Program Digest, 2019 IEEE International Ultrasonics Symposium (IUS)
Glasgow, Scotland, October 6-9, 2019

Comparison of Deep Learning and Classical Image Processing for Skin Segmentation

Felix Q. Jin*, Michael Postiglione*, Anna E. Knight*,
Adela R. Cardones†, Kathryn R. Nightingale*, Mark L. Palmeri*
* Department of Biomedical Engineering, Duke University, Durham, USA
† Department of Dermatology, Duke University, Durham, USA
Email: *felix.jin@duke.edu*

***Abstract*—Skin stiffness correlates with the progression of sclerotic skin diseases. Ultrasound shear wave elasticity imaging techniques can measure skin stiffness, but an accurate skin thickness measurement is required to compute the elastic modulus. We explored different automated methods to segment the skin for use in real-time skin elastography. Local gradient-based methods could not robustly segment the skin on our B-mode images, so we developed a new thresholding method to detect the edges of the skin. We also used our thresholding method to generate labels to train a deep neural network. We compared the performance of thresholding and the trained network for central thickness estimation on a held-out test set. Our thresholding method correctly segmented 58% of images, and the neural network correctly segmented 82%. More than half of thresholding failures on the test set were from overestimation of the bottom skin boundary. The neural network had significantly less overestimation failures and similar rates of failure due to bubbles and underestimation.**

***Index Terms*—deep learning, skin, segmentation**

I. Introduction

Sclerotic skin diseases such as systemic sclerosis and cutaneous graft-versus-host disease are characterized by stiff and fibrotic skin [1], [2]. The lack of a reliable and quantitative metric for disease severity has motivated the application of ultrasound elastography techniques for measuring skin stiffness [3]–[5].

Shear wave elasticity imaging (SWEI) tracks the speed of off-axis shear waves generated using an acoustic radiation force impulse [6]. In most tissues, the shear wave speed c_s can be directly related to the shear modulus μ using (1), where ρ is the tissue density.

$$\mu = \rho c_s^2 \tag{1}$$

However, this simple relation does not hold in skin because the thickness is on the order of a shear wave's wavelength, approximately 3 mm. In [3], a Lamb wave model was used to convert shear wave speed to tissue elasticity. This thin-plate model introduced a dependence on thickness [7], with the characteristic ratio being wavelength to plate thickness. In [4], skin thickness was used to normalize the elastic modulus. An accurate measurement of skin thickness is needed to correctly calculate the shear and elastic moduli.

This work was supported by the National Institutes of Health (R01EB002132, R42CA228159, T32GM007171), and the Duke Skin Disease Research Center Translational and Innovative Research Support Program.

On B-mode ultrasound, the dermis of the skin appears as a hyperechoic band, as shown in Fig. 1. Skin thickness is typically measured by manually drawing a region of interest around the visualized dermis. However, this process is time consuming and liable to human error. Our goal was to explore automated methods for skin segmentation to allow for real-time elastography and to improve measurement consistency.

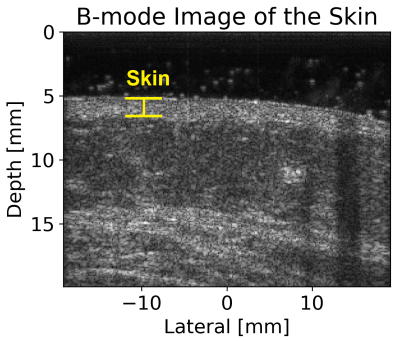

Fig. 1. A B-mode ultrasound image of the skin. The skin appears as the proximal hyperechoic band (labeled). Fatty subcutaneous tissue appears hypoechoic, whereas muscle, connective tissue, and fibrosis appear bright.

Previous work on automated skin segmentation for ultrasound imaging has relied on image gradients and iterative optimization. [8] used a modified active contours method to trace along the skin boundary with a snake. [9] used an iterative process to fit a curve that passes through regions of high gradient magnitude. In our image dataset, there was relatively poor contrast between the dermis and subcutaneous tissue, and these local gradient methods failed to leverage global pixel intensity statistics. Therefore, we explored a global binary thresholding method that does not use iterative optimization.

We also explored deep learning for skin segmentation. Deep learning methods have achieved state-of-the-art performance in a variety of image processing tasks [10], including medical image segmentation [11]. Deep convolutional neural networks have wide receptive fields to identify global features. The output of a trained network depends on both the input data and on its training history, which acts as a prior. We hypothesized that a deep learning approach to skin segmentation would

be more robust than a classic method because deep neural networks can combine local, global, and prior information.

The objectives of this study were to automate skin segmentation and compare the performance of deep learning and classical image processing. We developed a new thresholding approach for skin segmentation and used it to help train a neural network. We also analyzed the failure modes of these two methods.

II. Methods

A. Image Dataset

Our dataset consisted of unlabeled B-mode images of the skin collected during a previous IRB-approved study [3]. These images were acquired using a Siemens ACUSON S2000 scanner with a Siemens 14L5 transducer transmitting at 6.15 MHz. Pulse-inversion harmonic imaging was used to improve image quality [12]. During acquisition, the probe was held above a layer of ultrasound gel to position the skin near the focal depth of 5.5 mm. The images in our dataset were acquired from 18 study subjects at a variety of anatomic locations, both healthy and diseased. IQ data were envelope detected and fractional-power compressed to form the B-mode image. Images were 20 mm in depth and 38 mm laterally.

The dataset was partitioned according to Fig. 2. First, we removed all images where the skin was not visible. Next, this clean dataset was divided into a training (15 patients) and test group (3 patients). Creating segmentations by hand would have been time consuming. Instead, we used our automated thresholding method to propose segmentations, and manually evaluated these proposals. Only segmentations that were correct at the center were included in the labeled dataset. Finally, this labeled dataset was divided into a training set (12 patients) and validation set (3 patients).

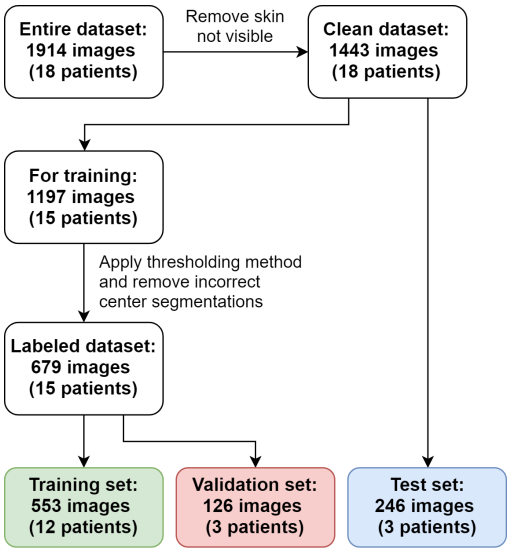

Fig. 2. Diagram of how the dataset was divided. Images in which the skin was not visible were excluded. The 18 patients were divided into 12 for training, 3 for validation, and 3 for testing. The training and validation images were labeled with the thresholding method and incorrect labels were thrown out.

B. Thresholding

Our proposed thresholding method used Otsu's method to generate thresholded images. Otsu's method [13] automatically selects the global binary threshold to maximize inter-class variance. The steps of our detection algorithm are shown in Fig. 3. Otsu's method was first applied to the entire image. Bright bubble artifacts were removed using morphological opening. The top skin edge was detected by a search starting at a pre-defined distance from the transducer. A cubic polynomial was fit to the detected top edge and any points with a residual error greater than one standard deviation were pruned.

To find the bottom edge, the image was truncated at the detected top edge and aligned at the top. Otsu's method was applied again to this truncated image. Holes in the thresholded mask were filled using morphological closing, and the skin's bottom edge was detected using a similar approach as above. For the bottom edge, a linear fit was used for pruning.

C. Deep Learning

We designed and trained a deep neural network to take as input B-mode images of the skin and output the skin's top and bottom edge locations at the center. We used a lightweight encoder-decoder architecture and a final coordinate regression layer. Our encoder-decoder was based on MobileNetV2 [14], but with fewer layers and channels. We used bilinear interpolation for upsampling and skip connections [11].

A 1x1 convolution reduced the decoder's output to two channels, representing heatmaps for the top and bottom edges. The lateral center slice was passed through a coordinate regression layer [15] to output a single scalar for each channel (top and bottom position). To segment an entire image, coordinate regression was applied to every A-line in the heatmap.

We implemented our neural network in PyTorch and trained on an NVIDIA Tesla V100 GPU. Training images were augmented by vertical shifts, horizontal flips, and random gamma correction between 0.8 and 1.25. We used a batch size of 32, learning rate of 1.0, weight decay of 1e-5, and trained for 50 epochs. As suggested in [15], we used variance regularization of the heatmaps to improve performance and encourage unimodal heatmaps.

To compare our trained neural network with the thresholding method, we manually evaluated central skin segmentations on the entire test set of 246 images in a randomized and blinded manner. Segmentations within 0.2 mm of the visualized boundary were marked as accurate. Lastly, we assigned a reason for each case where a method failed.

III. Results

An example segmentation using our thresholding method is shown in Fig. 3. An example heatmap and segmentation using the neural network is shown in Fig. 4.

Table I shows the performance of the thresholding method and the neural network for central skin segmentation on the test set. Thresholding accurately segmented 58% of the images, and the neural network accurately segmented 82%. Only

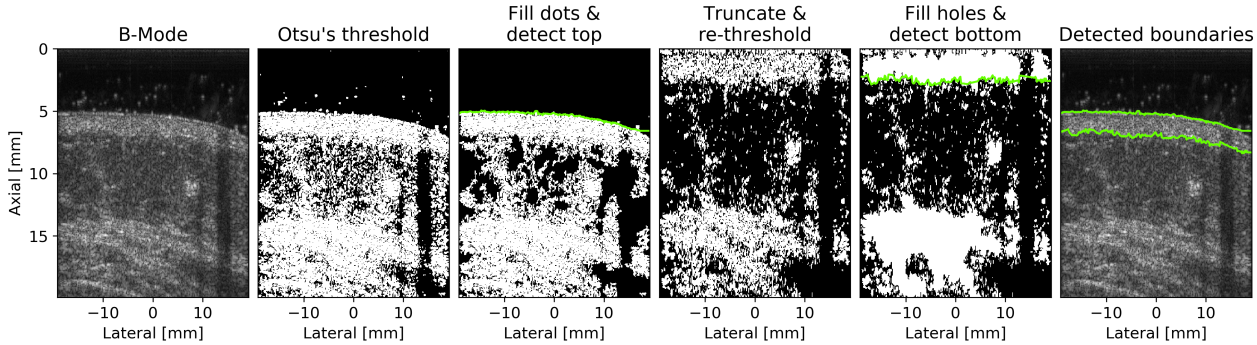

Fig. 3. The steps of our thresholding method. The input B-mode image is thresholded, and bright bubbles in the ultrasound gel are removed. The top edge is detected, and the original image is truncated according to this edge. The thresholding, filling, and detection process is repeated on the truncated image to obtain the bottom edge.

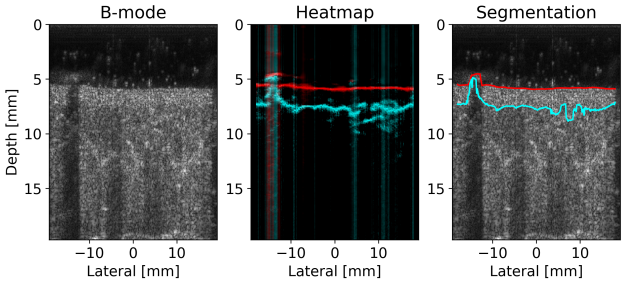

Fig. 4. Example image (left) is input to the neural network. The heatmap output (middle) shows the network's prediction of likely locations for the top and bottom edges (red and cyan respectively). Applying coordinate regression to the heatmap gives the final network segmentation (right).

2% of the images were correctly segmented by thresholding *but not* the neural network.

We manually categorized reasons for failure on the test set, shown in Table II. More than 80% of the thresholding method's failures were overestimation of the bottom boundary. The neural network failed due to overestimation less often (28 images) compared to thresholding (86 images). Both methods failed equally due to bubbles in the ultrasound gel and underestimation.

TABLE I
NUMBER OF IMAGES CORRECTLY SEGMENTED ON THE TEST SET

Method	Number	Percent
Both Methods	139	56%
Thresholding Only	5	2%
Neural Network Only	63	26%
Neither Method	39	16%
Total	246	100%

We also evaluated run times for our two methods. On a single thread of an Intel Core i7-6600U CPU (laptop processor), the thresholding method processed 40 images per second and the neural network processed 5.2 images per second. On an NVIDIA GeForce RTX 2070 GPU (consumer graphics card), the neural network processed 150 images per second.

TABLE II
FAILURE MODES OF THE TWO METHODS ON THE TEST SET

Reasons for Failure	Thresholding Failed	Network Failed
Overestimation	86	28
Bubbles in gel	13	13
Underestimation	3	3
Total	102	44

IV. DISCUSSION

The primary challenge of skin segmentation on B-mode ultrasound images is detecting the skin's distal boundary between the dermis and the subcutaneous tissue. Ideally, the dermis is hyperechoic and provides contrast against the underlying fatty and connective tissue, such as in the example Fig. 1. Shadowing artifacts, bubbles in the ultrasound gel, and poor contrast can make images difficult for automatic segmentation.

We initially tested local gradient-based methods, which have been used in the literature [8], [9]. For our images, these gradient methods were not sensitive enough to robustly detect the distal edge. Region-growing methods, such as the confidence-connected region growth (CCRG) algorithm [16], generated high quality segmentations when the image was ideal. However, CCRG was very sensitive to baseline image intensity and lacked the ability to jump over shadowing artifacts.

Our two-step thresholding method produced detailed skin segmentations while being robust to shadowing, bubble artifacts, and speckle noise. Thresholding was less sensitive to local artifacts because Otsu's method leverages the global pixel histogram to find a threshold value. As shown in Fig. 3, morphological opening and closing further reduced the impact of bubbles in the gel and dark speckles in the dermis respectively.

The most common failure mode of our thresholding method was overestimation of the bottom boundary. An example of this is shown in Fig. 5. Overestimation was frequently caused by bright subcutaneous structures. When hyperechoic tissue

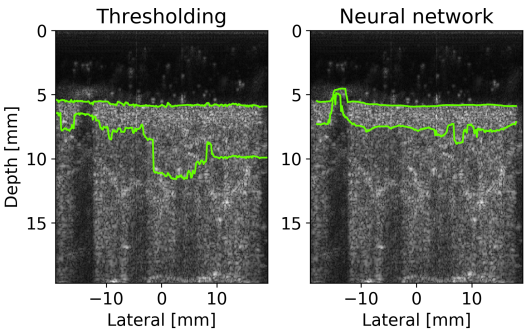

Fig. 5. Thresholding often fails due to bright subcutaneous structures. In this example, the neural network was less affected by these bright structures. However, the network was not immune to the shadowing artifact on the left.

was immediately below the dermis, the skin's distal edge became difficult to detect after thresholding.

To train a neural network, we required labeled images. We used the segmentations calculated from the thresholding method to avoid manually segmenting images. We visually evaluated the proposed segmentations and threw out images that were incorrect. This process reduced the size of our training dataset by about 40%, but it was still sufficient to train our neural network.

We manually evaluated the two methods on the test set for central segmentation. The network outperformed thresholding (82% versus 58%). Notably, only on 5 out of 246 images did the network fail but not the thresholding method. Thus, the network replicated the results of thresholding and generalized to more difficult images.

The distribution of training images was altered by keeping only images that were correctly segmented by thresholding. This selection process could have prevented the network from generalizing to more difficult images. Instead, we found that the neural network was able to correctly segment over half of the images on which thresholding failed.

An analysis of failure modes revealed that overestimation was the biggest problem for thresholding. The neural network was much less prone to this failure mode compared to the thresholding method. Fig. 5 shows a case where bright subcutaneous structures affected the thresholding method but not the neural network.

Other failure modes were bubbles in the ultrasound gel and underestimation of the bottom edge. In future clinical studies, we plan to address these issues in several ways. Using an ultrasound stacking pad instead of a layer of gel would reduce the impact of bubbles. Higher frequency imaging and time gain compensation could improve the contrast between the dermis and the subcutaneous tissue.

V. Conclusions

Skin thickness measurements are necessary to calculate skin tissue stiffness using ultrasound swear wave elastography. To allow for robust real-time elasticity imaging, we designed and compared two automated skin segmentation methods. Our thresholding method produced high quality segmentations of the skin and was relatively robust to bubbles in the gel, dark speckles in the dermis, and shadowing artifacts. However, thresholding frequently overestimated skin thickness due to bright subcutaneous tissues. To train a deep neural network, we created a training dataset using our thresholding results instead of manually labeling the images. Despite being trained only from thresholding predictions, the deep neural network generalized beyond the training set and was more robust. Our trained network outperformed the thresholding method on a held-out test set and was less likely to overestimate the distal skin boundary.

Acknowledgment

We received in-kind technical support from Siemens Medical Solutions, USA, Ultrasound Division.

References

[1] C. P. Denton and D. Khanna, "Systemic sclerosis," *The Lancet*, vol. 390, no. 10103, pp. 1685–1699, 2017.
[2] S. Aractingi and O. Chosidow, "Cutaneous graft-versus-host disease," *Archives of dermatology*, vol. 134, no. 5, pp. 602–612, 1998.
[3] S. Y. Lee, A. R. Cardones, J. Doherty, K. Nightingale, and M. Palmeri, "Preliminary results on the feasibility of using arfi/swei to assess cutaneous sclerotic diseases," *Ultrasound in medicine & biology*, vol. 41, no. 11, pp. 2806–2819, 2015.
[4] Y. Yang, F. Yan, L. Wang, X. Xiang, and Y. Tang, "Quantification of skin stiffness in patients with systemic sclerosis using real-time shear wave elastography: a preliminary study," *Clin Exp Rheumatol*, vol. 36, no. 113, pp. S118–S125, 2018.
[5] X. Zhang *et al.*, "An ultrasound surface wave technique for assessing skin and lung diseases," *Ultrasound in medicine & biology*, vol. 44, no. 2, pp. 321–331, 2018.
[6] K. Nightingale, S. McAleavey, and G. Trahey, "Shear-wave generation using acoustic radiation force: in vivo and ex vivo results," *Ultrasound in medicine & biology*, vol. 29, no. 12, pp. 1715–1723, 2003.
[7] H. Lamb, "On waves in an elastic plate," *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, vol. 93, no. 648, pp. 114–128, 1917.
[8] J.-M. Lagarde, J. George, R. Soulcié, and D. Black, "Automatic measurement of dermal thickness from b-scan ultrasound images using active contours," *Skin Research and Technology*, vol. 11, no. 2, pp. 79–90, 2005.
[9] Y. Gao *et al.*, "Automated skin segmentation in ultrasonic evaluation of skin toxicity in breast cancer radiotherapy," *Ultrasound in medicine & biology*, vol. 39, no. 11, pp. 2166–2175, 2013.
[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.
[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[12] J. R. Doherty, J. J. Dahl, and G. E. Trahey, "Harmonic tracking of acoustic radiation force-induced displacements," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 60, no. 11, pp. 2347–2358, 2013.
[13] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
[15] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "Numerical coordinate regression with convolutional neural networks," *arXiv preprint arXiv:1801.07372*, 2018.
[16] H. J. Johnson, M. McCormick, L. Ibáñez, and T. I. S. Consortium, *The ITK Software Guide*, 3rd ed., Kitware, Inc., 2013, *In press*.