International Journal of Computer Assisted Radiology and Surgery (2019) 14:775–783
https://doi.org/10.1007/s11548-019-01934-0

ORIGINAL ARTICLE

Automatic segmentation of bone surfaces from ultrasound using a filter-layer-guided CNN

Ahmed Z. Alsinan[1] · Vishal M. Patel[1] · Ilker Hacihaliloglu[2,3]

Received: 30 August 2018 / Accepted: 5 March 2019 / Published online: 13 March 2019

Abstract

Purpose Ultrasound (US) provides real-time, two-/three-dimensional safe imaging. Due to these capabilities, it is considered a safe alternative to intra-operative fluoroscopy in various computer-assisted orthopedic surgery (CAOS) procedures. However, interpretation of the collected bone US data is difficult due to high levels of noise, various imaging artifacts, and bone surfaces response appearing several millimeters (mm) in thickness. For US-guided CAOS procedures, it is an essential objective to have a segmentation mechanism, that is both robust and computationally inexpensive.

Method In this paper, we present our development of a convolutional neural network-based technique for segmentation of bone surfaces from in vivo US scans. The novelty of our proposed design is that it utilizes fusion of feature maps and employs multi-modal images to abate sensitivity to variations caused by imaging artifacts and low intensity bone boundaries. B-mode US images, and their corresponding local phase filtered images are used as multi-modal inputs for the proposed fusion network. Different fusion architectures are investigated for fusing the B-mode US image and the local phase features.

Results The proposed methods was quantitatively and qualitatively evaluated on 546 in vivo scans by scanning 14 healthy subjects. We achieved an average *F*-score above 95% with an average bone surface localization error of 0.2 mm. The reported results are statistically significant compared to state-of-the-art.

Conclusions Reported accurate and robust segmentation results make the proposed method promising in CAOS applications. Further extensive validations are required in order to fully understand the clinical utility of the proposed method.

Keywords Orthopedic surgery · Segmentation · Ultrasound · Bone · Deep learning

Introduction

Orthopedic procedures have been a prominent solution in treating interminable pain and disabilities, due to musculoskeletal diseases, e.g. osteoarthritis, spinal conditions, osteoporosis, and trauma injuries. In 1990, the World Health Organization reported 1.7 million hip fractures and projected the figure to increase to 6 million by 2050 [13]. Osteoporosis and related fracture treatments costs were estimated at $19.1 billion in 2004. Moreover, spine related injuries for the years 2002–2004 were estimated at $193.9 billion [13,17].

The need for high precision in the mentioned procedures is evidently required in order to minimize the intra- and post-operative complications. Computer-assisted orthopedic surgery (CAOS) systems enable higher precision by providing surgeons, intra-operatively, real-time feedback for guidance during the procedure. Imaging is one of the most important components of any CAOS system. The standard intra-operative imaging modality in CAOS is two-/three-dimensional (2D/3D) fluoroscopy. Navigation during surgery is difficult using 2D fluoroscopy imaging due to limited 3D information available in 2D scans. 3D fluoroscopy systems provide a solution to this problem, however, they are twice as expensive and currently are not as widely employed as their 2D alternative. Most importantly, both of these modal-

✉ Ilker Hacihaliloglu
ilker.hac@soe.rutgers.edu

Ahmed Z. Alsinan
ahmed.alsinan@rutgers.edu

Vishal M. Patel
vishal.m.patel@rutgers.edu

[1] Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, USA

[2] Department of Biomedical Engineering, Rutgers University, Piscataway, NJ, USA

[3] Rutgers University Robert Wood Johnson Medical School, New Brunswick, NJ, USA

ities operate with ionizing radiation which causes important safety concerns to both surgical team and the patient. In order to provide a safe intra-operative imaging alternative, ultrasound (US) has been incorporated into various CAOS systems. US provides real-time, non-radiation based 2D/3D imaging. However, low signal-to-noise ratio (SNR), imaging artifacts, limited field of view and being a user operated imaging modality have hindered the widespread use of US in CAOS procedures. Furthermore, the beamwidth in elevation direction strongly influences the bone surface response profile making the bone boundaries appear several mm in thickness [6]. In order to provide a solution to these difficulties, focus has been given to develop automated US bone segmentation and enhancement methods. Accurate segmentation is also very important for intra-operative registration which is an essential step for any CAOS system based on US.

Recently, methods based on deep learning have been proposed for the segmentation of bone surfaces from US data. In [16], the authors modify a deep learning network architecture, termed U-net and proposed in [15], for segmentation of bone surfaces from US data. Localization accuracy was not reported, however, the recall and precision rates for the proposed method were 0.87. In [1], a similar architecture based on U-net was investigated for segmenting vertebra bone surfaces. Reported precision and recall rates were 0.88 and 0.94, respectively. Bone surface localization accuracy was not investigated. Low-quality bone surfaces were excluded from the validation and testing procedure. In [20], U-net was utilized in the development of a classification network that simultaneously performs bone segmentation with a concatenated input. In [19], a network architecture based on the fully convolutional network architecture (FCN) [12] was investigated. The reported mean recall, precision, F1 score, accuracy and specificity were, respectively, 62%, 64%, 57%, 80% and 83%. The success of the deep learning methods is dependent on: (1) number of scans used for training, (2) anatomical variations (such as type of bone surfaces) present in the training data, (3) quality of the collected US data. High-quality bone US data is usually defined as a high-intensity bone response profile, corresponding to the bone surface, followed by shadow region.

In this study, we propose a convolutional neural network (CNN)-based approach for automated bone segmentation. Based on [4,5], and inspired by [15,18], we introduce a new CNN design that can perform accurate segmentation of bone structures in US images. We show that the performance of the CNN segmentation methods improves if the training is performed on data that incorporates local phase image features in addition to the intensity features of the B-mode US data. This novel approach attempts to alleviate the shortcomings of unimodal designs, and their susceptibility to noise and other imaging artifacts. Validation is performed on 546 in vivo scans obtained from 14 healthy subjects by scanning various bone surfaces. We also include quantitative and qualitative evaluation results on data sets obtained from a different US imaging platform which was not used during the training of the proposed method. Obtained results are also compared against U-net [15] trained using (1) B-mode US data only and (2) B-mode and local phase image features.

Methods

Data acquisition

After obtaining the institutional review board (IRB) approval, a total of 415 B-mode US images (categorized into three groups of bone structures: radius, femur, and tibia) from twelve healthy subjects, were collected using Sonix-Touch US machine (Analogic Corporation, Peabody, MA, USA) with a 2D C5-2/60 curvilinear probe and L14-5 linear probe. Depth settings and image resolutions varied between 3–8 cm and 0.12–0.19 mm, respectively. In addition, a second dataset of 131 images were obtained using a hand-held wireless ultrasound probe (Clarius C3, Clarius Mobile Health Corporation, BC, Canada) from two volunteers. This dataset was considered for validation/testing purposes only. All the bone surfaces were manually segmented by an expert ultrasonographer.

Local phase image features

The local phase image feature extraction is based on the computation of three different phase image features [4]. The combination of three different image phase features provides a more compact and robust enhancement. Next, we explain how these three features are extracted and combined.

Hacihaliloglu et al. [8] proposed a tensor-based phase feature descriptor called local phase tensor image feature ($\mathrm{LPT}(x, y)$). $\mathrm{LPT}(x, y)$ is obtained from B-mode US image, $\mathrm{US}(x, y)$, using even (T_{even}) and odd filter responses (T_{odd}) which represent the symmetric and asymmetric features found in $\mathrm{US}(x, y)$. T_{even} and T_{odd} filters are constructed using a gradient energy tensor (GET) filter [8]. The final $\mathrm{LPT}(x, y)$ image is obtained as follows [8]:

$$\mathrm{LPT}(x, y) = \sqrt{T_{\text{even}}^2 + T_{\text{odd}}^2} \times \cos(\varphi). \tag{1}$$

Here, φ represents instantaneous phase obtained from the symmetric and asymmetric feature responses, respectively [8]. $\mathrm{LPT}(x, y)$ provides a general enhancement independent of the specific feature type present in the acquired $\mathrm{US}(x, y)$ scans. This presents a more robust enhancement of complex shaped bone surfaces, such as the spine [8]. However, soft tissue interfaces close to the bone surface which have similar

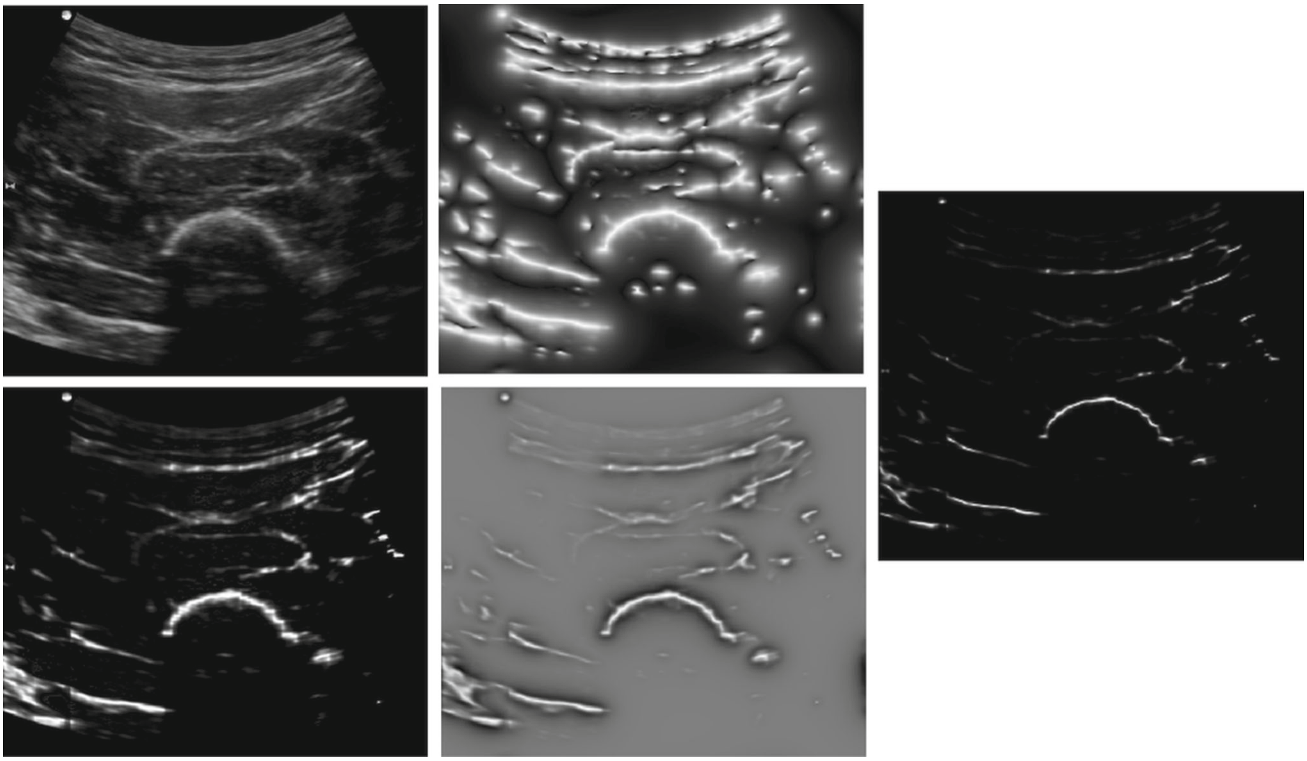

Fig. 1 First column top: US(x, y) image of in vivo femur bone. First column bottom: LPT(x, y) image. Second column top: LwPA(x, y) image. Second column bottom: LPE(x, y) image. Third column: $LP(x, y)$ image

intensity values are also enhanced during this process (Fig. 1). To suppress the enhancement of these soft tissue interfaces and obtain a more compact bone representation, monogenic image filtering was applied to LPT(x, y) image. This results in the extraction of two more local phase image features: local phase energy, LPE(x, y), and local weighted mean phase angle, LwPA(x, y). These two image phase features are obtained by combining the band-pass-filtered LPT(x, y) image, denoted as LPT$_B(x, y)$, with Riesz filtered components (represented by h_1 and h_2) resulting in the extraction of monogenic signal image, US$_M(x, y)$, as follows [4]:

$$\begin{aligned} US_M(x, y) &= [US_{M1}, US_{M2}, US_{M3}] \\ &= [\mathrm{LPT}_B(x, y), \mathrm{LPT}_B(x, y) * h_1, \mathrm{LPT}_B(x, y) * h_2]. \end{aligned} \quad (2)$$

In (2), $*$ represent the convolution operation. By accumulating the local energy of the image along multiple filter responses, the LPE(x, y) image encodes the underlying shape of the bone boundary. Averaging the phase sum of the response vectors over many scales generates the LPE(x, y) image as follows:

$$\mathrm{LPE}(x, y) = \sum_{sc} |US_{M1}| - \sqrt{US_{M2}^2 + US_{M3}^2}. \quad (3)$$

In (3), sc corresponds to the number of filter scales. The LwPA(x, y) image can be found as follows:

$$\mathrm{LwPA}(x, y) = \arctan\left(\frac{\sum_{sc} US_{M1}}{\sqrt{\sum_{sc} US_{M1}^2 + \sum_{sc} US_{M2}^2}}\right). \quad (4)$$

In (4), sc, represent the number of filter scales. The LwPA(x, y) image preserves all the structural details of the LPT(x, y) image, i.e., soft tissue interfaces and bone surfaces. Investigating the extracted local phase images (LPT(x, y), LPE(x, y), LwPA(x, y)) in Fig. 1 we can see that the bone surfaces have accurate localization in all the extracted phase images. However, the soft tissue interfaces do not have similar localization accuracy (they appear in different local regions in the image). Using this investigation the final local phase bone image, $LP(x, y)$, is obtained by multiplying the three phase feature images as: $LP(x, y) = \mathrm{LPT}(x, y) \times \mathrm{LPE}(x, y) \times \mathrm{LwPA}(x, y)$. Figure 1 shows all the extracted three local phase image features and the final $LP(x, y)$ image. Investigating Fig. 1 we can see that the final $LP(x, y)$ has compact representation of the bone surface with reduces soft tissue artifacts. The extracted local phase image, $LP(x, y)$, and the B-mode US image, US(x, y), are used during the proposed CNN-based bone segmentation methods which is explained in the next section.

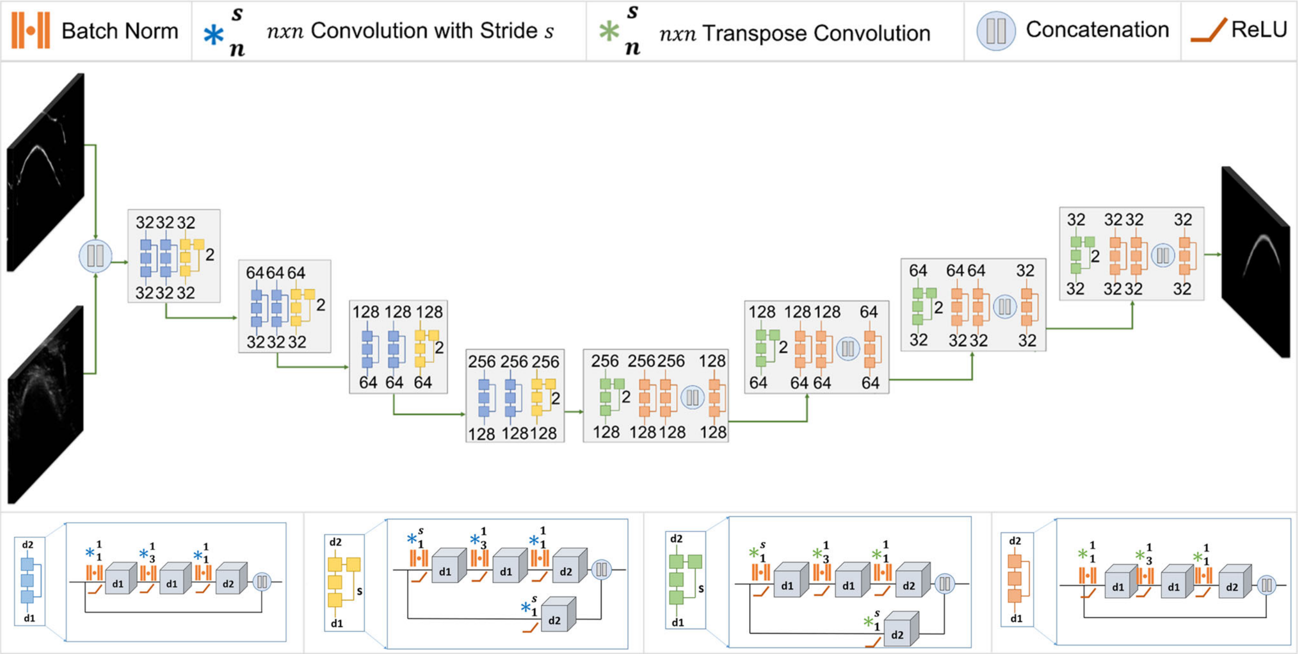

Fig. 2 An overview of the early-fusion CNN architecture. Input B-mode US image, US(x, y), is concatenated with the local phase filtered image, $LP(x, y)$, at the pixel level, and the result is processed through the network

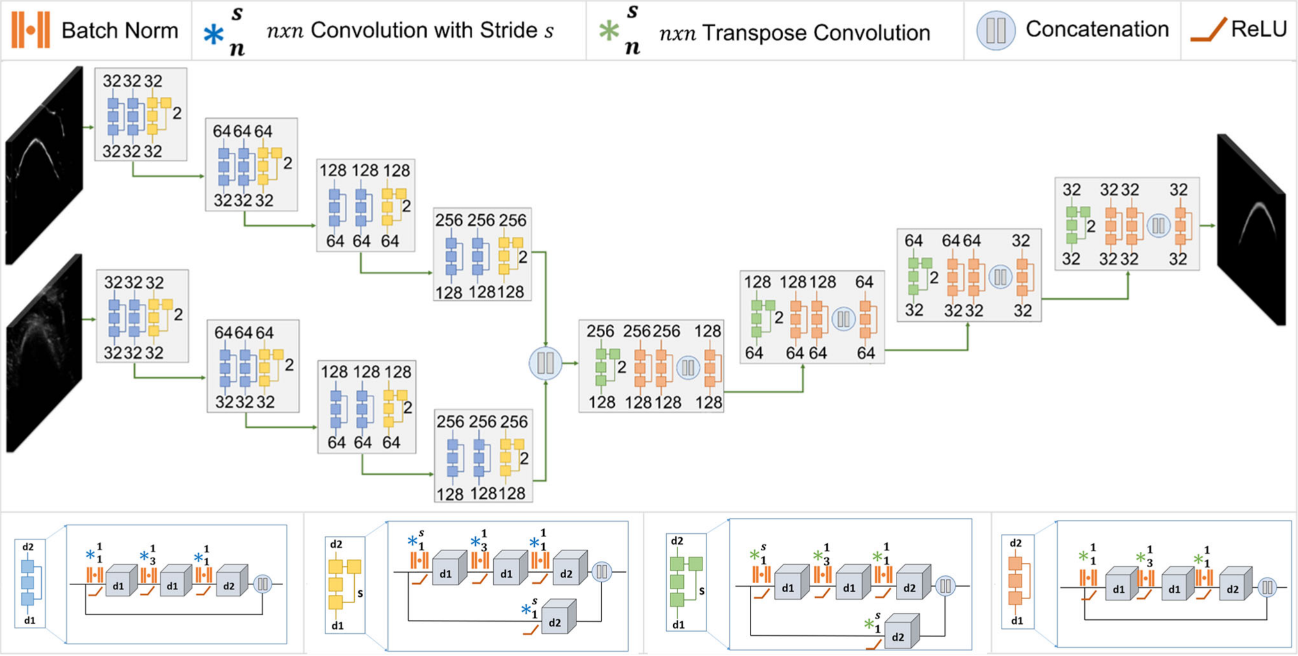

Fig. 3 An overview of the mid-fusion CNN architecture. Input B-mode US image, US(x, y), is processed through the primary encoder (bottom), while the local phase filtered image, $LP(x, y)$, is processed through the secondary encoder (top). Feature maps from both encoders are fused in a mid-fusion stage, and processed through the decoder

Network architecture

Our proposed CNN architecture is based on the common contractive-expansive design, as depicted in Figs. 2, 3 and 4. The encoder maps the input image into a low-dimensional latent space, and the decoder maps the latent representation into the original space. We first resize the input B-mode US image US(x, y) and its complementary local phase fil-

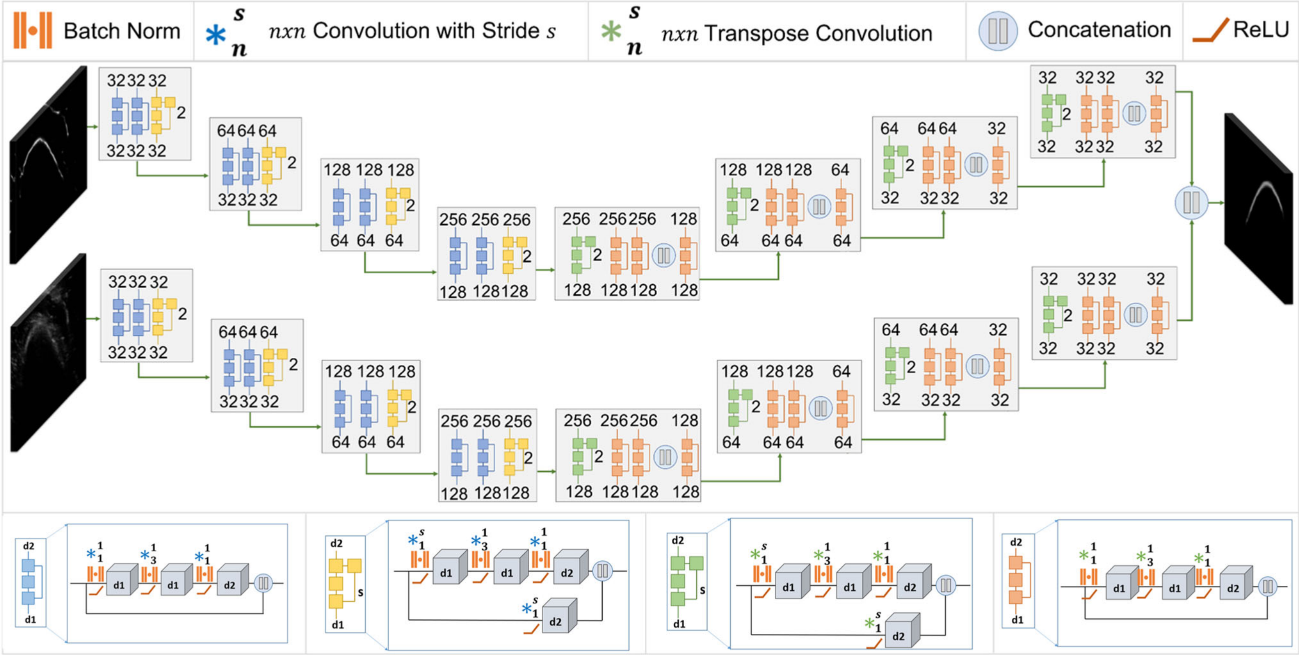
Batch Norm
$\overset{s}{\underset{n}{*}}$ nxn Convolution with Stride s
$\overset{s}{\underset{n}{*}}$ nxn Transpose Convolution
Concatenation
ReLU
32 32 32
64 64 64
128 128 128
256 256 256
2
d1
d2
s

Fig. 4 An overview of the Late-fusion CNN architecture. Input B-mode US image, US(x, y), is processed through the primary network (bottom), while the local phase filtered image, $LP(x, y)$, is processed through the secondary network (top). Feature maps from both networks are fused in a late-fusion stage

tered image $LP(x, y)$ to a standardized 256 × 256 size. After network operations, the images were resized to their original size before quantitative and qualitative validation. In our proposed design, each input image would connect to an independent primary network and a secondary network. In each network, the input image is processed through convolutional blocks, with each block consisting of several convolutional layers. Utilized in our design are four distinct blocks (colored in blue, yellow, green, and orange) that are depicted and labeled in the legend of each network design at the bottom of (Figs. 2, 3, 4). The networks in our design incorporated skip connection and projection blocks similar to [11]. In each of the four convolutional blocks, d1 and d2 indicate the depth of each convolutional layer, while s indicates stride. Within the network design, the specific depths of each convolutional block are specified. A skip connection block consists of 1 × 1 convolutions before, and after a 3 × 3 convolution, reducing and restoring channel dimensions, respectively. Each convolution is followed by batch normalization and rectified linear unit (ReLU) activation. The output of the skip connection block is obtained by concatenating its input with the aforementioned convolutions. Our projection blocks consist of a similar structure to the skip connections, with the difference being the output is the result of concatenating the aforementioned convolutions with the projected input through a 1 × 1 convolution. We employ projection blocks as a means of max-pooling when a stride of 2 convolution is used. On the other hand, transposed convolution blocks were implemented in the decoder path of each network. The design of the transposed convolution blocks are similar to the aforementioned skip and projection blocks with all convolution operations replaced by transposed convolutions. We also use a stride of 2 transposed convolutions to upsample the feature maps. In the primary network, the input image is a B-mode US image US(x, y), while in the secondary network, the input is a local phase filtered image, $LP(x, y)$, that proceeds through the aforementioned blocks. Feature maps extracted from both networks are fused (Figs. 2, 3, 4) in a fusion layer at various stages depending on the model. Specifically, we investigate early-, mid- and late-fusion network models. Our early-fusion model, depicted in Fig. 2, fuses the input B-mode US image, US(x, y), and the local phase filtered image, $LP(x, y)$, at the pixel level. The fused image is then processed through a single network. In Fig. 3, a feature level fusion model was implemented in which mid-level features from both primary and secondary networks are fused together. Finally, in Fig. 4, a classifier level model was implemented in which high-level features from each network are concatenated. A 3 × 3 convolution with sigmoid activation is performed on the output of the fused layer to generate the final segmented probability distribution. Throughout our network designs, concatenation fusion is used as the fusion operation [3]. Concatenation fusion does not define any correspondence as it stacks feature maps at the same spatial locations across the feature channels. However, subsequent layers define the correspondence by learning suitable filters that weight the layers.

Training and testing

The performance of the proposed designs were compared against each other and the networks proposed in [15], termed U-net, and [9]. The depths of both networks in [9,15] were increased to a scale close to our proposed designs. In order to further validate the effectiveness of our design, we trained the U-net network proposed in [15] using: (1) B-mode US image features only, (2) local phase image features only, and (3) both B-mode US and local phase image features. We have also designed a mid, and late level fusion U-net in order to further compare our proposed designs. Our proposed designs and the state-of-the-art networks in [9,15] were trained using a training set of 300 B-Mode US images (out of the 415 images obtained from Sonix-Touch US machine) and their corresponding local phase filtered images. The remaining 115 B-mode US images were reserved for testing the performance of the networks. During the random split of the SonixTouch dataset, same patient scans were not used for both training and testing. We repeated this process five times, with each training and testing data randomized from our datasets. In addition, we reserved another set of data for testing and validation only. This dataset consisted of 131 US scans from two different volunteers obtained using the Clarius C3 probe. The two volunteers were not part of the scanning process performed using the SonixTouch machine. Our proposed network designs were trained to minimize the cross-entropy loss. In all of our networks, we have used Adam Optimizer with batch size of 4 and a learning rate of 0.0002 for 35,000 iterations. Cross-entropy loss was used for the segmentation task. Based on [2,4,10,14], five error metrics were calculated in our testing set: namely, F-score, Rand error, Hamming Loss, as well as the IoU and average bone surface localization error. The evaluation metrics are computed on the estimated probability maps, with grayscale color maps, and compared to the manual segmentations.To measure how similar any two segmentation regions in an image, the Rand error, which takes into account shifts in boundary locations, was calculated as $R_e = 1 - R_i$, where R_i is the Rand index. The bone surface, in each scanline, was localized during a bottom-up ray casting method and by selecting the pixel values which are above the 60% of the maximum intensity value of the bone segmentation probability map. This value was optimized using scans from one subject and kept constant throughout the validation.The bone localization error is calculated as the average Euclidean distance (AED) error between the automatically segmented bone surfaces and the manual expert segmentation. The evaluation metrics were computed on the estimated probability maps and compared to the gold standard manual segmentations.

Results

Experiments were carried out using the Keras framework and Tensorflow as backend with an Intel Xeon CPU at 3.00 GHz and an Nvidia Titan-X GPU with 8GB of memory. On average, our networks converge in about 6 h during the training process. The total number of parameters in our late-fusion design was 9,440,531 with 9,409,811 trainable parameters and 30,720 non-trainable parameters. A single-stream Unet (with our implementation) had 5,470,627 trainable parameters. On average, our networks converge in about 6 h during the training process. Testing on average took 52 ms. However, this time does not include the local phase image extraction part which took on average 1 s.

Quantitative results

The aforementioned error metrics were calculated for each of the networks, and the results are tabulated in Table 1. As can be seen from Table 1, the average numerical error calculations show that that the late-fusion design had the lowest errors, and the highest average IoU and F-scores. A paired t test, for IoU F-score and AED results at a %5 significance level, between our designed networks and the networks proposed in [9,15] achieved p values less than 0.05 indicating that the improvements of our method are statistically significant. We have also performed same paired t test comparing our late-fusion to early- and mid-fusion networks and achieved p values less than 0.05. The AED results increased for all the networks analyzed when using the data obtained from the Clarius imaging platform. This is an expected results since this data was obtained from an imaging platform which was not part of the training process. However, incorporating local phase image features increases the success of the network proposed in [15].

The overall AED error for our late-fusion design is 0.1482 mm (standard deviation (SD) 0.028 mm). U-net network [15] using B-mode, local phase, and combine (B-mode and local phase features) achieved overall AED errors of 2.296 mm (SD 0.038 mm), 1.0319 mm (SD 0.059 mm), 0.7060 mm (SD 0.05 mm), respectively. The network proposed in [9] achieved overall AED error of 0.8612 mm (SD 0.0834 mm). Again a paired t test, at a %5 significance level, between our proposed late-fusion network and other network achieved p values less than 0.05 for overall AED errors.

Qualitative results

Qualitative results of our early, mid, and late-fusion network designs as well as the networks in [15] and [9] are shown in Fig. 5, where the red pixels indicate high prediction scores, while blue pixels indicate low prediction scores for the segmentation. The prediction outcome when only B-mode US

Table 1 Error metrics

Method	IoU%	*F*-score	Rand	Hamming	AED (mm)
Dataset I—Sonix-Touch US machine					
Ronneberger [15] B-mode US (BM) only	0.864986	0.912431	0.705538	0.135013	2.4344
Ronneberger [15] Local phase (LP) only	0.870279	0.915084	0.654647	0.129720	1.0443
Ronneberger [15] BM & LP early fusion	0.914163	0.944772	0.626641	0.085836	0.6375
Ronneberger [15] BM & LP mid fusion	0.928410	0.952721	0.655679	0.071590	0.4927
Ronneberger [15] BM & LP late fusion	0.948090	0.964523	0.626641	0.051913	0.2864
Hazirbas [9]	0.894572	0.931847	0.657970	0.105427	0.8582
Ours early fusion	0.972125	0.978072	0.448373	0.027874	0.1087
Ours mid fusion	0.957193	0.969739	0.484314	0.042806	0.1183
Ours late fusion	**0.972865**	**0.978388**	**0.439368**	**0.027134**	**0.1071**
Dataset II—Clarius C3 US probe					
Ronneberger [15] B-mode (BM) US only	0.820128	0.878605	0.647348	0.179871	2.1576
Ronneberger [15] Local phase (LP) only	0.869984	0.950463	0.746484	0.179871	1.0195
Ronneberger [15] BM & LP early fusion	0.904848	0.944324	0.760781	0.095151	0.7463
Ronneberger [15] BM & LP mid fusion	0.932476	0.956250	0.656760	0.067523	0.4682
Ronneberger [15] BM & LP late fusion	0.948085	0.964500	0.624421	0.051914	0.3755
Hazirbas [9]	0.842741	0.901746	0.670910	0.157258	0.8642
Ours early fusion	0.968001	0.976297	0.487246	0.031998	0.2186
Ours mid fusion	0.958058	0.969953	0.485360	0.041941	0.2579
Ours late fusion	**0.970965**	**0.977650**	**0.453752**	**0.029034**	**0.1893**

Bold values indicate statistically significant improvement

images were used in training, as the case in Fig. 5b for [15], had the lowest probability distribution among all others. The inadequate segmentation performance may be attributed to the nature of the US images used in the testing process. For low-quality US scans, where the bone surface has a low intensity profile and high-intensity soft tissue interfaces appearing above the bone surface, the performance of the network proposed in [15] declines. The importance of collecting high-quality US data and its affect on the segmentation outcome was also discussed previously in [1] who proposed a similar network architecture for segmenting vertebra bone surfaces from US data. Fig. 5d shows an improved prediction outcome when both B-mode US and local phase filtered images are used to train the U-net architecture proposed in [15]. The network in [9] had a lower probability distribution than our proposed fusion networks. This is because in the network proposed in [9], the fusion performed at the feature level is considered a slow fusion in which multiple feature maps are fused throughout the encoder. As shown in Fig. 5f, early fusion outperforms the network in [9] since the fusion happens at the pixel level in which the fused image would possess enhanced bone surfaces, while the soft tissue interfaces remain unaltered. Investigating Fig. 5 (first and third rows), we can also see that all the networks perform better when the testing data is from the same imaging platform where the training data is obtained. However, when using test data obtained from an imaging platform which the networks have not seen during training the performance decreases. However, we can still infer that our network designs, compared to U-net [15] and [9], perform better resulting with segmentation outcomes with high probability.

Bone localization results, against expert manual localization, are presented in Fig. 6. The specific B-mode US data presented in this figure (Fig. 6a) show low-quality bone scans. Investigating the localization results we can infer that U-net [15] trained only using B-mode data achieves the worst performance: large gap from the expert localization, missing bone boundaries, false positive bone localizations. Although the performance increases when the same network is trained together with B-mode and local phase features (Fig. 6d) false and true positive localization is still visible. Qualitatively, our late-fusion design achieves the best performance for this dataset.

Discussion and conclusions

In this study, three CNN architectures, for the task of bone segmentation from US data, were proposed. Our networks incorporate local phase images in conjunction with B-mode US data. We have investigated how to combine information from local phase images and B-mode US data by analyzing

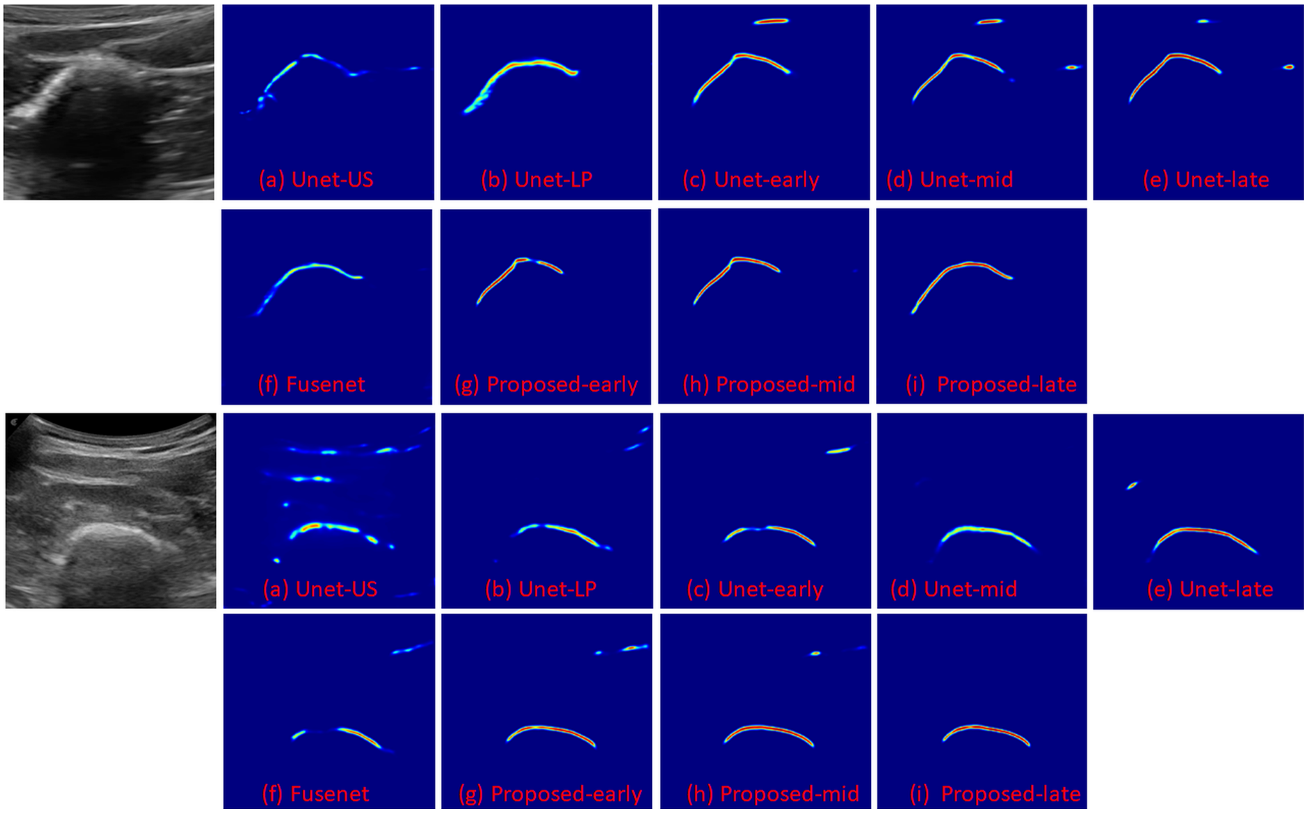

Fig. 5 First column in vivo US B-mode images of distal radius (top), and femur (bottom). Image are obtained from the Clarius platform. Network segmentation results obtained using: Ronneberger et al. [15] trained with **a** B-mode US images only (U-net), **b** local phase filtered images only (U-LP), **c** B-mode US and local phase filtered images using early-fusion (Unet-early), **d** B-mode US and local phase filtered images using mid-fusion (Unet-mid), **e** B-mode US and local phase filtered images using late-fusion (Unet-late). **f** Hazirbas et al. [9] trained with both B-mode US and Local phase filtered images (Fusenet). Our proposed designs **g** early-fusion, **h** mid-fusion, and **i** late-fusion

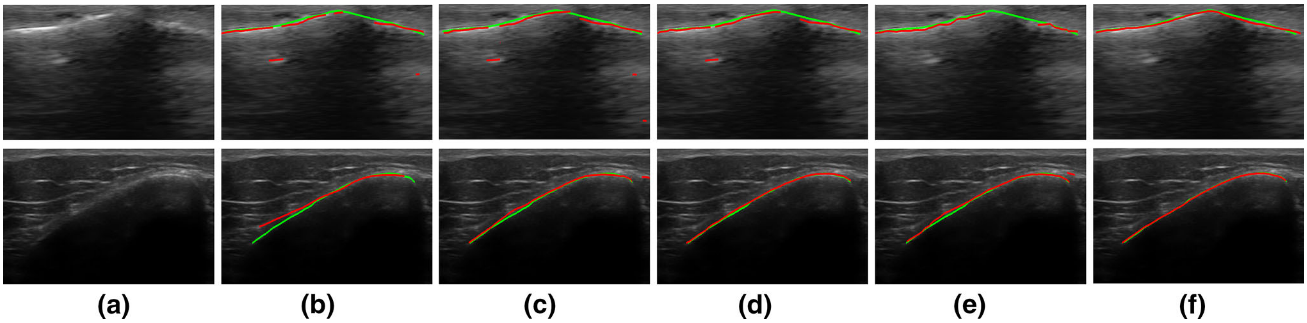

Fig. 6 Bone localization obtained from the proposed method (red) to manual expert localization (green). **a** In vivo B-mode US image of distal radius (top) and femur (bottom). **b** Ronneberger et al. [15] trained with B-mode US images only, **c** Ronneberger et al. [15] trained with local phase filtered images only, **d** Ronneberger et al. [15] trained with B-mode US and local phase filtered images, **e** Hazirbas et al. [9], and **f** our late-fusion design

different fusion strategies. Our results demonstrate that for the task of bone segmentation fusing B-mode US and local phase features at a later stage outperforms early and mid fusion, specifically for the dataset obtained from Clarius C3 US probe. Since local phase image features enhance the bone surface response in the US data, the B-mode US data and local phase image features are less correlated in the low-level features. The proposed late level fusion network models the correlations and interactions between high-level features of each modality, outperforming the other fusion networks. A similar investigation can also be observed with the U-net network late fusion design [15] (Table 1). We also show that incorporating local phase bone image features, using three different stages of fusion, improves the performance of state-of-the-art U-net network [15]. Conducted quantitative studies show significant improvement of our network with late fusion over state-of-the-art CNN methods [15].

In our network architecture, we use convolutional/ projection blocks. Our projection blocks allow semantic information to be more efficiently passed forward in the network while progressively increasing feature map sizes, compared to simple convolutions which is used in the U-net design [15]. The projection blocks allow us to have more comprehensive feature maps. This is one of the reasons why our fusion networks outperform fusion networks of U-net design [15].

One of the drawbacks of the proposed work is the computational time required for the extraction of local phase image features. This takes on average 1 s (MATLAB implementation) which needs to be improved for real-time CAOS procedures where US is used as an intra-operative imaging modality. Furthermore, during this work the expert manual segmentation was performed by a single expert user. The effect of intra- and inter-user expert bone segmentation on the segmentation results is also crucial. Our future work will involve (1) extensive clinical validation of the proposed method, (2) improving the computational cost of local phase feature extraction, (3) inter- and intra-user variability analysis for expert bone segmentation, and (4) extension of our network architecture to process volumetric US data [7].

Funding This work was supported in part by 2017 North American Spine Society Young Investigator Award.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

1. Baka N, Leenstra S, van Walsum T (2017) Ultrasound aided vertebral level localization for lumbar surgery. IEEE Trans Med Imaging 36(10):2138–2147
2. Cernazanu-Glavan C, Holban S (2013) Segmentation of bone structure in x-ray images using convolutional neural network. Adv Electr Comput Eng 13(1):87–94
3. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1933–1941
4. Hacihaliloglu I (2017) Enhancement of bone shadow region using local phase-based ultrasound transmission maps. Int J Comput Assist Radiol Surg 12(6):951–960
5. Hacihaliloglu I (2017) Localization of bone surfaces from ultrasound data using local phase information and signal transmission maps. In: International workshop and challenge on computational methods and clinical applications in musculoskeletal imaging. Springer, pp 1–11
6. Hacihaliloglu I (2017) Ultrasound imaging and segmentation of bone surfaces: a review. Technology 5(2):74–80
7. Hacihaliloglu I, Guy P, Hodgson AJ, Abugharbieh R (2014) Volume-specific parameter optimization of 3d local phase features for improved extraction of bone surfaces in ultrasound. Int J Med Robot Comput Assist Surg 10(4):461–473
8. Hacihaliloglu I, Rasoulian A, Rohling RN, Abolmaesumi P (2014) Local phase tensor features for 3-d ultrasound to statistical shape + pose spine model registration. IEEE Trans Med Imaging 33(11):2167–2179
9. Hazirbas C, Ma L, Domokos C, Cremers D (2016) Fusenet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Asian conference on computer vision. Springer, pp 213–228
10. Jain V, Bollmann B, Richardson M, Berger DR, Helmstaedter MN, Briggman KL, Denk W, Bowden JB, Mendenhall JM, Abraham WC et al (2010) Boundary learning by optimization with topological constraints. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2488–2495
11. Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N (2016) Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV), pp 239–248
12. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
13. Organization WH (2003) The burden of musculoskeletal conditions at the start of the new millennium: report of a who scientific group. WHO Technical Report Series 919
14. Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(336):846–850
15. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241
16. Salehi M, Prevost R, Moctezuma JL, Navab N, Wein W (2017) Precise ultrasound bone registration with learning-based segmentation and speed of sound calibration. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 682–690
17. United States Bone and Joint Initiative (2014) The burden of musculoskeletal diseases in the United States (BMUS), 3rd edn. Rosemont, IL. http://www.boneandjointburden.org. Accessed on 13March 2018
18. Valada A, Vertens J, Dhall A, Burgard W (2017) Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In: 2017 IEEE International conference on robotics and automation (ICRA). IEEE, pp 4644–4651
19. Villa M, Dardenne G, Nasan M, Letissier H, Hamitouche C, Stindel E (2018) FCN-based approach for the automatic segmentation of bone surfaces in ultrasound images. Int J Comput Assist Radiol Surg 13(11):1707–1716
20. Wang P, Patel VM, Hacihaliloglu I (2018) Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided CNN. In: Medical image computing and computer assisted intervention

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.