Ultrasound in Med. & Biol., Vol. 43, No. 10, pp. 2426–2437, 2017
© 2017 World Federation for Ultrasound in Medicine & Biology
Printed in the USA. All rights reserved
0301-5629/$ - see front matter

http://dx.doi.org/10.1016/j.ultrasmedbio.2017.04.022

● *Original Contribution*

# RANDOM FOREST-BASED BONE SEGMENTATION IN ULTRASOUND

Nora Baka,[*] Sieger Leenstra,[†] and Theo van Walsum[*]

[*] Biomedical Imaging Group Rotterdam, Departments of Radiology & Nuclear Medicine and Medical Informatics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands; and [†] Department of Neurosurgery, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands

(*Received* 19 *August* 2016; *revised* 14 *April* 2017; *in final form* 24 *April* 2017)

**Abstract—Ultrasound (US) imaging is a safe alternative to radiography for guidance during minimally invasive orthopedic procedures. However, ultrasound is challenging to interpret because of the relatively low signal-to-noise ratio and its inherent speckle pattern that decreases image quality. Here we describe a method for automatic bone segmentation in 2-D ultrasound images using a patch-based random forest classifier and several ultrasound specific features, such as shadowing. We illustrate that existing shadow features are not robust to changes in US acquisition parameters, and propose a novel robust shadow feature. We evaluate the method on several US data sets and report that it favorably compares with existing techniques. We achieve a recall of 0.86 at a precision of 0.82 on a test set of 143 spinal US images. (E-mail: t.vanwalsum@erasmusmc.nl) © 2017 World Federation for Ultrasound in Medicine & Biology.**

***Key Words:*** **Ultrasound, Machine learning, Spine, Vertebra, Intra-operative, Ultrasound guidance, Orthopedic procedure.**

## INTRODUCTION

Ultrasound (US) imaging is a safe alternative for guidance during minimally invasive orthopedic procedures. Its main advantage compared with X-ray guidance is the lack of ionizing radiation and its cost-effectiveness. However, ultrasound imaging has its own challenges, such as the relatively low signal-to-noise ratio, its inherent speckle pattern, shadowing and several types of artifacts. US guidance is therefore mainly performed by registering the acquired US images to the pre-operative computed tomography (CT) image on which the intervention was planned. Such registration usually requires the bone surface from both modalities, US and CT (Nagpal et al. 2015; Penney et al. 2006). Automatic bone detection algorithms are crucial for such navigation. In this article, we propose and evaluate a method for such automatic bone tissue interface detection from US. We propose learning the appearance of bone interfaces in the images based on annotated training examples and machine learning methods.

There have been several published approaches to solving bone classification from ultrasound, most of them using heuristic functions to calculate bone and non-bone interfaces. The most obvious heuristic is that the bone surface appears bright in the images. An additional intensity correction using the expected depth of the bony structure can suppress other bright interfaces, as proposed in Kowal et al. (2007), to effectively highlight bones. The down side of this method is that the expected depth of the bone must be known; otherwise, noise or soft tissue interfaces will be enhanced, and could be mistaken for bone. Hacihaliloglu et al. (2009) described a different approach, suited for bones at all depths. They proposed using phase symmetry in the frequency domain to find and enhance edges regardless of their brightness. This was determined to be very accurate in finding the bone outline. However, the method enhances all lines in the image, including fat–muscle or other soft-tissue interfaces. A property most widely used to distinguish bone for soft tissue interfaces is shadowing. Because of the large difference in acoustic properties of bone and soft tissue, almost all the sound energy is reflected from the bone surface. The lack of sound traversal through the bone creates the shadow, a region of dark intensities below the bone surface. Karamalis et al. (2012) proposed an algorithm to quantify the chance of sound

Address correspondence to: Theo van Walsum, PO Box 2040, 3000 CA Rotterdam, The Netherlands. E-mail: t.vanwalsum@erasmusmc.nl

reaching every image pixel, which can be used as an indicator for shadowing. Quader et al. (2014) combined the two features, phase symmetry and the shadow feature of Karamalis et al., and improved bone detection accuracy. Indeed, multiple properties can be combined to reliably characterize bone in US. Jain and Taylor (2004) proposed a Bayesian framework combining intensity, gradient, shadow, intensity profile along scanline and multiple reflections for bone segmentation; however, how the required conditional probabilities can be obtained is not straightforward. Foroughi et al. (2007) proposed a heuristic combination of intensity, shadow and the Laplacian filtered image to derive a probability map for bones. In a second step, a maximum of one pixel per image column was selected as bone, producing the final segmentation with dynamic programming. This post-processing method became popular in the field, as it could correct for small errors in bone probability images. A similar method was used by Jia et al. (2016) and Cao et al. (2016) with different heuristically calculated feature images. Jia et al. (2016) proposed calculating the bone probability images by multiplying in total seven feature images, including integrated back scattering and local energy. Although these methods exhibited good accuracy on their respective published test data, one might wonder if they are optimal given the manually created cost function.

Learning the combination and importance of features from training data seems a more structured way of characterizing bone interfaces in US. Penney et al. (2006) proposed learning the distribution function of bones versus background using two features, bone intensity and an artifact distance. Any pixel with an intensity $<40$ was defined as artifact. More recently, Berton et al. (2016) combined the bone probability feature of Foroughi et al. (2007), Hacihaliloglu et al. (2009), Hellier et al. (2010) with local binary pattern (LBP) and Gabor filtering for shadow, bone and soft tissue differentiation. In their work, they used a linear classifier on the already heuristically combined features.

In this article, we propose learning the bone probability map from simple features, using a patch-based classification approach. The contributions of this work are as follows:

- We present a bone segmentation scheme using a patch-based classification approach and perform an extensive evaluation of the method.
- We propose a novel shadow feature and evaluate it in comparison with different shadows and other features.
- We compare the presented method to the standard heuristic methods from the literature.
- We include multiple ultrasound data sets to assess classification robustness.

This work extends our previous work in which we compared linear and non-linear classifiers for bone segmentation from ultrasound images (Baka et al. 2016). This study differed in that it used a simplified classifier, proposed and evaluated a new shadow feature and provided more extensive evaluations between methods, between data sets and for parameters within the method.

## METHODS

Our aim is to segment the bone interfaces from US images. For this, we propose learning a classifier from a training set of annotated US images for bone segmentation. Once the classifier is learned, the bone probability map of an unseen image can be computed. The classification is done as follows. First, the image undergoes a preprocessing step. Subsequently, the feature images are computed. Each pixel of the image and the patch around it are then fed into a classifier, which gives the probability of that pixel being part of a bone–soft tissue interface. This step thus results in the bone probability map. If a single interface line segmentation is desired rather than a probability map, a dynamic programming post-processing step for segmentation can be added, as in Foroughi et al. (2007). Below we describe each part of the method in detail.

### *Random forest classifier*

Random forest classifiers are non-linear classifiers consisting of several decision trees, first proposed by Ho (1995). The output of the forest is the average prediction of its trees. To ensure that the trees are sufficiently dissimilar, every tree is trained on a subset of features and on a subset of data. Each tree consists of a series of nodes, that can either branch into two child nodes with a splitting rule or be a leaf node. When learning the tree, at each yet unsplit node, a splitting rule is computed, which best separates the positive and negative samples arriving at that node. At test time, the new sample is passed through each tree according to the splitting rules and ends up in a leaf node. The output probability of the sample from a tree is then equal to the percentage of positive training samples that are in that leaf node.

Random forest classifiers work well with a large number of features, and with selection of the best feature during training for splitting each node, they have an inherent feature selection property. This makes them good candidates for patch-based learning, as we propose for US segmentation. In this work, all the pixels surrounding a US pixel in an $n \times n$ window are taken as feature candidates. Additionally, differential features calculated by downsampling the window to a size of $5 \times 5$ and subtracting any two random samples from this smaller window are also taken as feature candidates. This was reported to increase accuracy by Lim et al. (2013) and Dollár and Zitnick (2013). Figure 1 is a schematic overview of a patchwise random forest classifier. For simplicity, this figure does not show the differential features.

In the case of bone segmentation from US images, there are considerably fewer positive samples (bone pixels) than negative samples (non-bone pixels). To reduce this asymmetry, we do not use all available samples for training, but sample $N_\mathrm{p}$ positive and $N_\mathrm{n}$ negative samples from all the training data set. To ensure the subsampling includes the relatively rare non-bone interface pixels, we employ a sampling mask $M_\mathrm{negative}$ such that the mask contains all pixels that have an $I_\mathrm{LoG} > 0.3$, or are part of a gridwise subsampling of the image, and are not in the neighborhood of the ground truth contour:

$$M_\mathrm{negative} = (I_\mathrm{LoG}>0.3) \cup M_\mathrm{grid30} \cup M_\mathrm{gridMid} \setminus \mathrm{dilate}\left(M_\mathrm{positive}, d\right) \tag{1}$$

Here $\bigcup$ and $\setminus$ are set union and set minus operators, and $I_\mathrm{LoG}$ is the negative Laplacian of Gaussian filtered image, as described under Feature Images. The empirically defined threshold of 0.3 highlights the tissue interfaces. The second component is a grid of pixels 3 mm from each other, and the next component is a denser grid at the middle of the image. $M_\mathrm{positive}$ is the mask of the positive samples, which is then dilated by $d$ = 2 mm and excluded from the negative mask. The negative samples for training are randomly selected from the allowed points of the mask $M_\mathrm{negative}$.

### *Pre-processing*

Ultrasound images inherently contain speckle. To reduce the effect of speckle on the subsequent classification, we applied a Gaussian filter on the images, with a standard deviation of $\sigma$ = 0.3 mm. We empirically found that blurring at this scale smoothed the speckle pattern while keeping neighboring tissue interfaces separated. The size of the smoothing kernel was not critical though, and a slightly smaller or larger smoothing would work as well. The speckle size in the depth direction of our images was about one pixel (0.1 mm).

Second, oblique bone interfaces are less bright in the images, as in this case only part of the emitted sound wave is reflected back to the transducer. To overcome this issue, Cao et al. (2016) enhanced the intensity of image locations that had an oblique gradient orientation. Computing a reliable gradient orientation on noisy images is though not trivial, as it is very sensitive to scale selection. We therefore propose enhancing oblique regions with template matching. We took as template the Gabor filter with orientations of 45° and −45°, wavelength $\lambda$ = 2 mm and Gaussian width $\sigma_\mathrm{g}$ = 1 mm. These parameters were set such that the filter resembles a single oblique edge with some room for deviation from the above angles. The filtered images were then thresholded to retain only the high-score regions and added to the blurred image such that

$$I_\mathrm{pre} = I_\sigma + \alpha\, F\left(\mathrm{Gabor}\left(\lambda, \sigma_\mathrm{g}, 45^\circ\right) * I_\sigma\right) + \alpha\, F\left(\mathrm{Gabor}\left(\lambda, \sigma_\mathrm{g}, -45^\circ\right) * I_\sigma\right) \tag{2}$$

where $I_\mathrm{pre}$ is the final pre-processed image, and $F$ represents a thresholding with 2 and a subsequent Gaussian blurring with $\sigma$. The threshold value was selected

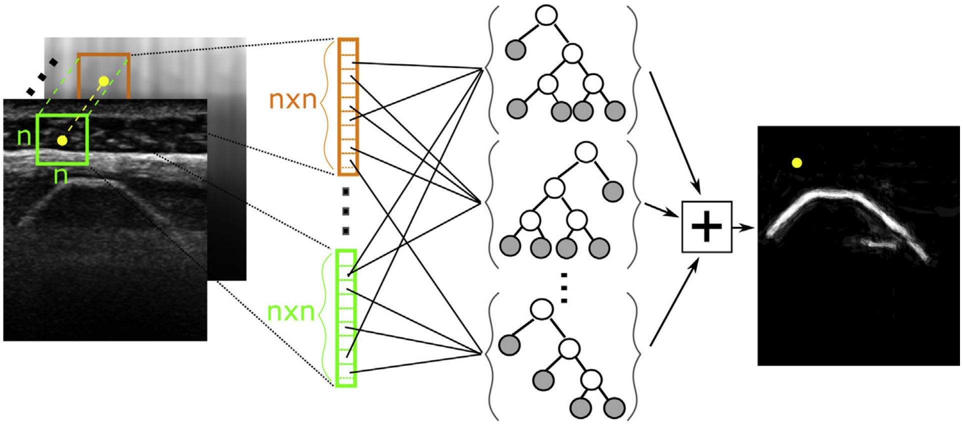

Fig. 1. Patchwise random forest classifier. To predict the output of a pixel (*yellow dot*), the $n \times n$ patch around the pixel is taken in each feature image. The pixel values in the patches are then concatenated to form a long feature vector. Each tree in the forest uses a random subset of these features to produce a probability estimate. The average combination of these probability estimates is the final probability of the center pixel. By doing this for all pixels in the image, the final bone probability image is calculated.

empirically, such that only the strongest responses remained, and the subsequent blurring was used to smooth out the edges. We used an $\alpha = 10$ weighting, such that the maximum value of the enhancement was about 50 for an image with a maximum intensity of 255. We also resampled the images to a pixel size of 0.1 mm.

### *Feature images*

We investigated several features, which we group in two categories, local and global features. Local features include the intensity and its derivatives, which can be computed locally in the image. Global features, on the other hand, require a larger part of the image or the entire image to be computed. Typically, shadow features belong to this category, but also location and relative location features. All investigated features are listed below and computed using the pre-processed images unless stated otherwise.

1. Intensity image $I_{\text{pre}}$
2. LoG image $I_{LoG} = -(\text{LoG} * I_{\text{pre}})$
3. Intensity shadow, which was calculated as the integral of the image intensities $I_{\text{pre}}$ under every pixel
4. Relative shadow, which was calculated as the integral of the relative image intensities under every pixel. The relative image intensity is calculated by subtracting the smoothed row average from every pixel, with negative pixels set at 0. In this work, we smoothed the row averages in the depth direction by a Gaussian with a 2-mm standard deviation. The method is not sensitive to this choice, and other tested values between 0 and 4 mm perform similarly on our data.
5. Karamalis shadow, as proposed by Karamalis et al. (2012). This shadow feature is inspired by acoustic wave propagation, giving the confidence that particles released along the top row of the image will arrive at each pixel of the image. As particles can slightly move sideways, calculation of this features requires knowledge of the entire image. We applied a small Gaussian blurring ($\sigma = 1$ pixel) prior to the shadow calculation, but did not do the oblique edge enhancement. All other parameters were set as proposed in Karamalis et al. (2012).
6. Artifact distance shadow. Proposed by Penney et al. (2006), the artifact distance shadow is calculated by defining all pixels with intensity lower than 40 as artifact and assigning them a feature value of 0. The value of the remaining pixels is calculated from their distance from the deepest non-artifact pixel per column. This feature was calculated from the original image, without pre-processing.
7. Border-to-border distance (BB). This feature discriminates structures that span the entire field of view (FOV), from left to right, from smaller structures. It is useful if we know that the imaged bone width is less than the FOV, which is the case with most bones. In this case the feature helps in discriminating fat–muscle interfaces from bones. We calculate the weighted distance function from the left border and from the right border of the image, where the weight is inversely proportional to the amount of structure in the image $I_{\text{weight}} = 1/(\max(0, I_{\text{LoG}}) + 0.01)$. The sum of the left and right distances results in low values for pixels that participate in a long structure, higher values for smaller structures and highest values where there is no structure at all.
8. Centrality. This feature quantifies the distance of structures from the vertical centerline of the field of view. We calculate this feature as the weighted distance function from the center of the image in both directions. The same weight image $I_{\text{weight}}$ was used as in the border-to-border distance. The feature is lowest for pixels that participate in a structure that is crossing the image centerline, and highest for pixels that are far from both the centerline and any structure in the image. This feature encodes the prior knowledge that structures in the middle of the image tend to be of greater importance than those on the border.
9. Depth. This feature is the distance (in mm-s) from the transducer. We use this feature as it has been suggested that the upper few millimeters below the transducer cannot be bone (Foroughi et al. 2007; Penney et al. 2006).

Figure 2 is an example of all the features used in this article, except the depth. For better visualization, features were scaled between 0 and 255.

### *Data*

We used two 2-D ultrasound data sets in this study. Both data sets contain images of the spinous process of vertebrae of the lower back, imaged in the sagittal plane. The data sets consist of both patient and volunteer data. An institutional review board waiver was granted for the US data collection. All data were processed anonymously, and patient approval was received from every patient.

Data set 1 was acquired at St. Elisabeth Hospital (Tilburg, Netherlands) with a Philips CX50 CompactXtreme system (Philips, Amsterdam, Netherlands) and a 2D L12-3 broadband linear array transducer. It consists of a training set of 106 vertebral US images from 15 subjects (data set 1A) and a test set of 56 images of 10 subjects (data set 1B). The acquisition focused on the lumbar vertebrae, but in some cases the thoracic T12 vertebra was also included. Each vertebra was imaged for about 3 s at a gain of 45 and contrast of 55. The default imaging depth was 5 cm, but it was adjusted if needed. Body mass

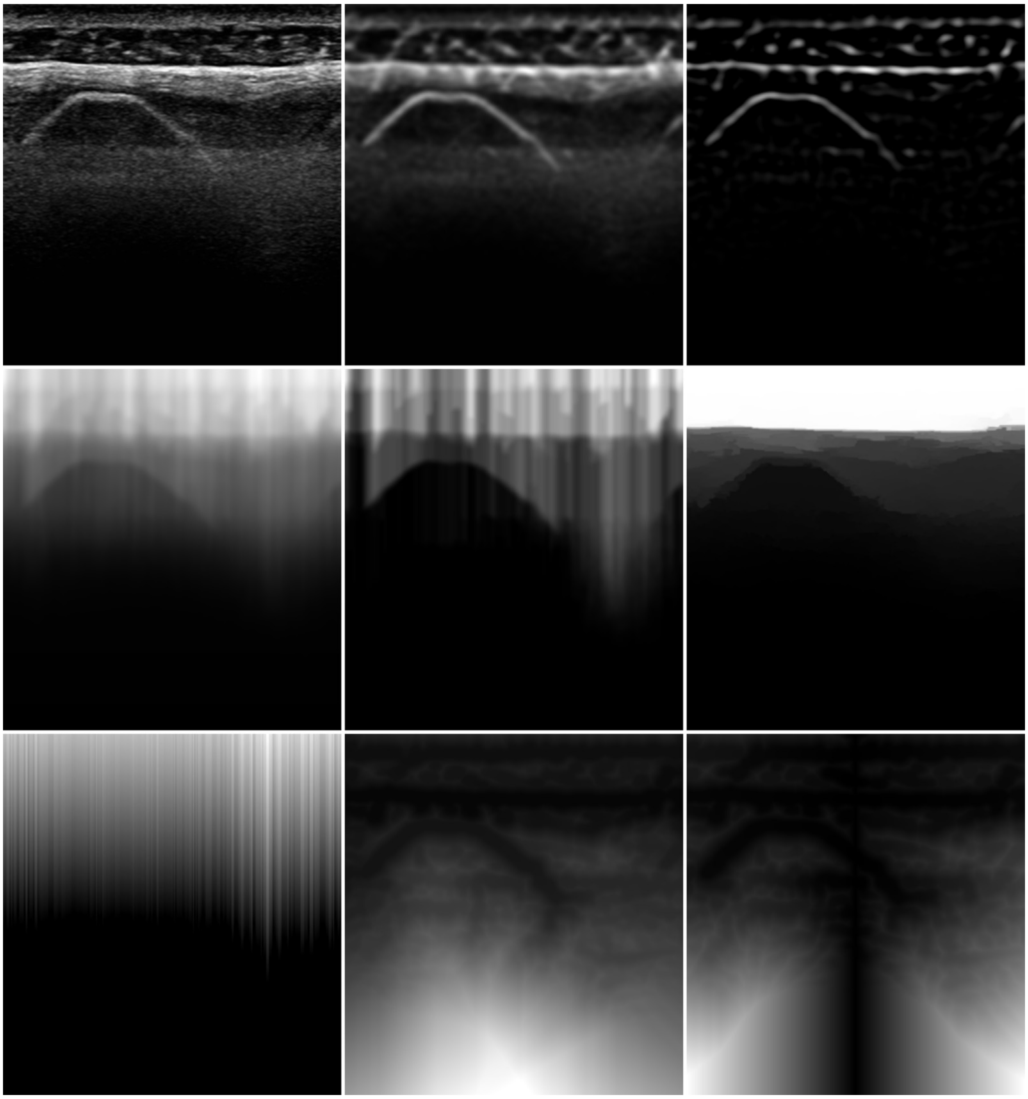

Fig. 2. Feature images. From left to right: (top row) original image, pre-processed image, Laplacian of Gaussian; (middle row) intensity shadow, relative shadow, Karamalis shadow; (bottom row) artifact distance shadow, border-to-border distance, centrality.

index (BMI) and age for the subjects in this data set were not recorded. The subjects were in a prone position (facing down) when imaged. For every vertebra, at least one image of the sequence was selected for manual annotation of the ground truth bone surfaces.

Data set 2 was acquired at Erasmus MC with a Philips iU22 machine and the 2D L12-5 linear transducer. It consists of a training set of 91 vertebra images of 11 subjects (data set 2A) and a test set of 87 images of 10 subjects (data set 2B). The division of the imaging data over a training set and test set was performed randomly per BMI category, to ensure sufficient training and test instances for all categories. The images were acquired with the general musculoskeletal protocol, using SonoCT and XRES adaptive image enhancement. The imaging depth was adjusted to between 3 and 5 cm, depending on the subject's anatomy, and a gain of 65 was used. Focus height was adjusted if needed, such that the imaged bone was in the focus region. Pixel size in most cases was around 0.1 mm. All subjects were in a sitting position when scanned. The acquisition focused on the lumbar vertebrae, but some neighboring thoracic and sacral vertebrae were also included. The subject population consisted of back pain patients aged 19 to 77 (median age: 47), with BMIs between 17.9 and 39.1 (median BMI: 26.81). For every vertebra, at least one image of the sequence was selected for manual annotation of the ground truth bone surfaces.

Manual annotation in both data sets was performed by placing control points on the bone surface

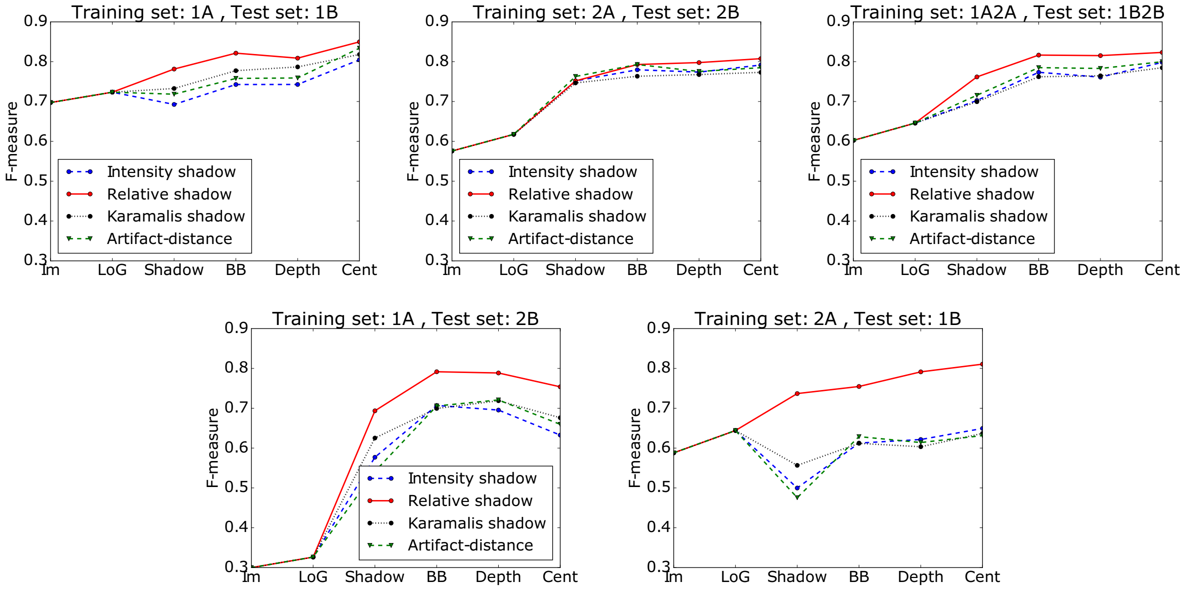

Fig. 3. Results of experiment 1. The plots indicate the added value of each feature, if the features are added in the order of appearance on the x-axis. Each classifier was given a maximum of one shadow feature. The different lines represent the accuracy with the different shadow features. Top row: Training and test sets belong to the same data set. Bottom row: Training and test sets are from two distinct data sets. BB = border-to-border distance, LoG = Laplacian of Gaussian.

and using spline interpolation in between. All visible bone surfaces were annotated, even if resulting in several disjoint lines on an image. The annotators were instructed to annotate as much of the bony structure as they were reliably able to determine. Annotators could scroll through the sequence to use information

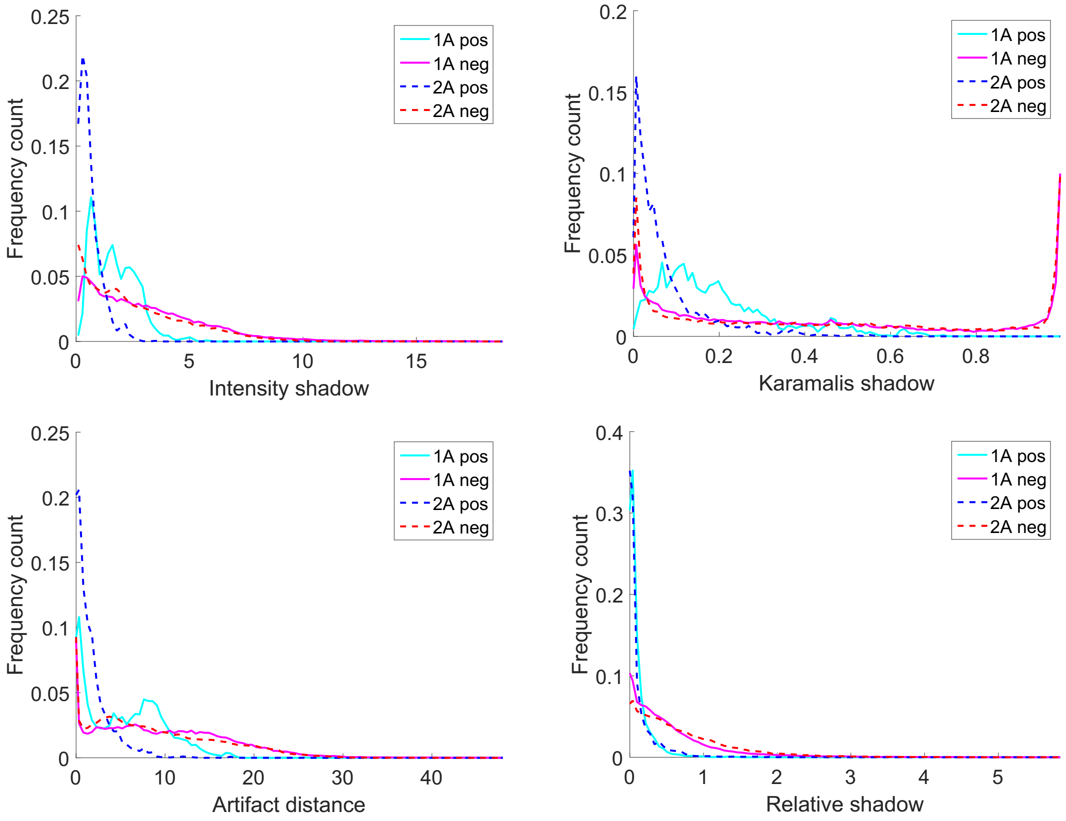

Fig. 4. Histograms for the shadow features for random positive and negative samples from data sets 1A and 2A.

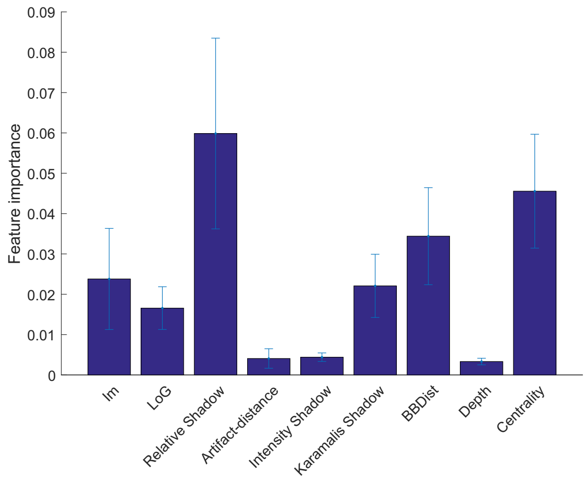

Fig. 5. Feature importance calculated from a forest with 12 trees, trained on data set 1A2A, with all features. The higher the value the more important is the feature. BBdist = border-to-border distance, LoG = Laplacian of Gaussian.

from neighboring frames when they doubted what they saw in the current frame.

### Experimental setup

We performed experiments to investigate the performance and robustness of the method with the different US data sets. Furthermore, we performed experiments to assess the value of the different features to segmentation accuracy. We introduce the evaluation measures in the next subsection, and present the experiments thereafter.

### Evaluation measures

We evaluated the bone probability map by first thresholding it to produce a binary segmentation. In all experiments, we set the classification threshold to 0.3. This threshold was empirically found to work best on the training set. We then calculated recall, precision and the *F*- measure of the classification as follows.

- Classifier recall ($\text{Rec}_\text{C}$) quantifies the ratio of the ground truth contour that was correctly classified over the complete ground truth contour. It is therefore calculated by dividing the sum of correctly classified contour pixels by the number of all contour pixels. This measure is also called sensitivity.
- Classifier precision ($\text{Prec}_\text{C}$) quantifies oversegmentation. As defining the exact location of the bone from the US images is not trivial (Jain and Taylor 2004), we do not count segmented pixels in the vicinity of the ground truth as oversegmentation. We calculate this measure by first dilating the ground truth with 2 mm, and then take the ratio of the sum of segmented pixels inside the dilated ground truth and the sum of all segmented pixels. This measure is also known as positive predictive value.
- Classifier *F*-measure ($\text{F}_\text{C}$): This measure combines the precision and recall by taking the harmonic mean

$$F = 2\frac{\text{prec}_\text{C} \times \text{rec}_\text{C}}{\text{prec}_\text{C} + \text{rec}_\text{C}} \tag{3}$$

- When we report the average *F*-measure, we first calculate the *F*-measure per image, and subsequently take the average. Because of the non-linear nature of the *F*-measure, it is possible to have a lower average *F*-measure than either the average recall or average precision.

### Experiments

We performed four experiments to assess the performance and robustness of the proposed bone segmentation method. The random forest parameters were set to the default values, unless noted otherwise. The default values were as follows: number of trees $N_\text{trees} = 12$, patch size $D_\text{patch} = 32$ pixels (3.2 mm), $N_\text{p} = 500{*}n_\text{Im}$ and $N_\text{n} = 1000{*}n_\text{Im}$, where $n_\text{Im}$ is the number of training images. The Gini splitting criterion was used during training, ensuring that the new branches after every split contained training samples with the most homogeneous labels possible.

In experiment 1, we assessed the value of the different features in three ways. First, we constructed several classifiers with an increasing number of features, in the order: intensity, LoG, shadow, BB distance, depth, centrality. For the shadow feature, we evaluated four variants (intensity, relative, Karamalis and artifact distance) separately. The feature order was chosen such that the first three features were the standard features used in most heuristic algorithms as well as Foroughi et al. (2007), Cao et al. (2016), and Jia et al. (2016), followed by the features proposed by us in Baka et al. (2016) and the depth feature. This order combines the features that are commonly used in the literature with the complexity of the feature. The questions we intended to answer with this experiment are as follows: Is it worth adding new features and more complexity or are the common features sufficient? Which shadow feature performs the best for bone segmentation? For evaluation, we use data sets 1 and 2. This allowed us to assess the robustness of the features and the classification to new data sets.

Second, we plotted the histograms of the shadow features to illustrate how the positive and negative sample distributions changed with the two data sets. As for every pixel an entire $n \times n$ patch is used as feature vector (see

Fig. 1), we decided to make the histogram for one shadow feature candidate out of the $n^2$. We chose the pixel 1 mm below the ground truth contour, as it is likely to be at the lower border of the bright bone interface and thereby minimizes the effect of bone interface brightness on the shadow value.

And finally, another way to compare the features is by calculating their importance from the forest (Breiman 2001; Louppe et al. 2013). This is possible, because during training, the forest is doing a feature selection. Feature importance was calculated by looking at all tree nodes $t$ in tree $T$ when a feature $F$ was selected as a splitting criterion, and counting how much the cost function $i\ (t)$ changed with this split:

$$\mathrm{Imp}(F,T) = \sum_{t\in T} \sigma(F,t)p(t)\Delta i(t) \qquad (4)$$

Here, $\sigma(F, t)$ is 1 for all nodes $t$ where feature $F$ was selected as splitting criterium, and 0 otherwise, and $p(t)$ is the proportion of samples that reached node $t$. We evaluated the feature importance for a forest with all features, including all four shadow features at the same time.

In experiment 2 we investigated the effect of the number of trees evaluated in the random forest, as well as the patch size and the number of training samples. This experiment was performed on the combined data sets 1 and 2. To assess the effect of number of trees and patch size, the classifiers were trained on training set 1A2A and evaluated on test set 1B2B. We evaluated 3, 6, 12, 24 and 48 trees, and varied the patch size from 8, 16 and 32–64 pixels. To evaluate the effect of training set size, we used cross-validation on the entire 1A1B2A2B data set. This allowed us to evaluate the accuracy with a training set larger than the combined 1A2A. The following training set sizes were evaluated: 13 patients (2-fold cross-validation on 1A2A), 23 patients (2-fold cross-validation on 1A1B2A2B), 31 patients (3-fold cross-validation on 1A1B2A2B) and 41 patients (10-fold cross-validation on 1A1B2A2B).

In experiment 3 we numerically evaluated the classification performance of the proposed method on the test set, and compared its performance with that of other published methods for ultrasound bone segmentation. For the proposed method, the relative shadow and all non-shadow features were used, except depth. The forest was trained on data set 1A2A. The comparison was made on the maximum bone probability gradient location for Hacihaliloglu et al. (2009) and on the dynamic programming-segmented bone surfaces for the methods of Foroughi et al. (2007) and Cao et al. (2016). The dynamic programming segmentation method of Foroughi et al. (2007) was used with the $\beta$ and $\gamma$ parameters set to 0. The parameters $\alpha$ = 0.1, threshold = 0.4 and jumpCost = 1.0 were optimized on the training set for the probability images of Foroughi et al. (2007). The same parameters were used for the proposed method. Cao et al. (2016) published a 3-D dynamic program for surface segmentation. As our images are 2-D, we used the 2-D version of the algorithm. We tested both the in-plane and out-of-plane parameter settings reported in the original article on our training set, as well as the parameter settings used with the method of Foroughi et al. As the latter parameter set performed the best on our images, this is what we chose for the comparison. The parameters for the method of Hacihaliloglu et al. (2009) were taken from the paper. Lastly, as a reference, we also evaluated the patch-based random forest framework with standard Gaussian scale-space features. Gaussian scale-space features are frequently used as a general filter bank, and have performed well in different segmentation tasks (Cheplygina et al. 2014; van der Lijn et al. 2012). In this work, we used a Gaussian scale space with three Gaussian sizes (0.3, 1.0 and 2.0 mm) and all directional and non-directional derivatives to second order.

For experiment 4 we describe the qualitative and semiquantitative results of using the classifier on US images of different parts of the anatomy acquired from different views. For this, we used three additional data sets. First, we acquired 7 transverse US images of the lower spine of a volunteer with our Philips iU22 system. Second, we used 5 knee US images of a volunteer from Jia et al. (2016), acquired with a GE LOGIQ US machine (GE Healthcare, Chicago, IL, USA) and a ML 6-15 probe. Finally, we ran the algorithm on 10 femur images from Penney et al. (2006) (Fig. 6 in their publication). Those images were acquired with a Philips ATL HDI 5000 US machine. For the semiquantitative results, manual contours were drawn on all images, and the recall, precision and $F$-measure were calculated. We still call this semiquantitative because of the small number of images included in the testing. This experiment was meant to assess the robustness of the classifier to different parts of the anatomy and different hardware. We used the classifier trained on the combined training set 1A2A, with the relative shadow and all other non-shadow features.

## RESULTS

Results for experiment 1 are illustrated in Figure 3, which depicts the $F$-measure of the classifier with an increasing number of features, where the newly added feature is shown on the x- axis. In the top row, the test set and training set acquisition parameters are the same, as described in the caption. The bottom row illustrates the cross-testing results when the training set and the

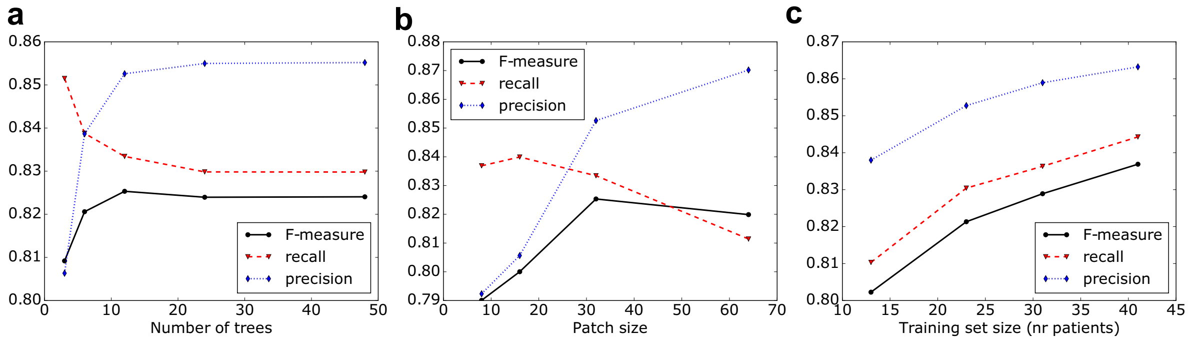

Fig. 6. Results of experiment 2: the effects on accuracy of (a) the number of trees, (b) the patch size in pixels and (c) the number of training data sets.

test set have different acquisition parameters. In Figure 4 are histograms of the shadow features, at a pixel 1 mm below the ground truth contour. Figure 5 illustrates feature importance when a forest with all features is trained on data set 1A2A. The bars represent the average importance from all trees in the forest, and the error bar, the standard deviation.

Results from experiment 2 are illustrated in Figure 6. In Figure 6a is the effect of increasing the number of trees; in Figure 6b, the effect of increasing the patch size; and in Figure 6c, the effect of a larger training set. Figure 6c depicts the evaluation measures with respect to the number of training subjects. For each subject, we have on average 7 images; the number of training images is therefore in the range 91 to 287.

In experiment 3 we compared the proposed method with methods from the literature. In Figure 7 are the bone probability maps and the segmentation with the different methods. The images were chosen randomly from the data set. Quantitative results are presented in Table 1.

In experiment 4 we tested the generalization ability of the proposed patchwise random forest method trained on data set 1A2A. We did this by using test images from of other parts of the anatomy and acquired with different scanners. Figure 8 illustrates some qualitative results on transverse spine and knee images. The probability maps reveal good correspondence with the bone responses in the ultrasound. The semiquantitative evaluation is outlined in Table 2.

## DISCUSSION

We proposed the use of patch-based random forests for the segmentation of bone images from US, together with US-specific feature images. We tested the proposed features and the method by separately evaluating the features, the method parameters, the performance compared with other methods and the robustness and generalization power of the proposed method on different data sets.

We found that the accuracy of the method depends on the features used. The shadow under the bone is one of the best characteristics distinguishing bone from soft tissue interfaces. This is because almost all the sound energy is reflected from the bone surface to the transducer. We evaluated four shadow features and found that the relative shadow proposed in this paper performed best in terms of accuracy as well as robustness, even when acquisition parameters in the training set differed from those in the test set (Fig. 3). It was the only shadow feature that consistently improved results irrelevant of the training and test data sets. This is most probably because the direct effect of image intensity is canceled by removing the row-average intensity during the calculation of the feature. The histograms in Figure 4 confirm that this is the only shadow feature where the distribution of the positive class (bone) is similar in both data sets. For the other three shadow features, the threshold value that best separates positive and negative classes is only applicable for an image acquired with the same protocol. Also, when all shadow features were used in a classifier, the relative shadow came out as most important. Figure 5 illustrates that the most important features for vertebral bone classification were the relative shadow, border-to-border distance and centrality, proposed by us in Baka et al. (2016) and in this article. This is a slightly different order than in Figure 3, as Figure 3 includes the bias from the order of adding the features. When evaluating feature importance, it is also essential to keep in mind that this is subject to the training data set. The centrality feature is very useful if a single bone in the middle of the field of view is the target of the segmentation. It is less useful when bones at the left or right border of the images are also of interest. This is visible in Figure 3d, where the centrality feature actually decreased accuracy. This occurred because data set 1 was acquired with a smaller transducer fitting only one bone in the field of view, whereas in data set 2,

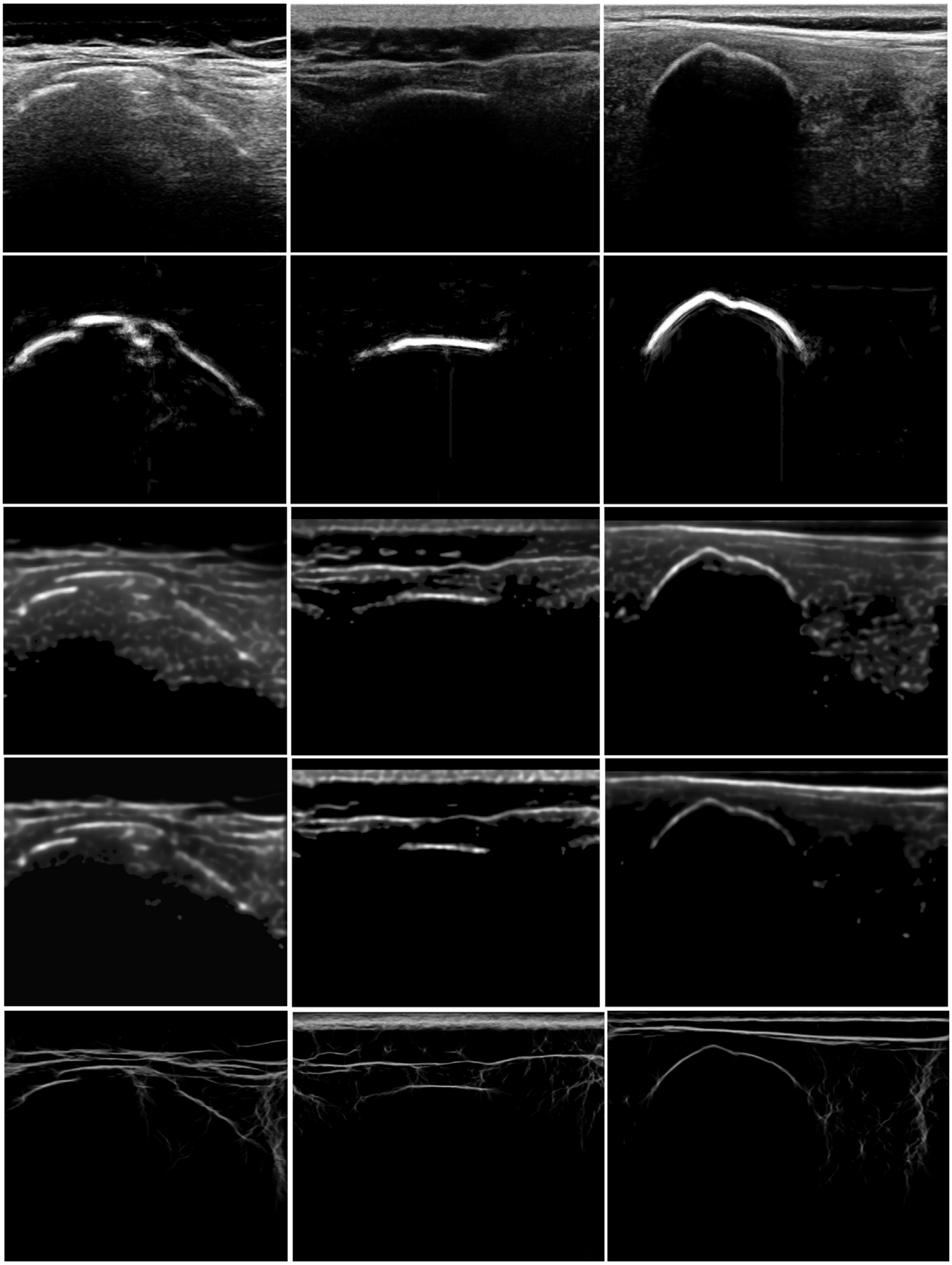

Fig. 7. Bone probability maps created by the different methods. Rows from top to bottom are the original images, random forest, Foroughi et al. (2007), Cao et al. (2016) and Hacihaliloglu et al. (2009). The first image is from data set 1B; the other two are from data set 2B.

several bones fit in the field of view. The least important feature was depth (Figs. 3 and 5). In fact, it often decreased rather than increased accuracy. This feature has been used to erase bright lines at the top of the images (Foroughi et al. 2007; Penney et al. 2006). Apparently, in the proposed method, the other features are sufficient for discarding those areas, and an explicit depth feature is not required.

We evaluated the effect of the two main parameters of the classification method: number of trees and patch size.

Table 1. Results of experiment 3: Evaluation of the proposed method on data set 1B2B, and comparison with other methods on the same data set

| | Segmentation | | | Classification | | |
|---|---|---|---|---|---|---|
| Method | Recall | Precision | *F*-Measure | Recall | Precision | *F*-Measure |
| Proposed method | 0.82 | 0.84 | 0.81 | 0.86 | 0.82 | 0.82 |
| Gauss scale space | 0.75 | 0.71 | 0.69 | 0.81 | 0.74 | 0.74 |
| Foroughi et al. (2007) | 0.53 | 0.44 | 0.45 | — | — | — |
| Cao et al. (2016) | 0.45 | 0.30 | 0.34 | — | — | — |
| Hacihaliloglu et al. (2009) | 0.34 | 0.16 | 0.22 | — | — | — |

The influence of the number of trees diminishes after about 12 trees (Fig. 6a). Evaluation of more trees did not result in any measurable improvement. The appearance of the probability images is somewhat smoother with larger numbers of trees. Figure 6b illustrates that the surrounding information is of great importance for the method. With too small a patch size less information is included, and the precision and *F*-measure are low. This highlights the ability of the random forest to extract useful information from a large pool of features. With too large a patch size (6.4 × 6.4 mm), we observed a slow decline in accuracy. This is most probably due to the insufficient training data for the classifier with increased complexity. Figure 6c illustrates the effect of training set size on performance. Both recall and precision constantly increased while adding new data sets. We thus expect improved results with further enlargement of the training set.

The performance achieved with the image intensity, Laplacian of Gaussian, relative shadow, border-to-border distance, depth and centrality features, was 86% recall at 82% precision (*F*-measure = 0.82). We compared this with values from the literature and found that the proposed method outperforms those methods in the literature by more than 35% (Table 1). We must be careful though when comparing these methods, as many of them were not proposed for the same type of image. For example, the method of Cao et al. was proposed for 3-D hand images, and was tested here on 2-D spine data. One of the largest differences in the data sets is the thicker soft tissue layer on top of the vertebrae than on the carpal bones, with more bright fat–muscle and muscle–muscle interfaces. This is similarly the case for Hacihaliloglu et al. (2009). The automatic parameter settings proposed in Hacihaliloglu et al. (2011) might improve the results, though the method has no way of distinguishing between soft tissue interfaces and bone interfaces, which still remains a major weakness. We also tested a standard filter bank, the Gaussian scale space features, as input for the patch-based random forest. The filter bank performed well, though the US-specific features outperformed the scale space features in both recall and precision (Table 1).

When using trained machine learning methods, it is very relevant to know how well the method generalizes to images acquired using different scanners, under different protocols or simply different parts of the anatomy. We investigated this numerically in experiment 1 and found good robustness for the proposed method on the two data sets used in this paper (Fig. 3). We went further to qualitatively and semi-quantitatively check the classification on images from publications of other groups with

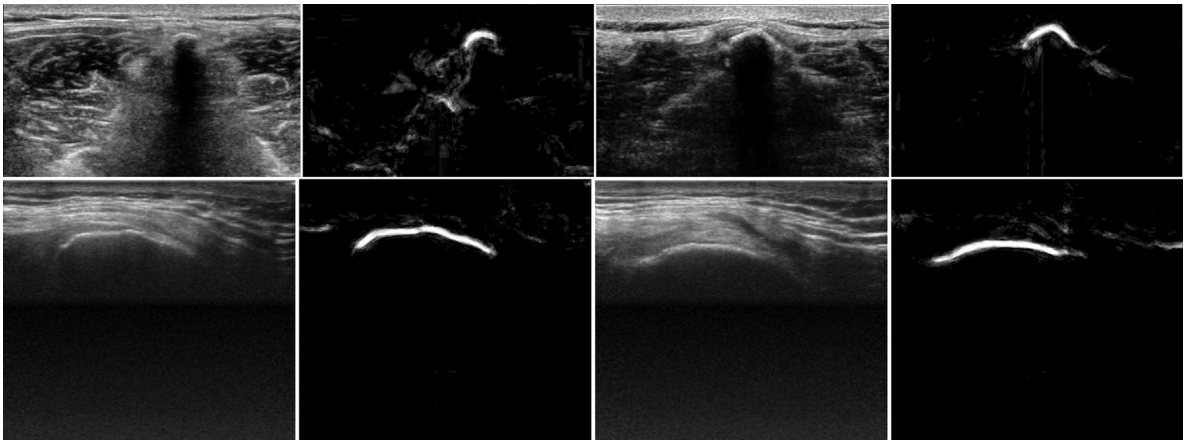

Fig. 8. Bone probability maps created by the proposed method on different parts of the anatomy acquired from different views. Top row: transverse spine. Bottom row: knee images form Jia et al. (2016). The images appear in the original image, probability map order.

Table 2. Results of experiment 4: Semiquantitative robustness evaluation of the proposed method on different parts of the anatomy imaged with different scanners

| Anatomy | Recall | Precision | *F*-measure | US device |
|---|---|---|---|---|
| Transverse spine | 0.99 | 0.70 | 0.79 | Philips iU22 |
| Femur (Penney et al. 2006) | 0.84 | 0.88 | 0.80 | Philips ATL HDI5000 |
| Knee (Jia et al. 2016) | 1.00 | 0.91 | 0.96 | GE LOGIQ |

different scanners and transducers and other parts of the anatomy. We tested knee, femur and transverse spine images acquired with GE and Philips scanners. Figure 7 and Table 2 indicate promising robustness to scanning protocols and different parts of the anatomy.

The random forest classification framework is advantageous when speed is of interest. The computation of the probabilities can be easily parallelized, as every tree can be computed independently. The running time of the proposed method depends on the size of the images, and was around 0.35–0.6 s, when using four parallel threads and the relative shadow. The majority of the time is needed to calculate the feature images. The method was implemented in MATLAB and C++ using Piotr's Computer Vision MATLAB Toolbox (Dollár 2016; Dollár and Zitnick 2013), and run on a Windows PC with Intel i7-6700 CPU with 8 GB memory. Further time gains might be achieved by using a structured forest rather than a random forest, as proposed in Dollár and Zitnick (2013).

## CONCLUSIONS

We proposed and evaluated a patch-based random forest framework for bone segmentation from ultrasound images, including a novel shadow feature. We tested the method on several data sets and found that the proposed shadow metric made the method robust to changes in acquisition parameters. We also evaluated the effect of several parameter settings. The method performed favorably compared with existing methods described in the literature. The method thereby has the potential to enable the rapid bone-based ultrasound–CT registration needed for ultrasound-guided interventions.

*Acknowledgments*—We thank Rui Jia and the Biomedical Image Analysis Laboratory of the University of Oxford for sharing the knee ultrasound data set, Graeme Penney from Kings College London for sharing the femur data set, and Ilker Hacihaliloglu for sharing the phase symmetry code.—The work was financially supported by the Dutch Science Foundation STW, Project 14542.

## REFERENCES

Baka N, Leenstra S, van Walsum T. Machine learning based bone segmentation in ultrasound. In: Yao J, Vrtovec T, Zheng G, Frangi A, Glocker B, Li S, (eds). Computational methods and clinical applications for spine imaging. CSI 2016. Lecture notes in computer science, vol. 10182. Berlin/Heidelberg: Springer; 2016.

Berton F, Cheriet F, Miron MC, Laporte C. Segmentation of the spinous process and its acoustic shadow in vertebral ultrasound images. Comput Biol Med 2016;72:201–211.

Breiman L. Random forests. Mach Learn 2001;45:5–32.

Cao K, Mills DM, Thiele RG, Patwardhan KA. Toward quantitative assessment of rheumatoid arthritis using volumetric ultrasound. IEEE Trans Biomed Eng 2016;63:449–458.

Cheplygina V, Sorensen L, Tax DM, Pedersen JH, Loog M, de Bruijne M. Classification of COPD with multiple instance learning. In: 2014 22nd International Conference on pattern recognition. New York: IEEE; 2014. p. 1508–1513.

Dollár P. Piotr's computer vision matlab toolbox (PMT). Available at: https://pdollar.github.io/toolbox/; 2016. Accessed January 15, 2016.

Dollár P, Zitnick CL. Structured forests for fast edge detection. In: 2013 IEEE International Conference on computer vision. New York: IEEE; 2013. p. 1841–1848.

Foroughi P, Boctor E, Swartz MJ, Taylor RH, Fichtinger G. Ultrasound bone segmentation using dynamic programming. Proc IEEE Int Ultrason Symp 2007;2523–2526.

Hacihaliloglu I, Abugharbieh R, Hodgson AJ, Rohling RN. Bone surface localization in ultrasound using image phase-based features. Ultrasound Med Biol 2009;35:1475–1487.

Hacihaliloglu I, Abugharbieh R, Hodgson AJ, Rohling RN. Automatic adaptive parameterization in local phase feature-based bone segmentation in ultrasound. Ultrasound Med Biol 2011;37:1689–1703.

Hellier P, Coupé P, Morandi X, Collins DL. An automatic geometrical and statistical method to detect acoustic shadows in intraoperative ultrasound brain images. Med Image Anal 2010;14:195–204.

Ho TK. Random decision forests. In: Proceedings, 3rd International Conference on document analysis and recognition, Vol. 1. New York: IEEE; 1995. p. 1. 278–282.

Jain AK, Taylor RH. Understanding bone responses in B-mode ultrasound images and automatic bone surface extraction using a Bayesian probabilistic framework. In: Walker WF, Emelianov SY, (eds). Proc SPIE 2004;5373: p.131–142.

Jia R, Mellon SJ, Hansjee S, Monk AP, Murray DW, Noble JA. Automatic bone segmentation in ultrasound images using local phase features and dynamic programming. In: 2016 IEEE 13th International Symposium on biomedical imaging (ISBI). New York: IEEE; 2016. p. 1005–1008.

Karamalis A, Wein W, Klein T, Navab N. Ultrasound confidence maps using random walks. Med Image Anal 2012;16:1101–1112.

Kowal J, Amstutz C, Langlotz F, Talib H, Ballester MG. Automated bone contour detection in ultrasound B-mode images for minimally invasive registration in computer-assisted surgery and in vitro evaluation. Int J Med Robot Comput Assist Surg 2007;3:341–348.

Lim JJ, Zitnick CL, Dollar P. Sketch Tokens: A learned mid-level representation for contour and object detection. In: 2013 IEEE Conference on computer vision and pattern recognition. IEEE; 2013. p. 3158–3165.

Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. In: Advances in neural information processing systems 26 (NIPS 2013). Curran Associates; 2013. p. 431–439. Available at: https://papers.nips.cc/.

Nagpal S, Abolmaesumi P, Rasoulian A, Hacihaliloglu I, Ungi T, Osborn J, Lessoway VA, Rudan J, Jaeger M, Rohling RN, Borschneck DP, Mousavi P. A multi-vertebrae CT to US registration of the lumbar spine in clinical data. Int J Comput Assist Radiol Surg 2015;10:1371–1381.

Penney G, Barratt D, Chan C, Slomczykowski M, Carter T, Edwards P, Hawkes D. Cadaver validation of intensity-based ultrasound to CT registration. Med Image Anal 2006;10:385–395.

Quader N, Hodgson A, Abugharbieh R. Confidence weighted local phase features for robust bone surface segmentation in ultrasound. In: Clinical image-based procedures: Translational research in medical imaging. Lecture notes in computer science, vol. 8680. Springer; 2014. p. 76–83. Available at: http://www.springer.com/gp/computer-science/lncs.

van der Lijn F, de Bruijne M, Klein S, den Heijer T, Hoogendam YY, van der Lugt A, Breteler MMB, Niessen WJ. Automated brain structure segmentation based on atlas registration and appearance models. IEEE Trans Med Imaging 2012;31:276–286.