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

Deep learning in spatiotemporal filtering for super-resolution ultrasound imaging

Katherine Brown, Kenneth Hoyt *

Department of Bioengineering, University of Texas at Dallas, Richardson, TX, USA
* kenneth.hoyt@utdallas.edu

***Abstract*—Super-resolution ultrasound (SR-US) imaging shows great promise as a clinical technique that can improve ultrasound (US) resolution by an order of magnitude. Current algorithms for SR-US suffer from high complexity and long computation times, precluding real-time imaging. Neural networks can be viewed as general function approximators and can process images at frame rates suitable for a real-time application. The goal of this study was to evaluate the effectiveness of deep networks to learn algorithms for tissue signal suppression while improving performance of SR-US for visualization of microvascular networks. In this study, deep 3D convolutional neural networks (3DCNNs) were chosen to perform spatiotemporal filtering to suppress the tissue signal and perform microbubble (MB) segmentation in place of conventional signal processing methods, e.g. singular value decomposition (SVD), singular value filtering (SVF), or difference filtering (DIF). For each method, a 3DCNN was trained with the respective conventional signal processing algorithm as ground truth. Three 3DCNN architectures with 6 to 7 layers and an input size of 9 x 9 x 9 pixels were evaluated. *In vivo* data was collected from a cancer-bearing murine model and images were captured with a clinical US scanner equipped with a 15L8 linear array transducer. *In vivo* data were used to train the networks and testing was based on an *in vivo* dataset not used in training. The deep networks reached testing accuracies of 97.1% for the DIF implementation with promising performance improvements. The average processing frame rate for *in vivo* images was 50 Hz with graphical processing unit (GPU) acceleration. Deep learning shows potential for effective spatiotemporal filtering, improving performance of SR-US towards a real-time imaging modality.**

Keywords—contrast agents, contrast-enhanced ultrasound, deep learning, microbubbles, super-resolution ultrasound.

I. INTRODUCTION

Super-resolution ultrasound imaging (SR-US) and ultrasound (US) localization microscopy hold promise as imaging techniques that improve resolution of microvasculature by an order of magnitude [1]–[3]. In these imaging techniques, as well as in Doppler US imaging, spatiotemporal filtering is often used to reduce the tissue clutter signal [4]. Spatiotemporal filters based on principle component analysis (PCA) techniques, e.g. singular value decomposition (SVD) and singular value filtering (SVF), are effective yet computationally intensive, which limits their use in real-time applications.

The real-time aspect of US imaging is important in clinical procedures such as the guidance of needles during biopsies and drug delivery [5], [6]. The spatiotemporal filtering step in SR-US contributes to the high computational burden of the method that requires off-line processing. With its ability to visualize angiogenic networks, which are biomarkers of cancer, a real-time SR-US would have broad clinical potential.

Multilayer feedforward neural networks are universal function approximators [7]. Their precision is limited by the number of hidden nodes in the network and the amount of data used in training. Additionally, when neural networks are deployed on a graphics processing unit (GPU), they can process images at hundreds of frames per second. Thus, their use in medical imaging is growing.

The goal of this study was to evaluate the effectiveness of a deep 3D convolutional neural network (3DCNN) to implement spatiotemporal filters in the context of SR-US processing. Three 3DCNN architectures of increasing size were evaluated as approximators of three spatiotemporal filters, namely, a finite impulse response (FIR) difference filter (DIF), SVD and SVF. Contrast-enhanced US (CEUS) images from *in vivo* murine breast cancer tumors were used to train deep networks to implement each filter. Both the accuracy of the deep networks in spatiotemporal filtering, and the accuracy of the final SR-US image based on the deep network implementation as compared to conventional filtering were considered in the assessment of results.

II. MATERIALS AND METHODS

A. Ultrasound Imaging

CEUS imaging was performed with a clinical US scanner (Acuson Sequoia 512, Siemens Healthcare, Mountain View, CA) equipped with a 15L8 linear array transducer in a contrast imaging mode at 14 MHz. A low mechanical index (MI) was used to minimize any microbubble (MB) contrast agent disruption. The frame rate for image collection was 15 Hz.

Female athymic nude mice were implanted with 2 million human breast cancer cells (MDA-MB-231, ATCC, Manassas, VA). These mice were imaged after about 4 weeks of tumor growth. A bolus of MB contrast agent (2.5 x 10^7 MBs in 60 μL saline; Definity, Lantheus Medical Imaging, N Billerica, MA)

This study was supported by National Institutes of Health (NIH) grants K25EB017222, R01EB025841, and Cancer Prevention Research Institute of Texas (CPRIT) grant RP180670.

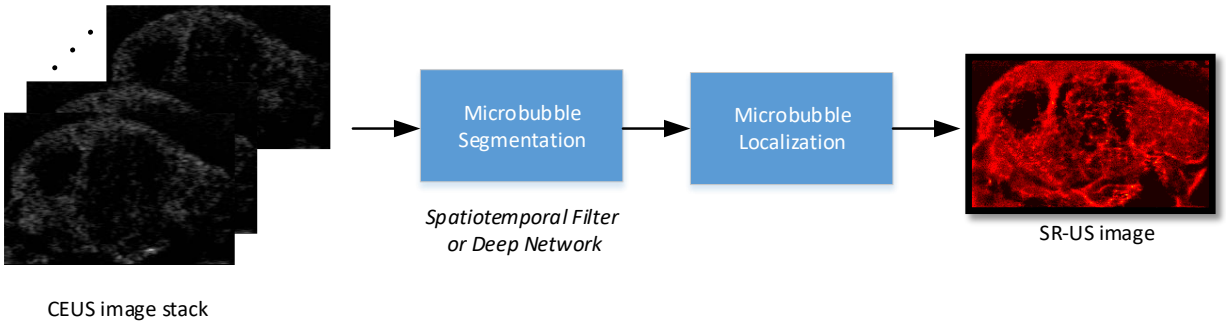

Fig. 1: Diagram of super-resolution ultrasound (SR-US) image processing flow. Microbubble (MB) segmentation from a sequence of contrast-enhanced ultrasound (CEUS) images is done with conventional spatiotemporal filtering or by prediction of a deep network.

was injected via a tail vein catheter in anesthetized subjects. Image sequences of 9000 frames were acquired over 10 minutes from five subjects.

B. SR-US Image Processing

An image processing flowchart for this study is depicted in Fig. 1. An input CEUS image stack is processed with a tissue clutter suppression method to segment MBs from the background tissue signal. The ground truth method for segmentation was a conventional implementation of the DIF, SVF and SVD spatiotemporal filters in MATLAB software (Mathworks Inc., Natick, MA). For the deep learning method, MB segmentation was performed by prediction of the trained deep network. Following segmentation, the center of each MB was localized by finding the centroid of an area around the MB. The accumulation of the locations of all MB from all the input frames results in a final SR-US image.

The DIF filter implemented was a FIR-based design that amounts to frame subtraction. This filter applied to successive frames of images in the image stack. SVD processing was implemented as described previously by our group [8]. Briefly, SVD was computed on the entire image stack and the low singular values representing the stationary tissue signal were removed. The number of singular values to be removed was determined experimentally by finding the value that maximized the contrast-to-noise ratio (CNR). SVF was implemented as described by Mauldin et al. [9]. In brief, singular values were determined for the principal component analysis (PCA) basis functions and reassembled with weighting values computed from the singular value spectrum. The first PCA basis functions with the highest singular values represent stationary tissue signal and were rejected. The image stacks were processed piecewise in matrices of 5 x 5 x 9000 pixels (5x5 in the spatial domain, 9000 frames), sliding across the image to its full spatial extent (on average 260 x 170 pixels). Precise MB centers in the filtered image stacks were obtained by finding the center of area of isolated MBs. The aggregation of these localizations formed the final SR-US image.

C. Proposed 3DCNN Architecture

The 3DCNN architectures used in this study had a patch-based architecture with three or four convolutional layers and one fully connected layer. A 9 x 9 x 9 pixel patch was used as the input data size. The patch encompassed the full point spread function (PSF) of the MB in the spatial domain of the US system and included 9 sequential frames in the slow-time dimension to capture MB motion. The three deep network architectures studied, A1, A2 and A3, are summarized in Table I. The architectures were chosen in such a way that the number of hidden nodes is increased by increasing the number of convolution layers (e.g. between A1 with 3 convolution layers and A2 with 4 convolution layers) as well as by increasing the number of features per convolution layer (e.g. between A2 and A3). Batch normalization was employed after each convolution layer to reduce training time. Each network was implemented in MATLAB.

D. Training, Validation and Testing

Training of the networks was performed with *in vivo* datasets of 2 million total images with an initial learning rate of 0.0003. The Adam optimizer was used with a mini-batch size of 512 and a validation size of 2,000 images. Training proceeded for 20 epochs or until the validation loss was higher than the previous lowest value for 10 iterations, whichever occurred first.

Labels for pixel patches for training each of the three reference spatiotemporal filters on networks A1, A2 and A3 were prepared by the appropriate conventional implementation

Table I: Deep 3-dimensional convolutional neural network (3DCNN) architectures A1, A2, and A3, showing a 9 x 9 x 9 input patch size, 3 or 4 convolutional layers each using a 3 x 3 x 3 kernel followed by a single fully connect layer and having two output states.

Type	Input	Kernel	Features			Output
			A1	A2	A3	
Input	9 x 9 x 9					9 x 9 x 9
Convolution	9 x 9 x 9	3 x 3 x 3	8	8	16	7 x 7 x 7
Convolution	7 x 7 x 7	3 x 3 x 3	16	16	32	5 x 5 x 5
Convolution	5 x 5 x 5	3 x 3 x 3	16	16	32	3 x 3 x 3
Convolution	5 x 5 x 5	3 x 3 x 3	-	32	64	3 x 3 x 3
Fully connected	3 x 3x 3	1 x 1 x 1	128	128	128	2 x 1

of each filter as the ground truth. Patches were a 50/50 mixture of MB-present or MB-not-present and were presented to the network in a random order. Validation was performed using the leave-one-out method in which one *in vivo* dataset was not used in training and reserved for validation.

Training performance was measured with accuracy, sensitivity, and specificity on a dataset not used in training. The computation of accuracy was the number of correct predictions divided by the total number of predictions. Sensitivity was computed as the number of classifications of MB-present divided by the number of MB-present in the test dataset as determined by the ground truth method. Specificity was computed as the number of classifications of MB-not-present divided by the number of patches in the test dataset labeled as MB-not-present.

After training, the deep networks were used to create SR-US images based on the prediction of MB segmentation for a representative *in vivo* dataset that had not been used in training. Potential MB locations were determined by thresholding the images and a pixel patch created around that location to create input patches for the deep network. If the prediction indicated a MB was present, pixel values from the original image in a 9 x 9 pixel patch at that location were copied into a working frame. The working frames were processed in the same manner as reference frames for the remaining steps of SR-US processing.

To assess the entire SR-US result of the deep learning methods of spatiotemporal filtering, a comparison was made of the SR-US image stack formed as the precursor to aggregation of localizations in forming the final image with that formed in the reference filtering methods. Accuracy was calculated as the number of correct pixel matches following localizations of MB with the deep learning method versus the reference method.

III. Results

The performance of the ground truth methods for the DIF, SVF and SVD on the representative *in vivo* test dataset run on a single central processing unit (CPU) is summarized in Table II. The DIF filter had the fastest processing time, completing in just seconds, while SVF and SVD took several minutes. The number of MBs localized was highest for SVD and lowest for SVF. SR-US images formed with the filters are shown in Fig. 2. It appears that the greater number of MB localized by SVD contribute to a higher quality image which has greater details and texture visible in the microvasculature of the tumor.

The accuracy, sensitivity, and specificity of MB segmentation by the deep networks trained on three filters and three architectures is summarized in Table III. Across all architectures, the DIF spatiotemporal filter had the highest accuracy, followed by SVF and SVD. The highest overall accuracy was for the DIF filter at 97.1% with architecture A2. The highest accuracy achieved for the SVF filter was 89.4% with architecture A1. The highest accuracy achieved for SVD was 83.4% with architecture with the greatest number of hidden nodes, A3, however the result had the lowest overall of the three filter realizations. For a given filter implementation, the highest sensitivity and highest specificity did not both match the architecture with the highest accuracy, but differences were only slight.

TABLE II. Performance results for spatiotemporal filters with conventional method on single central processing unit (CPU).

Filter	CPU time (sec)	MB localized
DIF	2.3	433,758
SVF	259	352,506
SVD	263	480,425

TABLE III. Microbubble (MB) segmentation results for filtering with deep 3DCNN networks for three spatiotemporal filters, a difference filter (DIF), singular value decomposition (SVD) and singular value filtering (SVF), assessed with three architectures A1, A2 and A3. The highest values for accuracy, sensitivity and specificity for each of the different spatiotemporal filters are displayed in bold.

	Accuracy	Sensitivity	Specificity
DIF			
A1	97.0%	**95.7**%	98.4%
A2	**97.1**%	95.5%	**98.6**%
A3	95.5%	92.8%	98.9%
SVF			
A1	**89.4**%	**88.3**%	90.6%
A2	89.3%	87.4%	**91.2**%
A3	89.1%	88.3%	89.9%
SVD			
A1	82.7%	83.1%	**81.6**%
A2	82.3%	85.7%	80.1%
A3	**83.4**%	**85.9**%	80.9%

TABLE IV. MB Localization results for SR-US images based on deep networks implementations of three spatiotemporal filters in Fig. 2. The DIF used the A2 architecture, SVF the A1 and SVD the A3. The highest values for accuracy and MB localized are in bold.

	Accuracy	MB localized
DIF (A2)	99.7%	247,893
SVF (A1)	**99.8%**	246,995
SVD (A3)	99.7%	**249,817**

Architecture selection was informed by early studies. We observed in these studies that an accuracy above 90% could not be achieved with a data input size smaller than 9 x 9 x 9 pixels. The architecture of the layers was varied during these studies and it was found that A1 converged well. Additionally, we observed that realizations of SVD decreased in accuracy as features of the deep network architecture increased beyond A3 (e.g. 4 convolutional layers of size 16, 32, 64, 128 features, and 16, 32, 64, 128, 256 features).

The *in vivo* SR-US images formed by the reference methods for each spatiotemporal filter are shown in Fig. 2 along with those formed with the MB segmentation prediction results of realizations of each filter with a trained deep 3DCNN network. The network architecture having the highest accuracy for each spatiotemporal filter was used in forming the SR-US images. MB localization accuracy for each figure pair is shown in Table IV. This data reveals nearly identical results, in accordance with the similar level of details visible in the SR-US images based

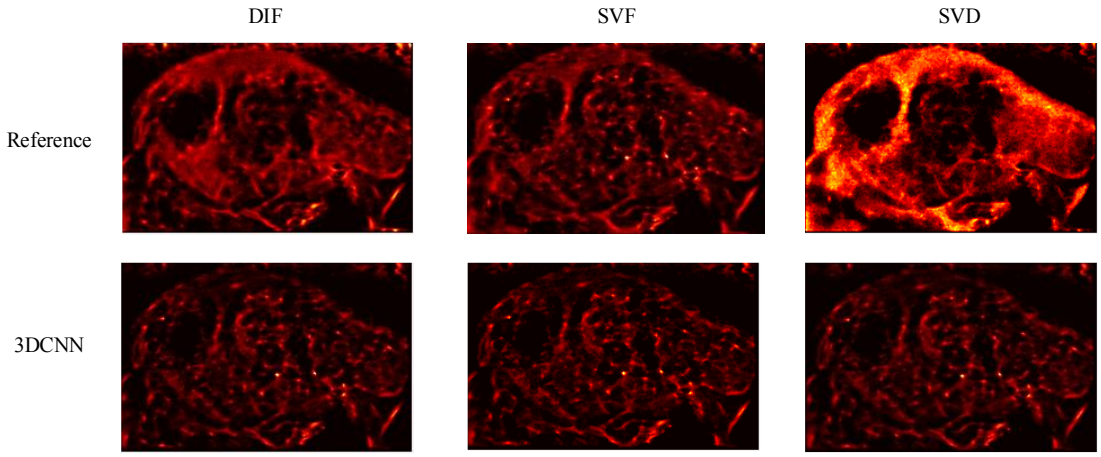

Fig. 2: SR-US images of *in vivo* murine tumors processed with three different spatiotemporal filters. The first row of reference images was produced with conventional signal processing methods for tissue clutter rejection with a difference filter (DIF), singular value filtering (SVF) and singular value decomposition (SVD). The second row of images was generated using a deep 3D convolutional neural network (3DCNN) to perform MB segmentation and reject the tissue signal after training with one of the three different filter methods.

on deep learning. The SVF image based on deep learning most closely matches its reference method. The DIF images also are quite similar. The SVD images are the least similar. Interestingly, the SVD image based on deep learning, while least like its reference image is quite similar to the SVF reference. The average processing frame rate for *in vivo* images was 50 Hz with GPU acceleration

IV. Discussion

The results of this study revealed that more complex spatiotemporal filters are difficult to realize with a deep network as measured by the accuracy of MB segmentation. Using computer processing time as a proxy for complexity, DIF is relatively simple while SVF and SVD are significantly more complex. Thus, it was expected that the simple FIR DIF filter would be easier to realize than the more complex SVF and SVD filters, and this was borne out in the study. While the computer processing time was nearly identical for SVF and SVD, suggesting similar complexity, the deep networks were slightly more effective in realizing SVF. This may be due to the limited reception field in the first layers of deep networks which is a better match to the method of SVF processing which works on 5 x 5 pixel spatial dimensions at a time. SVD processing works on the full image spatial dimensions and would presumably need a larger receptive field to capture its operation.

Theoretically, improving the accuracy of a deep network to realize SVD might be achieved with a deeper network. The residual network architecture proposed by He *et al.* improves training and allows deeper networks [10]. An implementation of a residual network to realize the SVD spatiotemporal filter will be explored in future work.

V. Conclusion

The results of this study show that deep learning is effective for implementing spatiotemporal filters of moderate complexity. The performance of MB segmentation by a deep network on *in vivo* CEUS datasets is promising for a real-time SR-US applications. Further refinements in network architecture may allow improvements in the realization of complex spatiotemporal filters.

References

[1] J. Foiret, H. Zhang, T. Ilovitsh, L. Mahakian, S. Tam, and K. W. Ferrara, "Ultrasound localization microscopy to image and assess microvasculature in a rat kidney," *Sci. Rep.*, vol. 7, no. 1, p. 13662, 2017.

[2] K. Christensen-Jeffries, R. J. Browning, M.-X. Tang, C. Dunsby, and R. J. Eckersley, "In vivo acoustic super-resolution and super-resolved velocity mapping using microbubbles," *IEEE Trans. Med. Imaging*, vol. 34, no. 2, pp. 433–440, 2015.

[3] C. Errico *et al.*, "Ultrafast ultrasound localization microscopy for deep super-resolution vascular imaging," *Nature*, vol. 527, no. 7579, pp. 499–502, 2015.

[4] C. Demené *et al.*, "Spatiotemporal clutter filtering of ultrafast ultrasound data highly increases Doppler and fUltrasound sensitivity," *IEEE Trans. Med. Imaging*, vol. 34, no. 11, pp. 2271–2285, 2015.

[5] K. J. Chin, A. Perlas, V. W. S. Chan, and R. Brull, "Needle visualization in ultrasound-guided regional anesthesia: Challenges and solutions," *Reg. Anesth. Pain Med.*, vol. 33, no. 6, pp. 532–544, 2008.

[6] König Katharina, Scheipers Ulrich, Pesavento Andreas, Lorenz Andreas, Ermert Helmut, and Senge Theodor, "Initial experiences with real-time elastography guided biopsies of the prostate," *J. Urol.*, vol. 174, no. 1, pp. 115–117, 2005.

[7] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[8] D. Ghosh *et al.*, "Monitoring early tumor response to vascular targeted therapy using super-resolution ultrasound imaging," *Proc IEEE Ultrason Symp*, 1–4, 2017.

[9] F. W. Mauldin, D. Lin, and J. A. Hossack, "The singular value filter: a general filter design strategy for PCA-based signal separation in medical ultrasound imaging," *IEEE Trans. Med. Imaging*, vol. 30, no. 11, pp. 1951–1964, 2011.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, pp. 770–778, 2016.