional guidance. The methods are evaluated on spine phantom image data collected by our proposed “Patch” scanner, and our systematic ablation experiment shows that improved accuracy can be achieved with the proposed architecture. Equipped with this surface estimation network, our wearable ultrasound system can potentially provide intuitive and accurate interventional guidance for clinicians in an augmented reality setting.

Keywords: Bone surface estimation, deep learning, volume reconstruction, wearable ultrasound, lumbar puncture

1. INTRODUCTION

Ultrasound imaging has been introduced in recent years to guide clinicians to perform spine interventions such as lumbar puncture, but the complex shapes of the self-shadowing lumbar vertebrae make it challenging for clinicians to correctly interpret the images, thus reducing the intervention success rate\textsuperscript{1}. The situation is aggravated when combined with other risk factors such as obesity\textsuperscript{2} and abnormal vertebrae\textsuperscript{3}, leading to failures and repeated needle insertion attempts which increases the risk of complications including cerebrospinal fluid (CSF) leak and headache\textsuperscript{4}. Therefore, in our earlier work\textsuperscript{5}, we presented a wearable ultrasound device “Patch” to allow volumetric imaging of patient lumbar vertebrae from multiple insonification angles for improving the vertebrae surface visibility. Ultimately, the imaging device will be incorporated into an augmented reality setting and aid the clinicians performing spine intervention.

The key idea of the “Patch”-based solution\textsuperscript{5} is to use a miniaturized robotic mechanism moving a phased array transducer in a two-degree-of-freedom (2-DOF) workspace and imaging the lumbar vertebrae from multiple angles, as shown in Fig. 1. The miniaturized robotic mechanism can be attached to the back of the patient (with tape or belt), which is highly desirable to avoid spatial displacement between patient anatomy and the imaging device, and allows the clinician to perform the needle insertion with both hands free. In addition, because the bone surface reflection signal is dependent on the incident beam angle, a multi-angle volumetric imaging setup can potentially collect images for each

Figure 1. Wearable patch ultrasound scanner setup (a) proposed solution of patch device-based lumbar puncture guidance. (b) equivalent 2-DOF workspace with rotation and translation. (c) Data acquisition experiment setup on phantoms.
piece of the vertebrae surface at an appropriate angle. As a result, our “patch” wearable device can provide better imaging and reconstruction of the vertebrae due to its ability to precisely control and measure the scanning motion.

One crucial step toward successful interventional guidance for lumbar puncture is the extraction and presentation of the imaging information obtained. Specifically, it is highly desirable to present the patient lumbar anatomy as a 3D surface model in an augmented reality setting. However, given the high noise level and various imaging artefacts, it remains a challenging task to correctly identify vertebrae surfaces in ultrasound images. Berton et al.\textsuperscript{6} leveraged random forests-based feature extraction to segment the spinous process in the cross-section view images, but it was limited to only delineating the top part of the spinous process where high reflection is typically observed. Wang et al.\textsuperscript{7} instead utilized a U-Net type neural network combined with filter-based feature maps to estimate bone surfaces, but had not demonstrated its application to the highly complex vertebrae anatomy.

In this work, we propose to utilize a spatiotemporal U-Net architecture with raw image and aggregated hand-crafted feature map as inputs, along with a visibility-based ground truth label generator, to estimate the visible surfaces to the maximum extent based on the image information. With estimated surface locations and device kinematics readings, we show that a 3D surface representation of the vertebrae can be successfully reconstructed for interventional guidance.

2. METHOD

2.1 Aggregated bone surface feature map generation

Before training our backbone neural network, we decided to implement a hand-crafted feature extractor to provide one channel of high-quality input to the U-Net estimator. This feature extraction workflow is demonstrated in Fig. 2. Given the original B-Mode phased array image input, we first inverse scan converted it back to the polar coordinate, and two feature extractors are separately executed and then aggregated as final output.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Aggregated feature map generation workflow. Original B-mode image from a spine phantom is first scan-converted back into polar coordinate space, and shadow detection and edge detection are executed and aggregated into single feature map to be fed into the neural network along with the original image as the input.}
\end{figure}

Phase feature extraction: The high intensity response from the tissue-bone boundary is one prominent feature to localize the bone surface. Researchers have proposed using local phase image information to extract the bone boundary based on its symmetric signature\textsuperscript{8}. Specifically, a set of 2-D Log-Gabor filters were used and each computes the difference between odd and even filter responses. Intuitively, a high intensity narrow-bar feature (ridge) in the image could be
defined as any point at which all the Fourier spectral components have phase congruency at 90°, thus the highest even filter response and lowest odd filter response for the quadrature filters (in our case, the Log-Gabor filters). Following the suggested values from literature, we construct our Log-Gabor filter bank with 2 wavelength scales and 6 orientations (30 degrees apart from each other). An example of a filter result is shown in Fig. 2 green box, where bar features in the image are generally getting high responses.

**Shadow feature extraction:** Shadow is another determining feature for bone structures, and it can differentiate bones from other bar feature-inducing reflectors such as muscle, tendon, and background speckle noise. Here we implemented a confidence map based shadow feature extractor. The confidence map denotes the information uncertainty in the attenuated (shadow) region, and it can be solved by modeling the problem within a random walk framework: for a pixel in the output feature map, its intensity corresponds to the probability of a random walk starting from this pixel to reach the virtual transducer element, under user defined walking constraints and penalty values. One example can be seen in Fig. 2 (orange box): the leftmost figure is the computed confidence map from the original image (inverse scan-converted); and we use a Sobel edge detector and a pre-defined threshold value to generate a binary bone boundary map, as in the middle figure; after applying a Gaussian filter with kernel size 6 we get the image on the right.

As it can be observed that the outputs from both the phase symmetry feature extractor and the shadow feature extractor are usually noisy due to the misleading bar and hypoechoic features present in the image, we propose to aggregate both the phase symmetry feature map and the shadow feature map via element-wise multiplication to achieve mutual agreement between two feature maps. The aggregated output feature map is further scan-converted into Cartesian space as the additional input channel to our spatiotemporal U-Net.

### 2.2 Spatiotemporal U-Net with image and aggregated feature channel input

The lumbar vertebrae have a repetitive anatomical appearance pattern longitudinally, and our patch ultrasound device is also scanning in a back-and-forth pattern with a slightly different angle for each sweep as shown in Fig. 1(b). Therefore, to capture this temporal pattern and the inherent 3D anatomical context information, we propose to incorporate a recurrent neural network (RNN) into a U-Net type segmentation model for spatiotemporal context learning. Inspired by VesNet, the model architecture is shown in Fig. 3 below. The two input channels consist of the original B-Mode image channel and the aggregated feature map channel, which are stacked along the channel dimension.

**Figure 3.** Spatiotemporal U-Net architecture. ConvGRU unit is taking encoded feature maps from the previous frame data. The spatial attention unit tries to utilize deeper level information with higher context value to guide segmentation of small targets, while the channel attention unit aims to adapt the weights for B-Mode and aggregated feature channel automatically.

The key components of the network architecture include:

1. **Encoder-Decoder Backbone:** Our model utilizes a U-Net-like backbone in which features can be learned from multiple spatial scales. To optimize for the mobile computing, we reduce the original U-Net to a four-layer
architecture, with a total number of \(~300k\) network parameters, B-Mode and Feature inputs are both normalized between \((0, 1)\). Batch normalization and PReLU activation are used and resize-up-convolutions\(^{13}\) have replaced up-sampling to minimize checkerboard artifacts.

(2) ConvGRU unit: to capture the temporal context (and the 3D anatomical context) information, we embedded the convolutional gated recurrent unit (ConvGRU\(^{12}\)) in the skip connections at different scales. Compared to the traditional GRU, replacing the dot product with convolutional operations can preserve the inherent spatial connectivity within the feature maps.

(3) Spatial attention unit: For some frames, the surface target is small compared to the size of the image. Therefore, as suggested by the literature\(^{14}\), we implemented the soft spatial self-attention unit that utilizes the higher semantic information from deeper layers with access to global spatiotemporal context to direct the feature activations from shallower layers, suppressing irrelevant background and attending to regions of interest.

(4) Channel attention unit: To capture the interdependency between the B-Mode channel and the feature channel, we employ the convolutional block attention modules (Cbam)\(^{15}\) to best exploit the multi-channel information.

2.3 Visibility-based surface ground truth label generation

For successfully training the proposed network on spine ultrasound images, we implemented a visibility-based ground truth label generator. First, after acquiring 3D volumetric scans of the vertebrae, we manually annotate the spinal surface ultrasound responses where high confidence is ensured. Then we utilize an iterative closest point (ICP)\(^{16}\) algorithm to register the annotated surface points to the CT model of the spine. With the registered US-CT data, we can utilize the pose information of each contributing slice of the US volume and create ray beams corresponding to the imaging poses, as shown in Fig. 4(b), to identify points where ray beams first hit the CT spine surface. Here we chose a sampling step of \(0.3^\circ\) rotating the ray around the transducer origin, and within each step the closest point is kept as a visible point, as shown in Fig. 4(c). Optionally, a fan shape region-of-interest is applied to crop out the final surface points and the points are further dilated to form a connected surface as in Fig. 4(d). Finally, a 2D Gaussian filter is applied to represent the uncertainty in registration Fig. 4(e), and is used during the network training process. During the testing phase, we again binarize the label as in Fig. 4(f) to compute dice coefficient with the predicted surface.

![Figure 4. Visibility-based ground truth label generation process. (a) original image, (b) registered cross-section view of the spine model, (c) sampled visible surface points, (d) ROI cropping and binary dilation, (e) Gaussian smoothed label used for training, and (f) binarized surface label (with 0.5 threshold) for dice score computation.](image)

3. EXPERIMENTS AND RESULTS

3.1 Experiment environment setup

Two sets of CT data of patients with no spine abnormalities were obtained from a public dataset\(^{18}\). Then, vertebrae L3-L5 were manually segmented from the CT and 3D printed. The printed spine models were embedded within a container filled with gelatin, as shown in Fig. 1(c). Metamucil fiber is also added in the gelatin background material for creating a
realistic soft tissue scattering appearance. An example B-Mode ultrasound image of the phantom is shown in Fig. 4(a). In total, 3367 frames of images were acquired from 4 sets of scan experiments: each set of scans is located on different sides of the two different spine models, and each scan contains 8 linear sweep motions by the patch scanner with distinct angles (15° apart from each other) with the spine midline. Because we know the exact spatial transformation of each image, the images were first post-processed by performing a volume reconstruction followed by cross-section interpolation based on the original image location. The processed images are used as the B-Mode channel input, with examples shown in Fig. 5. For the following experiments, we split the dataset into training and testing groups, where each group contains consecutive image sequences from the acquired data, so we do not lose the temporal correlation for RNN training. In total, 2000 images were used for training and 1367 for testing.

3.2 Data augmentation and training methods

Our dataset only contains a limited number of ultrasound images collected on the same phantom to ensure that the trained network is tested on images unseen by the network during the training process. As a result, it is likely that the network will overfit the limited training data. Therefore, extensive data augmentation is needed for the generalizability of the network, and we employed the following augmentation methods: (1) Spatial augmentation: we perform drastic geometric augmentations including random translation, rotation, cropping, scaling and flipping. (2) Contrast augmentation: we randomly adjust the image contrast by intensity clipping and stretching. (3) Channel augmentation: for training the two channel networks, we will randomly drop out the feature channel input to force the network to learn features from the B-Mode channel. To train the surface estimator, we utilized a weighted Dice-loss (w-Dice) function, which is computed by:

\[
\text{wDice} = 2 \times \left( \frac{\text{reduce}_\text{sum}(X \odot Y)}{\text{reduce}_\text{sum}(X \oplus Y)} \right)
\]

where \(X\) and \(Y\) are network output’s sigmoid activation and ground truth label map, respectively. In the operation, \(\odot\) denotes element-wise multiplication and \(\oplus\) denotes element-wise addition. With \(\text{reduce}_\text{sum}\) operation the formula output essentially represents the dice coefficient between two heatmaps such that heatmap pixel intensity is used as weights. In the end, for all the following experiments, the spatiotemporal bone surface estimator is trained with the ADAM\(^{17}\) optimizer with an initial learning rate of 1e-4 for 60 epochs.

3.3 Results and Discussion

To evaluate the effectiveness of each proposed mechanism used in our surface estimator, we have conducted an ablation study and the controlled variable experiment results are summarized in Tab. 1.

| Experiment ID | Loss Type | Network Type | RNN Reset Type | Test Data | Input Channel | Avg. Dice score |
|---------------|-----------|--------------|----------------|-----------|---------------|----------------|
| 1             | w-Dice    | RNN          | Fixed-length   | Unseen image | BMode+Feature | 0.422          |
| 2             | w-CE      | RNN          | Fixed-length   | Unseen image | BMode+Feature | 0.416          |
| 3             | w-Dice    | CNN          | N/A            | Unseen image | BMode+Feature | 0.287          |
| 4             | w-Dice    | RNN          | Align-with-scan| Unseen image | BMode+Feature | 0.493          |
| 5             | w-Dice    | RNN          | Align-with-scan| Unseen anatomy| BMode+Feature | 0.436          |
| 6             | w-Dice    | RNN          | Align-with-scan| Unseen image  | BMode only    | 0.331          |
| 7             | w-Dice    | RNN          | Align-with-scan| Unseen image  | Feature only  | 0.405          |

Key observations from the ablation experiments include:

1) *Exp1 vs. Exp2*: Our weighted Dice (w-Dice) loss (using label/prediction intensity as weights) outperforms the traditional weighted Cross-Entropy (w-CE) loss.

2) *Exp1 vs. Exp3*: Compared to a pure CNN (without recurrent units, enlarged to a comparable number of trainable parameters), networks with ConvGRU units show better performance, which suggest that the recurrent part can potentially learn from the temporal context (as well as the full anatomical context information) to make better vertebrae surface estimation.
(3) **Exp1 vs. Exp4:** If we align the RNN reset length (# of frames to run before reset) with patch scanning motion, i.e., reset whenever patch changes scanning direction, the network can learn the repetitive anatomy better, this demonstrates that using a miniaturized robotic device which has full access to the motion data can potentially improve the imaging and reconstruction quality.

(4) **Exp4 vs. Exp5:** If the training and testing images come from the same anatomy, the results will generally be better than testing on an unseen anatomy, but the difference is still minor. We will be making more phantoms with varying anatomy and enhanced realism to ensure network generalizability.

(5) **Exp4 vs. Exp6:** With the addition of our aggregated feature channel (with <1% increase of overall network size), the network can obtain a major improvement comparing to using B-Mode image only.

(6) **Exp4 vs. Exp7:** Although *Exp7* can achieve a decent score when trained with feature channel only, it still cannot surpass the results with both channel active, suggesting the importance of having full access to the available information.

A qualitative result example of our surface estimator is also shown in Fig. 5. We can visually confirm that the 3D vertebrae surface model is well recovered and can potentially be used to guide clinicians to perform lumbar intervention. Besides, the network is very computationally efficient since each frame of inference takes around 13ms on a laptop with a RTX3070 graphics card.

4. **CONCLUSIONS**

In this work we presented a method for estimating the complex shaped vertebra surface from ultrasound images acquired by a wearable patch-like ultrasound scanner. With the visibility-based label generator to increase the correlation with image data, our results showed that our spatiotemporal context learning-based surface estimator can achieve promising results for guided lumbar intervention using the reconstructed 3D model. In addition, the proposed network can leverage the wearable device’s awareness of scanning motion to enhance surface estimation.

In the future, we will expand our database for network training and conduct experiments with animal data. Ultimately, we will integrate the surface estimator with our angle-based reconstruction framework, and use the wearable ultrasound imaging system with an augmented reality-based guidance system to evaluate overall LP-guidance performance.
5. ACKNOWLEDGEMENTS

This work was supported by National Science Foundation [Award number 1653322], and student fellowship program sponsored by Analog Devices Inc.

REFERENCES

[1] Gueziri, Houssem-Eddine, Carlo Santaguida, and D. Louis Collins. "The state-of-the-art in ultrasound-guided spine interventions." Medical Image Analysis 65 (2020): 101769.
[2] Edwards, Cory, Enrique C. Leira, and Pedro Gonzalez-Alegre. "Residency training: a failed lumbar puncture is more about obesity than lack of ability." Neurology 84, no. 10 (2015): e69-e72.
[3] Rosiak, Grzegorz, Anna Lusakowska, Krzysztof Milezarek, Dariusz Konecki, Anna Fraczek, Olgierd Rowinski, and Anna Kostera-Pruszczczyk. "Ultra-low radiation dose protocol for CT-guided intrathecal nusinersen injections for patients with spinal muscular atrophy and severe scoliosis." Neuroradiology 63, no. 4 (2021): 539-545.
[4] Seeberger, Manfred D., Mark Kaufmann, Sven Staender, Markus Schneider, and Daniel Scheidegger. "Repeated dural punctures increase the incidence of postdural puncture headache." Anesthesia & Analgesia 82, no. 2 (1996): 302-305.
[5] Xu, Keshuai, Baichuan Jiang, Abhay Moghekar, Peter Kazanzides, and Emad Docter. "AutoInFocus, a new paradigm for ultrasound-guided spine intervention: a multi-platform validation study." International Journal of Computer Assisted Radiology and Surgery 17, no. 5 (2022): 911-920.
[6] Berton, Florian, Farida Cheriet, Marie-Claude Miron, and Catherine Laporte. "Segmentation of the spinous process and its acoustic shadow in vertebral ultrasound images." Computers in biology and medicine 72 (2016): 201-211
[7] Wang, Puyang, Vishal M. Patel, and Ilker Hacihaliloglu. "Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided CNN." In International conference on medical image computing and computer-assisted intervention, pp. 134-142. Springer, Cham, 2018
[8] Hacihaliloglu, Ilker, Rafeef Abagheribieh, Antony J. Hodgson, and Robert N. Rohling. "Bone surface localization in ultrasound using image phase-based features." Ultrasound in medicine & biology 35, no. 9 (2009): 1475-1487.
[9] Karamalis, Athanasios, Wolfgang Wein, Tassilo Klein, and Nassir Navab. "Ultrasound confidence maps using random walks." Medical image analysis 16, no. 6 (2012): 1101-1112.
[10] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234-241. Springer International Publishing, 2015.
[11] Jiang, Baichuan, Alvin Chen, Shyam Bharat, and Mingxin Zheng. "Automatic ultrasound vessel segmentation with deep spatiotemporal context learning." In International Workshop on Advances in Simplifying Medical Ultrasound, pp. 3-13. Springer, Cham, 2021.
[12] Siam, Mennatullah, Sepehr Valipour, Martin Jagersand, and Nilanjan Ray. "Convolutional gated recurrent networks for video segmentation." In 2017 IEEE international conference on image processing (ICIP), pp. 3090-3094. IEEE, 2017.
[13] Odena, Augustus, Vincent Dumoulin, and Chris Olah. "Deconvolution and checkerboard artifacts." Distill 1, no. 10 (2016): e3.
[14] Abraham, Nabila, and Naimul Mefraz Khan. "A novel focal tversky loss function with improved attention u-net for lesion segmentation." In 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), pp. 683-687. IEEE, 2019.
[15] Woo, S., J. Park, J. Y. Lee, and I. So Kweon. "Cham: convolutional block attention module. In proceedings of the European conference on computer vision (ECCV): 3-19." (2018).
[16] Besl, Paul J., and Neil D. McKay. "Method for registration of 3-D shapes." In Sensor fusion IV: control paradigms and data structures, vol. 1611, pp. 586-606. Spie, 1992
[17] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
[18] Saltz, Joel. "Stony Brook University COVID-19 Positive Cases." The Cancer Imaging Archive (2021).
[19] Jiang, Baichuan, Keshuai Xu, Abhay Moghekar, Peter Kazanzides, and Emad M. Docter. "Insonification Angle-based Ultrasound Volume Reconstruction for Spine Intervention." In 2022 IEEE International Ultrasonics Symposium (IUS), pp. 1-4. IEEE, 2022.