2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Virtual Scanning Framework for Robotic Spinal Sonography with Automatic Real-time Recognition of Standard Views

Keyu Li, Yangxin Xu, Li Liu, and Max Q.-H. Meng*, *Fellow*, *IEEE*

***Abstract*— Ultrasound (US) imaging is widely used to assist in the diagnosis and intervention of the spine, but the manual scanning process would bring heavy physical and cognitive burdens on the sonographers. Robotic US acquisitions can provide an alternative to the standard handheld technique to reduce operator workload and avoid direct patient contact. However, the real-time interpretation of the acquired images is rarely addressed in existing robotic US systems. Therefore, we envision a robotic system that can automatically scan the spine and search for the standard views like an expert sonographer. In this work, we propose a virtual scanning framework based on real-world US data acquired by a robotic system to simulate the autonomous robotic spinal sonography, and incorporate automatic real-time recognition of the standard views of the spine based on a multi-scale fusion approach and deep convolutional neural networks. Our method can accurately classify 96.71% of the standard views of the spine in the test set, and the simulated clinical application preliminarily demonstrates the potential of our method.**

***Index Terms*— Robotic Ultrasound System, Spinal Ultrasound Imaging, Ultrasound Image Classification, Standard View Recognition.**

I. INTRODUCTION

As a safe and non-invasive medical imaging modality, ultrasound (US) is widely used to visualize the spinal anatomy to assist in the diagnosis and intervention of the spine [1]. However, since the clinician has to manually scan the patient's back, the process is usually time-consuming and imposes heavy physical and cognitive burdens on the operator. Moreover, the frontline sonographers are vulnerable to infectious diseases due to direct patient contact, especially during a pandemic such as COVID-19 [2].

In the past two decades, an increasing number of robotic systems have been developed to automatize the US imaging process [3], which can provide an alternative to the standard handheld technique. While existing works have demonstrated the potential of using robots to acquire meaningful images, the real-time interpretation of the images during the robotic acquisition is rarely addressed in previous work. As a result, the clinicians have to further process the robotically acquired data to extract useful images, such as the standard view planes that contain important information of the anatomy for diagnosis. [4] uses well-engineered features to detect the carotid landmarks in the image acquired by the robot, but the method is difficult to generalize to other organs.

To this end, an integration of real-time annotation of the acquired US images in the robotic scan can improve ease of use and reduce the cognitive burden of clinicians. In recent years, deep learning has been intensively studied and applied in many US image analysis tasks, which can eliminate the need for complex feature engineering [5]. Some attempts have been made to detect the spinal anatomical landmarks in US images with learning-based methods [6][7]. However, these methods either do not consider a realistic robotic scanning process or cannot recognize multiple standard views of the spine during the scan.

In this work, we present a virtual scanning framework to simulate the autonomous robotic spinal sonography, and incorporate AI-powered real-time recognition of three standard views of the spine, namely, the paramedian sagittal lamina view (PSL), paramedian sagittal articular process view (PSAP) and transverse spinous process view (TSP) of the lumbar spine, as shown in Fig. 1. Specifically, we develop a framework to (i) automatically plan the scanning path according to the clinical routines to cover the spinal region, adapt to patient surface and search for the desired views of the spine, (ii) simulate the 6-DOF control of a US transducer and B-mode image acquisition based on real-world US data acquired by a robotic system, and (iii) incorporate multi-scale fusion and deep convolutional neural networks (CNNs) for real-time recognition and retrieval of three standard views of the spine during the scan. Our method can be easily integrated with existing robotic systems to facilitate fully autonomous spinal sonography, and can also generalize to the robotic US imaging of other human tissues. The demonstration video can be found at `https://youtu.be/wLYTuUV3w0o`.

II. METHOD

A. Simulation Environment

In order to simulate the robotic sonography considering different patient anatomy, a total of 41 3D-US volumes that cover the L1-L5 lumbar vertebrae of 17 volunteers aged 20 to 26 are acquired using a KUKA LBR iiwa 7 R800 (KUKA Roboter GmbH, Augsburg, Germany) and a C5-1B convex transducer (Wisonic Clover diagnostic US machine, Shenzhen Wisonic Medical Technology Co., Ltd, China) mounted

This work was partially supported by National Key R&D program of China with Grant No. 2019YFB1312400, Hong Kong RGC GRF grant #14210117, Hong Kong RGC TRS grant T42-409/18-R and Hong Kong RGC GRF grant #14211420 awarded to Max Q.-H. Meng.

K. Li, Y. Xu and L. Liu are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: kyli@link.cuhk.edu.hk; yxxu@link.cuhk.edu.hk; liliu@cuhk.edu.hk).

Max Q.-H. Meng is with the Department of Electronic and Electrical Engineering of the Southern University of Science and Technology in Shenzhen, China, on leave from the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, and also with the Shenzhen Research Institute of the Chinese University of Hong Kong in Shenzhen, China (e-mail: max.meng@ieee.org).

*Corresponding author.

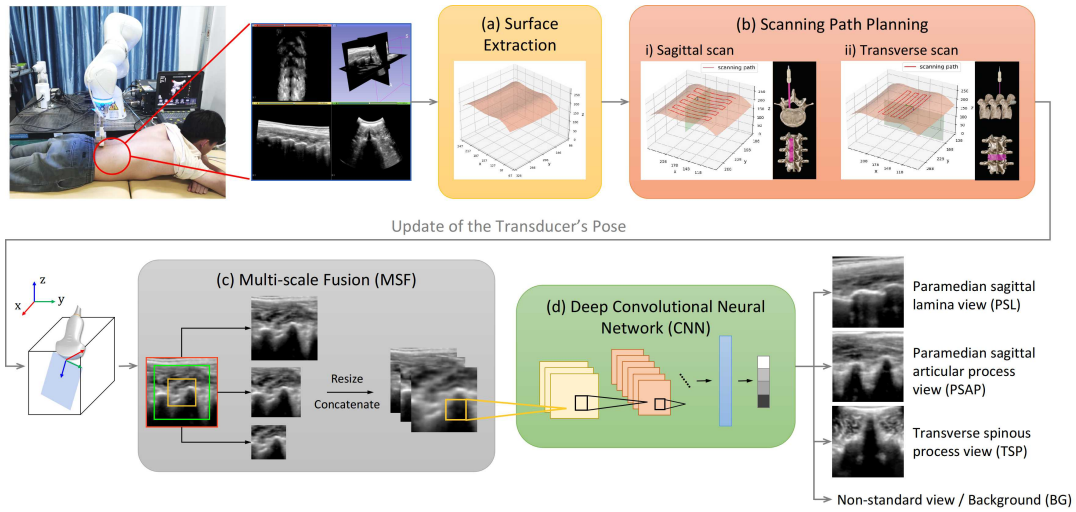


Fig. 1. Workflow of the proposed scanning framework. B-mode images of the spine acquired with a robotic system are reconstructed as 3D-US volumes and set as virtual patients in the simulation. (a) The patient surface is first extracted from the volume data and (b) the scanning path is planned according to the clinical routines to cover the spinal region and adapted to the patient surface. After the 6-DOF probe pose is determined by the scanning path, (c) a 2D-US image is acquired and multi-scale fusion is performed to enhance the location sensitivity of the (d) deep convolutional neural network (CNN) to recognize and retrieve the standard view planes of the spine.

at its end-effector, as shown in Fig. 1. The volunteers are in the prone position on a horizontal examination bed during the acquisition. The robot linearly moves the probe from the start point to the end point specified by a clinician under Cartesian impedance control, and the acquired B-mode images are reconstructed as 3D volumes, which serve as virtual patients in our simulation. The average size of the resulting US volumes of the spine is $350 \times 397 \times 274$ and the size of each voxel is $0.5 \times 0.5 \times 0.5mm^3$. In order to simulate the US scans with common 2D US transducers in clinical routines, we assume the virtual transducer to have a rectangular field of view. Once the 6-D pose of the probe is determined, a 2D US image of size 150×150 is sampled in the volume data. 33 data volumes obtained from 14 subjects are used as the training set and 8 data volumes obtained from 3 subjects are used as the test set. The standard view images and the associated probe poses are manually annotated by a medical expert. The 2D images for training and testing the standard view recognition algorithm are collected by sampling frames from the volumes. Finally, a total of 715 images (PSL: 171, PSAP: 173, TSP: 158, BG (background): 213) and 213 images (PSL: 57, PSAP: 79, TSP: 32, BG: 45) are collected for training and testing, respectively.

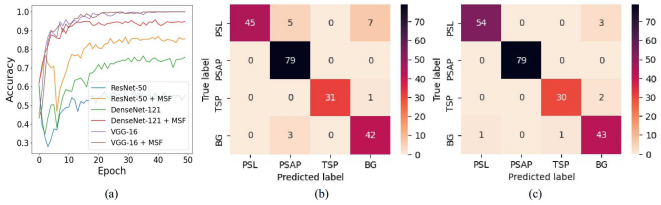
B. Surface Extraction

Since the US wave can barely penetrate the air, close contact between the transducer and the patient should be maintained to ensure sufficient acoustic coupling during the scan. In order to position the transducer to track the patient surface, we use an image intensity-based method to extract the surface of the virtual patient before planning the scanning path. To extract the surface equation $z = surface(x, y)$ for each patient V, for each pair of (x, y), we approximate the surface point as the point with the largest z-coordinate whose gray value is not zero. The estimated surface is then smoothed using a uniform filter, as shown in Fig. 1 (a). Note that this intensity-based method is only used to estimate the patient surface in our simulation. In real-world applications, the patient surface can be extracted using some external sensing devices such as an RGB-D camera [8].

C. Scanning Path Planning

Since the spinal region is too large to be scanned with a single sweep using a common transducer, we plan the scanning path in the X-Y plane with an "S" shape, containing a set of parallel lines to cover the spinal region, and then adjust it to the patient surface, as shown in Fig. 1 (b). We consider the sagittal scan and transverse scan in the workflow, as they are two commonly used axes of scan in spinal sonography [1].

In the sagittal scan (see Fig. 1 (b)(i)), the imaging plane of the US transducer is kept parallel to the longitudinal plane of the body (X-Z plane). The *paramedian sagittal lamina view* (PSL) and *paramedian sagittal articular process view* (PSAP) of the lumbar spine can be acquired during the scan. In order to ensure coverage of the standard views, a set of parallel lines perpendicular to the sagittal plane with an interval of 5 mm are generated, resulting in an S-shaped scanning path that covers the center region with 40% length and 40% width of the volume. In the transverse scan (see Fig. 1 (b)(ii)), the imaging plane is set parallel to the transverse plane of the body (Y-Z plane), and the *transverse spinous process view* (TSP) of the lumbar spine can be acquired during the scan. Since the spinous process is located in the middle of the spine, the region to be covered is smaller than that in the sagittal scan. Therefore, the scanning path are generated to cover the center region with 40% length and 20% width of the virtual patient. In real-world applications, the region to be covered can be estimated using surface



(a) (b) (c)

Fig. 2. (a) Learning curves of different classification methods (training accuracy). (b) Confusion matrix of VGG-16 on the test images. (c) Confusion matrix of VGG-16+MSF on the test images.

TABLE I
CLASSIFICATION SCORES OF DIFFERENT MODELS

Model	Accuracy	Precision	Recall	F1-score
ResNet-50	61.03	42.44	49.47	45.49
ResNet-50+MSF	91.55	90.40	88.09	88.90
DenseNet-121	82.63	82.59	82.66	80.71
DenseNet-121+MSF	94.37	93.54	92.55	93.00
VGG-16	92.49	93.70	92.29	92.56
VGG-16+MSF	**96.71**	**96.13**	**96.01**	**96.03**

landmarks of the patient, and the same path planning method can be applied.

After the scanning path is planned in the horizontal plane for coverage of the region of interest, the z-coordinate of each point is calculated according to the extracted surface in Section II.B to make the virtual transducer compliant with the patient surface. Therefore, the 6-DOF pose of the probe can be determined.

D. Standard View Recognition

1) Multi-scale Fusion: The spatial location of anatomical landmarks usually plays an important role in the medical image analysis tasks. Some methods incorporated multi-scale patches, modified the network architecture, or directly added explicit location information in the network to improve the segmentation performance on the brain MRI images [9]. In our task, the standard views of the spine are also highly relevant with the spatial location of the pixels. Therefore, we adopt a multi-scale fusion (MSF) approach to make the classification network more sensitive to location features. A larger-scale image can usually capture more context information, but the localization accuracy of the anatomic landmarks will be decreased, while a smaller-scale image can locate the landmarks more accurately but may not contain enough information for classification of the view. Therefore, in order to combine the advantages of different scales, we extract three different scales of the original US image (100%, 75% and 50% of the original size), resize them to the same size (64×64) and accumulate them as different channels, as shown in Fig. 1 (c). Note that this fusion approach will not increase the overall complexity of the networks.

2) Deep Neural Network Fine-tuning: Three state-of-the-art classification networks, i.e., ResNet-50 [10], DenseNet-121 [11] and VGG-16 [12] are used as our basic classifiers. We first forgo the top fully connected layer of each model, and initialize the networks with publicly available weights pre-trained on the ImageNet dataset [13]. Then, a fully connected layer with 4 outputs corresponding to 3 standard views and the background is added to each network and initialized with random weights. Each network is fine-tuned on our training data using stochastic gradient descent and cross-entropy loss with a batch size of 16 for 50 epochs to achieve stable performance. For the ResNet-50 and DenseNet-121 models, the learning rate is initialized as 0.01 and reduced by 0.2 when the validation loss stagnates for 3 epochs. For the VGG-16 models, as the network is deeper and harder to train, we use a smaller learning rate of 0.0001. The learning curves of different networks are shown in Fig. 2 (a). After fine-tuning, the VGG-16 models achieves the highest classification accuracy of 100% on the training set. For the other two CNN architectures, the MSF approach improves the training accuracy by a large margin.

III. EXPERIMENTS

A. Standard View Classification

In order to quantitatively assess the performance of our approach, we first evaluate the accuracy of standard view classification on the test set with different methods, as reported in Table I. Among all the models, the VGG-16+MSF achieves the best performance in terms of classification accuracy, precision, recall and F1-score. It can be observed that the MSF approach can effectively improve the classification performance of all three network models without increasing the complexity and number of parameters. Furthermore, we present the confusion matrices on the test data for VGG-16 and VGG-16+MSF in Fig. 2 (b) and (c), for a comparison. As shown in Fig. 2 (b), nearly all the PSAP and TSP planes are classified accurately, while the PSL and background images are occasionally misclassified by the VGG-16 model. However, the VGG-16+MSF model can better classify the standard and non-standard views, which shows that our location sensitive approach can yield better results in the standard view recognition task.

B. Simulated Scan with Real-time Recognition and Retrieval of Standard Views

To further compare the methods with and without MSF, and preliminarily demonstrate the effectiveness of our method in real-world US scans, we apply our virtual scanning framework to the robotically acquired volumetric data of 3 unseen test subjects using the VGG-16-based models. Some snapshots of the virtual scan performed on a test subject are shown in Fig. 3. The recognition threshold for the standard views is set as 50%. The acquired images are annotated by a human expert for comparison. It can be seen that the VGG-16 model occasionally makes incorrect predictions on the acquired images, while the VGG-16+MSF model can accurately recognize the standard views during the scan. After the scan, the standard view images with the highest recognition probability are retrieved by our method and compared with those acquired by a human expert in Fig. 4. All the standard view images retrieved by our method are

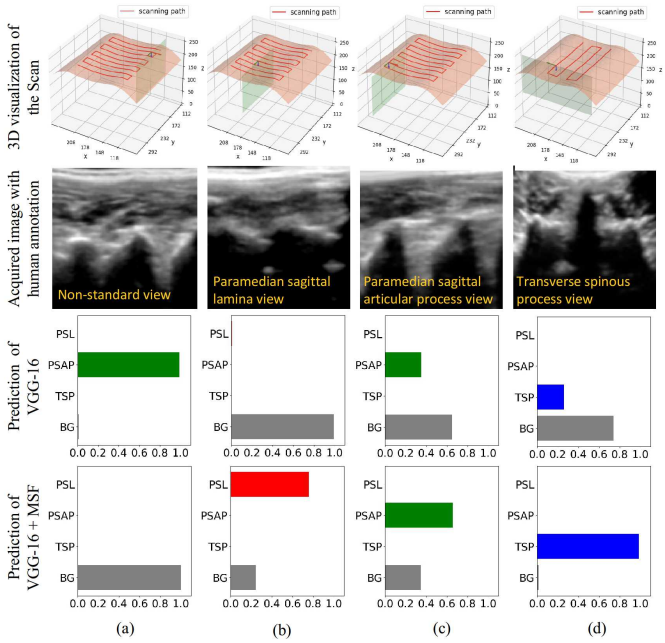


Fig. 3. Snapshots of the virtual sagittal scan (a)(b)(c) and transverse scan (d) of the spine. The first and second rows show the 3D visualization of the scan and the acquired B-mode images with the recognition results by a human expert (yellow). The third and fourth rows show the prediction results of the VGG-16 and VGG-16+MSF models, respectively.

similar to those acquired by the human expert. However, the PSAP and TSP views retrieved with the VGG-16 model are slightly deviated from the image center compared with those retrieved with VGG-16+MSF. This demonstrates that the MSF-based method can make the network more sensitive to the location features and extract more accurate standard views like a sonographer.

IV. CONCLUSIONS

In this paper, we present a framework to simulate the autonomous robotic spinal sonography with automatic real-time recognition of multiple standard views of the spine. The scanning path is planned to simulate the sagittal and transverse scans in clinical routines and follow the patient surface. A multi-scale fusion based deep learning approach is presented to recognize the standard views in real time during the scan. The proposed framework can be implemented as a plug-in module and easily integrated in existing robotic systems to enable fast and autonomous US imaging. For future application of the method on real patients in a clinical setting, an ethical approval would be needed.

REFERENCES

[1] M. K. Karmakar and K. J. Chin, *Spinal Sonography and Applications of Ultrasound for Central Neuraxial Blocks*. New York, NY: McGraw-Hill Education, 2017. [Online]. Available: accessanesthesiology.mhmedical.com/content.aspx?aid=1141735352

[2] A. Kaminski, A. Payne, S. Roemer, D. Ignatowski, and B. K. Khandheria, "Answering to the call of critically ill patients: Limiting sonographer exposure to covid-19 with focused protocols," *Journal of the American Society of Echocardiography*, 2020.

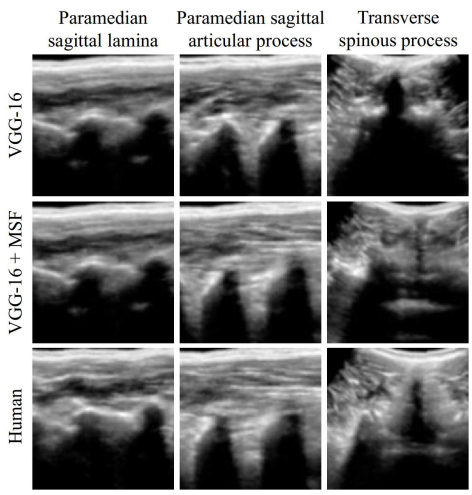


Fig. 4. Standard view images retrieved by our virtual robotic scanning framework based on VGG-16 (first row) and VGG-16+MSF (second row), and corresponding images acquired by a human expert (third row).

[3] K. Li, Y. Xu, and M. Q.-H. Meng, "An overview of systems and techniques for autonomous robotic ultrasound acquisitions," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, no. 2, pp. 510–524, 2021.

[4] R. Nakadate, J. Solis, A. Takanishi, E. Minagawa, M. Sugawara, and K. Niki, "Implementation of an automatic scanning and detection algorithm for the carotid artery by an assisted-robotic measurement system," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 313–318.

[5] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep learning in medical ultrasound analysis: a review," *Engineering*, 2019.

[6] M. Tirindelli, M. Victorova, J. Esteban, S. T. Kim, D. Navarro-Alarcon, Y. P. Zheng, and N. Navab, "Force-ultrasound fusion: Bringing spine robotic-us to the next" level"," *arXiv preprint arXiv:2002.11404*, 2020.

[7] K. Li, J. Wang, Y. Xu, H. Qin, D. Liu, L. Liu, and M. Q.-H. Meng, "Autonomous navigation of an ultrasound probe towards standard scan planes with deep reinforcement learning," *arXiv preprint arXiv:2103.00718*, 2021.

[8] C. Hennersperger, B. Fuerst, S. Virga, O. Zettinig, B. Frisch, T. Neff, and N. Navab, "Towards mri-based autonomous robotic us acquisitions: a first feasibility study," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 538–548, 2016.

[9] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. W. van Uden, C. I. Sanchez, G. Litjens, F.-E. de Leeuw, B. van Ginneken, E. Marchiori, and B. Platel, "Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities," *Scientific Reports*, vol. 7, no. 1, pp. 1–12, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.