Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Automated quality assessment in three-dimensional breast ultrasound images

Julia Schwaab
Yago Diez
Arnau Oliver
Robert Martí
Jan van Zelst
Albert Gubern-Mérida
Ahmed Bensouda Mourri
Johannes Gregori
Matthias Günther

SPIE.

Julia Schwaab, Yago Diez, Arnau Oliver, Robert Martí, Jan van Zelst, Albert Gubern-Mérida, Ahmed Bensouda Mourri, Johannes Gregori, Matthias Günther, "Automated quality assessment in three-dimensional breast ultrasound images," *J. Med. Imag.* **3**(2), 027002 (2016), doi: 10.1117/1.JMI.3.2.027002.

Automated quality assessment in three-dimensional breast ultrasound images

Julia Schwaab,[a,*] Yago Diez,[b] Arnau Oliver,[c] Robert Martí,[c] Jan van Zelst,[d] Albert Gubern-Mérida,[d] Ahmed Bensouda Mourri,[e] Johannes Gregori,[a] and Matthias Günther[a,f]
[a]mediri GmbH, Vangerowstr. 18, Heidelberg 69115, Germany
[b]Tohoku University, Tokuyama Laboratory, 6-3-09 Aramaki-Aoba Aoba-ku, Sendai 980-8579, Japan
[c]University of Girona, Campus Montilivi, Ed. P-IV, Girona 17071, Spain
[d]Radboud University Medical Center, Geert Grooteplein Zuid 10, Nijmegen 6525 GA, The Netherlands
[e]University libre de Bruxelles, Franklin Rooseveltlaan 50, Brussels 1050, Belgium
[f]Fraunhofer MEVIS, Universitätsallee 29, Bremen 28359, Germany

Abstract. Automated three-dimensional breast ultrasound (ABUS) is a valuable adjunct to x-ray mammography for breast cancer screening of women with dense breasts. High image quality is essential for proper diagnostics and computer-aided detection. We propose an automated image quality assessment system for ABUS images that detects artifacts at the time of acquisition. Therefore, we study three aspects that can corrupt ABUS images: the nipple position relative to the rest of the breast, the shadow caused by the nipple, and the shape of the breast contour on the image. Image processing and machine learning algorithms are combined to detect these artifacts based on 368 clinical ABUS images that have been rated manually by two experienced clinicians. At a specificity of 0.99, 55% of the images that were rated as low quality are detected by the proposed algorithms. The areas under the ROC curves of the single classifiers are 0.99 for the nipple position, 0.84 for the nipple shadow, and 0.89 for the breast contour shape. The proposed algorithms work fast and reliably, which makes them adequate for online evaluation of image quality during acquisition. The presented concept may be extended to further image modalities and quality aspects. [DOI: 10.1117/1.JMI.3.2.027002]

Keywords: automated breast ultrasound imaging; image processing; machine learning; image quality.

Paper 15195RRR received Oct. 5, 2015; accepted for publication Mar. 30, 2016; published online Apr. 25, 2016.

1 Introduction

Three-dimensional (3-D) automated breast ultrasound (ABUS) is gaining importance in breast cancer screening programs as an adjunct to x-ray mammography.[1] It has been shown that its use may lead to early detection of small invasive cancers that are occult on mammography in women with dense breasts.[2–4] Furthermore, ABUS is a radiation-free technique, which is relatively inexpensive and effective, since images are acquired by technicians and interpreted later by radiologists—in contrast to hand-held ultrasound, which needs to be performed by experienced clinicians.

However, the quality of the images highly depends on the acquisition procedure. Bad skin contact or slight misplacement of the transducer during ABUS acquisition produces imaging artifacts, which may obstruct a complete diagnostic evaluation. This may lead to a recall of the woman for subsequent additional imaging, which increases screening costs. Recall rates of up to 19% due to BI-RADS category 0 rated images (Breast Imaging Reporting and Data System of the American College of Radiology) have been reported,[5] which means that these images were incomplete or of low quality and that a possible abnormality could not be clearly seen or defined. These numbers can be explained by the fact that technicians need some time to train before they are able to produce artifact-free images since the positioning of the transducer frame is an essential factor for image quality. Automated image quality assessment (AQUA) could support the technicians in recognizing image artifacts during or directly after image acquisition. By doing so, technicians could repeat the scan with corrected parameters while the woman is still in the examination room. In the described scenario, correctly detected artifacts would help to anticipate and potentially avoid recalls caused by insufficient image quality.

While there are several studies investigating image quality assessment of (breast) MRI[6–8] and of hand-held ultrasound images,[9] very little work has been performed to investigate image quality assessment of ABUS images. In Ref. 10, an algorithm to reduce motion artifacts in ABUS images based on nonlinear registration was developed. Generally, ultrasound image quality is considered from the technical point of view. The descriptions in Refs. 11–13 focus on the functionality of the equipment (beam former, transducer) but not on the usage of the system in daily routine. More recently, we investigated the incidence and influence of diverse ABUS artifacts in a reader study.[14] In that previous work, the investigated artifacts had been defined by radiologists, technicians, and physicists, aiming at those that were disturbing diagnostics. In the present work, we concentrate on three of the most relevant aspects that could be avoided in the majority of cases by rescanning: the acoustic shadow caused by the nipple, the position of the nipple relative to the rest of the breast in the image, and the shape of the breast contour on the image. If we manage to achieve high specificity in artifact detection, avoiding unnecessary rescans,

*Address all correspondence to: Julia Schwaab, E-mail: j.schwaab@mediri.com

such a tool could not only lower the number of recalls that cost time and money but also help to train the technicians.

The contribution of this work is the development of an automated image quality assessment system to automatically detect the previously mentioned artifacts. Such a system will support technicians during image acquisition by giving a warning if imaging artifacts disturbing the clinical interpretation of the images are present. A repetition of the affected scans with corrected parameters can then be performed while the patient is still in the examination room.

2 Background

2.1 *Automated Breast Ultrasound Imaging*

ABUS images are acquired by a wide linear array ultrasound transducer sliding continuously over one breast, which is gently compressed by a dedicated membrane while the patient lies in a supine position. During the sliding motion of the transducer, the ultrasound scanner acquires more than 300 transversal images covering a large segment of the breast. These single slices are stacked to form a 3-D ultrasound image that can be examined in multiplanar reconstructions.[15] Depending on the size of the breast, three to five views of each breast are acquired. The positioning and compression of the breast are standardized to some extent and include anterior–posterior (AP), lateral (LAT), medial (MED), superior (SUP), or inferior (INF) views, the breast being gently pushed in these directions, respectively. The latter one (INF) is acquired very rarely and was not contained in our datasets.

2.2 *Automated Breast Ultrasound Image Quality Aspects*

The focus was put on the three most frequent quality aspects that could be avoided by a repeated scan. The first problem is an incorrect nipple position within the image. In some cases, the nipple is pushed very close to the edge of the breast in coronal view [see Fig. 1(a)]. This may cause severe posterior acoustic shadows, obscuring anatomical structures behind the nipple, which can usually be avoided by proper repositioning of the transducer. The second issue is the shadow of the nipple [see Fig. 1(b)]. In the area around the nipple, there is commonly no perfect contact between transducer and skin, resulting in an acoustic shadow behind the nipple on the ultrasound image. Air-filled ducts may contribute to this effect. In most cases, the image is nevertheless usable for diagnostics, but sometimes, the shadow covers noteworthy parts of the breast tissue. Applying more contact gel in a repeated scan often resolves this problem. The third aspect is also correlated to the positioning of the transducer and the breast. If the breast is not supported correctly by the provided cushions, there might be a lack of contact and the outer regions of the breast will not be not imaged [see Fig. 1(c)]. This results in large background areas in the image as well as irregular breast contour lines.

2.3 *Automated Image Quality Assessment*

In this work, we propose an automated image quality assessment system checking the images during or directly after the acquisition. The current standard and the proposed additional workflow step are indicated in Fig. 2. The early automatic detection of image quality issues will initiate a repeated acquisition if indicated. This will only take a few minutes. If a problem that disturbs the diagnosis was only detected later by the radiologist, the woman would have to be recalled, which would take several days. In order to build a convenient application for clinical practice, we first gathered expert definitions of artifacts and had real image data annotated by clinicians. Approved image processing algorithms were employed to extract features characteristic of distinct quality aspects. Feature design was based on a training

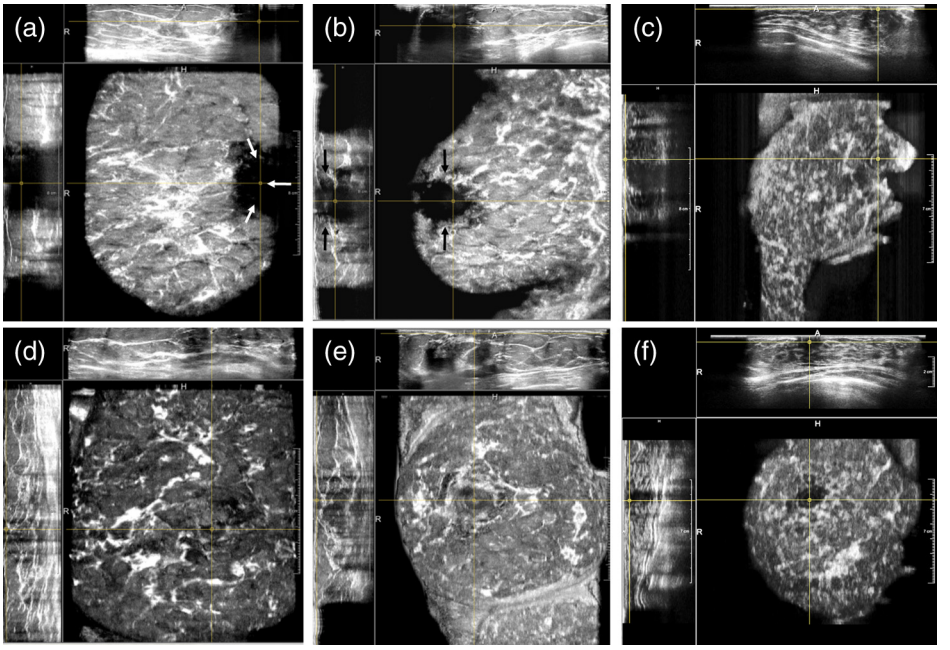

Fig. 1 Sample ABUS images with artifacts. (a) Nipple is too close to the contour of the breast, (b) nipple shadow is too prominent, (c) breast contour is too irregular. (d)–(f) Correctly acquired images of the same breasts as in (a)–(c), respectively. For each example, original transversal view (top) as well as reconstructed sagittal (left) and coronal views (main) are shown.

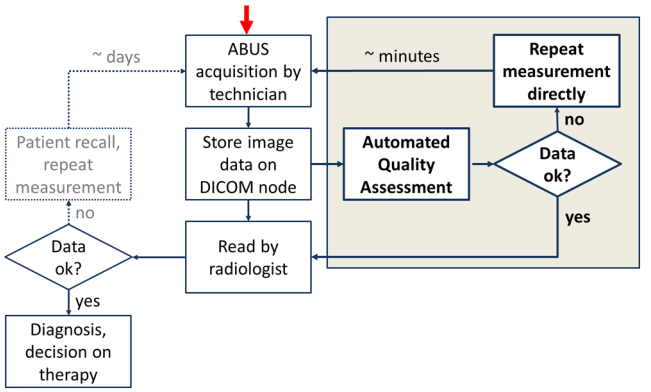

Fig. 2 Workflow diagram for general ABUS screening application. Automated image quality assessment (gray box) could be performed directly after the data were sent to the PACS system.

dataset (dataset A, introduced below) and aimed at translating the physical properties of the artifacts into computable values taking into account the radiologists' descriptions. Classifiers were used to reproduce the manual annotations based on the most meaningful subset of available features. In order to be used in clinical environments and produce results before the patient leaves the facility, the algorithms had to have a low run time (a few minutes at most). Similarly, in order not to produce unnecessary inconveniences for patients, the false positive rate was sought to be very low (clearly below 10%).

3 Methodology

The presented software development approach is based on machine learning and evaluated against the expert assessment of two clinicians. In what follows, first we explain the features computed for the detection of each image quality aspect and subsequently present the learning algorithm used. All image processing routines were implemented in C++ using the open source National Library of Medicine Insight Segmentation and Registration Toolkit (ITK, www.itk.org). All computations were performed on a Windows 7 machine with an Intel® Core™ i7-2627M processor at 2.7 GHz and with 6 GB of RAM.

3.1 *Relative Nipple Position*

The position of the nipple relative to the rest of the breast in the image is important because it relates to acoustic shadows that hamper the clinical interpretation of the image. The absolute nipple position in the image was given by the technician during image acquisition and stored as a private DICOM tag as specified by the standard acquisition protocol. The ABUS images were prepared for feature extraction in several preprocessing steps. First, a two-dimensional (2-D) coronal breast mask was computed similarly to the approach proposed by Wang et al.[16] Therefore, a coronal mean projection of a stack of 120 slices close to the skin was performed. However, the top 50 slices from the skin were excluded from the breast mask computation to avoid responses from skin tissue. The projection image was smoothed using a Gaussian filter with a sigma of 0.2 mm and binarized by applying Otsu's thresholding method.[17] In order to close holes within the breast mask or along its edges, the binary image was dilated and holes were filled before it was eroded again. Finally, the breast contour line was computed in 2-D based on the mask image, as shown in Fig. 3. Note that the nipple coordinates x_T and y_T are generally assumed to be the same for all coronal slices, and the z-coordinate of the nipple is always on top of the image, since there is direct contact between transducer and nipple. The breast contour line is the same in all slices due to the compression of the breast and the properties of ultrasound. Using this contour and the given nipple position, nine features were extracted.

- c_{view}: The view of the considered image strongly influences the absolute nipple position and may affect the impact of a nipple being close to the contour line of the breast. Thus, a categorical feature c_{view} that can be one of the four available standard views (AP, LAT, MED, SUP) was extracted from the information provided in the header of the DICOM file.
- x_T and y_T: The given nipple coordinates (x_T, y_T), which are the same for all coronal slices, were considered possibly important features since the absolute nipple position in the image may correlate with the position relative to the breast. As the appearance of ABUS images differs a lot depending on the breast size and the transducer position, the absolute nipple position is, however, not coupled directly to the nipple position relative to the breast image.
- d_{min}: The shortest Euclidean distance d_{min} between the nipple position (x_T, y_T) and the breast mask contour line was computed.
- c_{io}: It was determined whether the nipple was located inside or outside the breast mask. The latter case can occur when the shadow around the nipple is very dark and close to the breast contour such that this region is, by mistake, not included in the breast mask. A categorical feature $c_{\text{io}} \in \{1, -1\}$ was included.
- d^*_{min}: The signed distance between nipple position and contour line was computed as $d^*_{\text{min}} = d_{\text{min}} \cdot c_{\text{io}}$.
- A_B: The total 2-D physical area of the breast A_B was computed using the pixel size and the number of pixels within the breast mask.
- $A_{B/I}$: The ratio of the physical 2-D area of the breast to the total image size was calculated as $A_{B/I} = A_B / A_{\text{Image}}$.
- d_{COM}: The center of masses $(x_{\text{COM}}, y_{\text{COM}})$ of the breast area and the Euclidean distance d_{COM} between $(x_{\text{COM}}, y_{\text{COM}})$ and (x_T, y_T) was determined.

3.2 *Nipple Shadow*

In order to estimate the size of a possible nipple shadow, it was assumed that the shape of the shadow could be approximated by a cylinder around the nipple with the axis going in the anteroposterior direction. As the nipple is (approximately) a disk in the coronal plane, once it has stopped the US wave, it produces a cylindrical acoustic shadow. The nipple position (x_T, y_T) was obtained from the DICOM header as given by the technician during acquisition. The size of the dark cylindrical region around the nipple position was estimated by counting cylinder segments (rings) that had low pixel intensity. The radius of the different cylinder segments varied from 4.0 to 20.0 mm in steps of 4 mm (see Fig. 4). In the anteroposterior direction, the height of each cylinder segment was ~2.0 mm. The highest layer was positioned starting at 6 mm below the skin, avoiding potentially

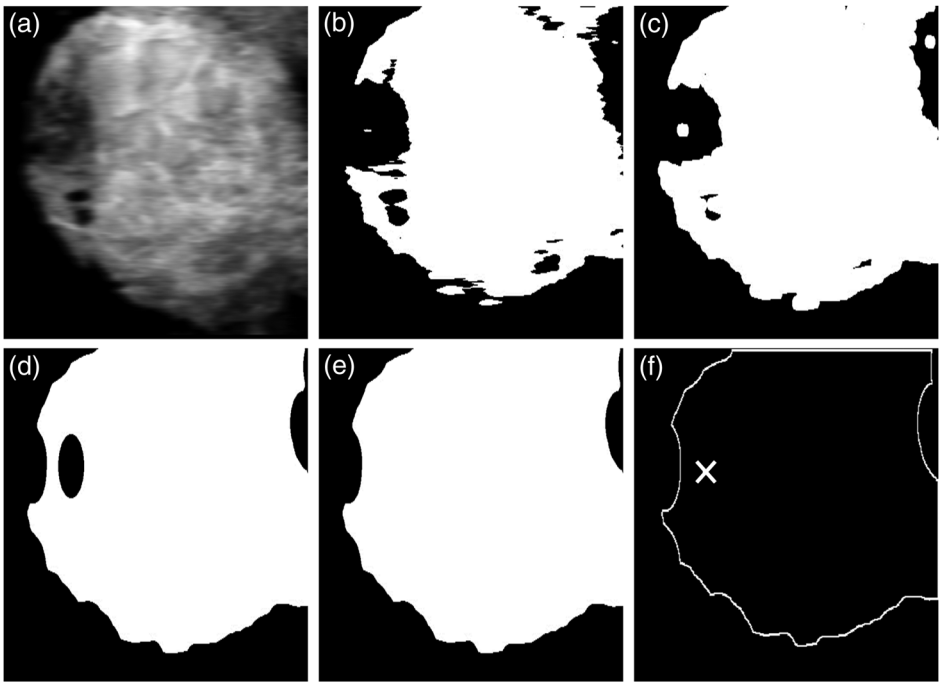

Fig. 3 Steps toward nipple position classification: (a) smoothed coronal projection, (b) binary mask after Otsu threshold, (c) dilated mask, (d) closed mask, (e) holes are filled, (f) eroded and contoured mask with a marker at the nipple position (set by technician during image acquisition).

disturbing high-intensity signals due to skin fat or sound reflections within the coupling layers of the transducer. The deepest layer ended at 26.0 mm below the skin. It was empirically determined that these measures were useful to describe the extent of the nipple shadow. The following seven features were extracted:

- c_{view}: The view of the considered image affects the absolute nipple position and the possibilities of supporting the breast properly by cushions.
- x_{T} and y_{T}: The coordinates $(x_{\text{T}}, y_{\text{T}})$ describing the absolute position of the nipple in coronal plane were included.
- $N_{\text{I<50}}$ and $N_{\text{I<60}}$: The segments showing a lower mean intensity than a specific threshold value were counted. The intensity threshold was set to 50 and 60, respectively, yielding two features, $N_{\text{I<50}}$ and $N_{\text{I<60}}$, for every image. In the present 8-bit grayscale images, these threshold values yielded reasonable differentiation between tissue and shadow signals.

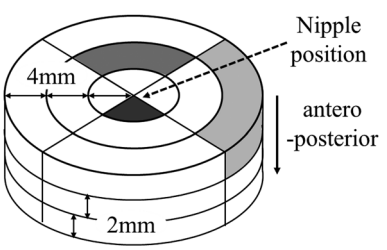

Fig. 4 Arrangement of cylinder segments that were used to estimate the size of a nipple shadow. The symmetry axis was at the nipple position. Three out of ten used layers and three out of five used rings (radii) are shown.

- N_{Pix}: The amount of pixels N_{Pix} in the cylinder segments that had a mean intensity below 60 was counted. This number accounted for the different sizes of the considered cylinder segments.
- σ^2_{bright}: The variance σ^2_{bright} of brightness in one cylindrical region of 4.0 mm radius around the nipple was calculated since ultrasound shadow signals tend to have a lower variance than signals reflected from structured tissue. The cylinder went from the skin to a depth of 25.0 mm in the anteroposterior direction.

3.3 *Breast Contour Shape*

In order to extract the breast mask and its contour line, several preprocessing steps were performed. They were similar to those described in Sec. 3.1 but with a focus on the breast contour line. A 4.0-mm stack of coronal slices starting at a distance of 7.0 mm from the skin was used for breast mask generation. The top 7.0 mm of coronal slices were excluded since they often contain spurious signals caused by sound reflections within contact fluid on parts of the transducer that do not have skin contact. Coronal slices lying deeper than 11.0 mm were not included in order to avoid signals from the ribs that can already appear from this depth on, depending on the breast size and the transducer positioning. A total of 17 features were extracted:

- c_{view}: The view direction was taken into account since breast positioning and cushion support depend on the intended view.
- A_{B}: The physical area A_{B} in 2-D coronal view of the breast mask was assessed as a first indicator for the amount of tissue being imaged.

- $A_{B/I}$: The relative size of the breast mask $A_{B/I} = A_B/A_{Image}$ compared to the total size of the image was computed. The higher this value, the higher the probability that the breast was imaged completely.
- x_C and y_C: The position (x_C, y_C) of the breast mask centroid was computed as an indicator for the position and "mass distribution" of the breast within the image.
- l_1, l_2, and F: An ellipsoid was fitted to the breast contour line, and the lengths l_1 and l_2 of the ellipsoid axes were determined. The flatness F was computed as the ratio l_1/l_2 to indicate whether the breast contour was extremely elongated in one direction or rather roundish.
- p_{Mask}: The perimeter p_{Mask} of the breast mask was determined and corresponded to the length of the breast contour line. The higher the p_{Mask}, the more curves and irregularities might be in the contour line.
- p_{Circle} and r_{Circle}: The perimeter p_{Circle} and the radius r_{Circle} of a circle that has the same surface as the breast mask were computed.
- N_{Border} and p_{Border}: The amount of pixels N_{Border} that belong to the breast mask and are touching the edges of the image, as well as the physical length p_{Border} of these pixels (perimeter on border), were measured. The higher these measures, the higher the probability that the imaged breast is very large.
- R_{Border}: The ratio $R_{Border} = p_{Border}/p_{Mask}$ of the breast mask perimeter along the border and the total breast mask perimeter were computed.
- R_{Round}: The roundness $R_{Round} = p_{Circle}/p_{Mask}$ was determined as the inverse ratio between the actual perimeter of the mask and the perimeter of a circle with the same surface. Since the circle is the geometrical shape with the lowest ratio between perimeter and surface, R_{Round} being close to 1 is a strong indicator for a round and smooth breast contour line. If R_{Round} is very small, the determined breast contour line is supposed to be "inefficient," meaning that it has many turns and irregularities.
- p_1 and p_2: The first two principal moments p_1 and p_2 of the breast mask were determined.

3.4 *Learning Step*

We first evaluated each of the three image quality aspects individually and afterward merged all above described features in order to detect images of generally insufficient quality, i.e., a fourth classifier was trained. This joint classification approach was motivated by the fact that a large portion of the positive images was affected by more than one artifact. The manual annotation of two experienced clinicians served as ground truth for classifier training.

Classification tasks were performed on dataset A (introduced below) using the random forests classifier,[18] as provided by the OpenCV library (version 2.4.10).[19] While the number of trees was set to 100, the number of considered random features for decision tree construction was determined internally by the classifier, as proposed in Ref. 18 as $\log_2(M) + 1$, where M is the number of given features. The maximum depth of each tree was set to 15, and the minimum sample count required at each node to be split was set to 10% of the total number of samples. Ten repetitions of 10-fold stratified cross-validation (CV) were conducted to evaluate the performance of the classification. For each repetition, the instances were randomly partitioned into 10 folds under the constraint that images of the same patient were within one fold to avoid bias.

The resulting receiver-operating characteristic (ROC) curves were fitted by a binormal function, as implemented in MATLAB and Statistics Toolbox (Release 2011a, The MathWorks, Inc., Natick, Massachusetts, United States). First, for each repetition of CV, the merged ROC curve of all 10 folds was computed by sorting all instances into one curve. These were used to determine the mean ROC curve and the 95% confidence interval (CI) of all 10 repetitions. The area under the ROC curve (AUC) was estimated from the fitted curves, whereas single values of sensitivity and specificity were retrieved from the original (unfitted) classifier outputs. To compare the performance of the joint approach to that of the single classifiers, the significance of the difference between the corresponding AUCs was computed as a p-value using the method described in Ref. 20, as well as the Bonferroni correction[21] to account for multiple (three) comparisons. This means that the computed p-values were multiplied by 3 and then compared to a confidence level of $\alpha = 0.05$. The number of actually positive and negative instances was used to compare two ROC curves.

3.5 *Additional Performance Tests*

In order to investigate the robustness and potential overfitting of the trained classifiers, an independent test dataset (called B) was employed. These data were acquired in a different clinic and manually annotated by a different reader group than dataset A. After training the four classifiers on the complete dataset A, they were applied to dataset B and compared to the manual rating results. Classifier decision thresholds were chosen such that the specificity in the training step was 97%. To obtain statistics, the data were bootstrapped 100 times. The data were also used to examine the difference between the joint classification, which is based on all features at once, and the straightforward combination of the three single classifier outputs to a combined rating. Ground truth for this comparison was the combination of the manual annotations, i.e., if at least one artifact was detected concordantly by both readers, the case was considered positive. As an additional performance measure, the inter-rater agreement between the two readers (R2 versus R3) as well as between the automated image quality assessment and the manual rating (AQUA versus R2&R3) was computed as Cohen's κ.[22]

4 Results

4.1 *Dataset*

In total, 815 ABUS volumes acquired from 114 women were obtained in routine clinical care and split up into two datasets, A and B. The images were acquired using either the Somo-v automated 3-D breast ultrasound system (U-systems, Sunnyvale, California) or the ACUSON s2000 ABVS (Siemens, Erlangen, Germany). Details on the size and spatial resolution of the images are given in Table 1. According to the acquisition protocol, the nipple position (in coronal view) was indicated manually by the technicians after each measurement and stored in the DICOM header of the corresponding file so that it could easily be used for further image processing. In some images, the

Table 1 Size and resolution of the ABUS volumes used in this study.

Device manufacturer	Max. image size (cm^3)	Min. voxel size (mm^3)
Siemens	$15.4 \times 16.8 \times 6.00$	$0.21 \times 0.52 \times 0.07$
U-systems	$14.6 \times 16.8 \times 4.86$	$0.29 \times 0.60 \times 0.13$

nipple is not visible at all. These cases were excluded from the analysis of the relative nipple position and the nipple shadow. Detailed description of the datasets is given in Table 2. The Institutional Review Board waived the need for informed consent and approved the use of anonymized images for this study.

All images were classified separately by two clinicians with several years of experience in ABUS imaging. Dataset A was annotated by Readers 1 and 2, whereas dataset B was annotated by Readers 2 and 3; i.e., one reader was the same and one was different for the two datasets. Among others, the above-mentioned quality aspects—nipple position, nipple shadow, and breast contour shape—were taken into account during manual classification.[14] The detailed rating results for dataset A are shown in Fig. 5. The distribution of artifacts was similar in dataset B. Considerable inter-rater disagreement has already been observed in another study dealing with quality rating of ultrasound images.[23] It renders classifier training difficult, but excluding the unclear, i.e., differently rated, cases from the study would mean excluding the critical cases and might bias the results. As the focus of the proposed application was put on a high specificity, we decided to consider only those cases "positive" that were rated as such by both readers. For the joint rating, an image was considered positive if at least one artifact was detected concordantly by both readers. All other cases were considered "negative" and hence usable for diagnostics. Throughout this report, "positive" and "negative" only refer to the rating of the image quality and are not correlated to any diagnostic findings, i.e., tumors or lesions.

4.2 *Relative Nipple Position*

Repeated cross-validation yielded an AUC of 0.99 [see Fig. 6(a)]. Different operating points on the ROC curve can be chosen for the final application by varying the decision threshold for the classifier. Depending on the intended purpose, the user may give more weight to specificity or sensitivity. As summarized in Table 3, at a specificity of 0.99, the sensitivity was 0.36. The point closest to the upper left corner of the plot ("best operating point") represents a specificity of 0.905 ± 0.009 (mean $\pm$ 95% confidence interval) and a sensitivity of 0.93 ± 0.01. For comparison, the performance of each reader when compared to the other, respectively, is displayed in the plots. It can be seen that the automatic classification performed very similarly to the readers. This is a general trend that also accounts for the other considered quality aspects. In Fig. 7, extreme outlier cases are shown. A false positive case is shown in Fig. 7(a), where the breast is very large and not completely visible in the image. In this case, the breast mask fails to describe the true contour of the breast. The breast in Fig. 7(b) is small and skinny, which impedes proper ultrasound coupling. As a consequence, a bright rectangle caused by reflections is visible in the upper right corner of the image, and breast mask segmentation using the Otsu filter fails. Figure 7(c) shows a false negative case caused by the irregular breast contour shape of the breast, which in turn produces an erroneous breast mask. The average computing time for all nine features was $3\ s \pm 2\ s$ per volumetric image, whereas the computing time for the classification was in the order of milliseconds (also for the other quality aspects) and thus negligible.

4.3 *Nipple Shadow*

Automatic classification yielded an AUC of 0.84 [see Fig. 6(b)]. At a specificity of over 0.99, sensitivity was 0.24. The best operating point was described by a specificity of 0.82 ± 0.02 and a sensitivity of 0.73 ± 0.02. Figure 8 shows three sample outlier cases. The false positive case in Fig. 8(a) is

Table 2 Training and test dataset properties.

Dataset	Clinic	Device manufacturer	Readers	Number of women	Patient age (mean $\pm$ stdev)	Nipple position known	Nipple not visible
A	1	Siemens	1 & 2	23	49 ± 11	312	19
	1	U-Systems	1 & 2	14	56 ± 9	28	9
B	2	Siemens	2 & 3	67	47 ± 12	394	53

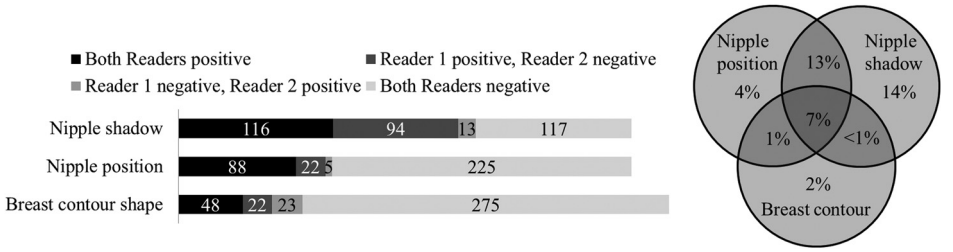

Fig. 5 (a) Incidence of the three considered artifacts in the training dataset. "Positive" and "negative" describe the cases with and without artifacts. (b) Distribution and overlap of artifacts in all positive cases.

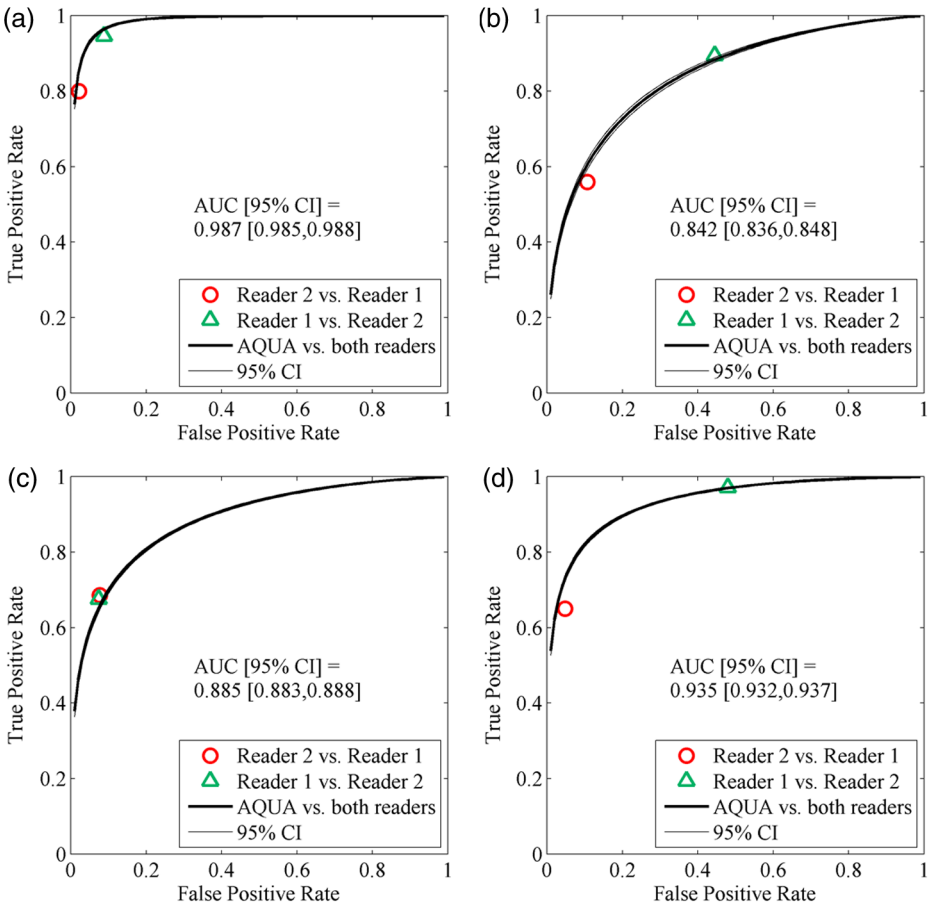

Fig. 6 ROC curve plots (a) for the relative nipple position, (b) for the nipple shadow, (c) for the breast contour shape, and (d) for the joint approach.

Table 3 Summary of classification results. AUC is the mean of 10 repetitions of 10-fold stratified cross-validation. The confidence intervals (CI) are computed from the 10 repetitions as well. *p*-values describe the Bonferroni-corrected (multiplied by 3) significance of difference between the AUC of the single classifiers when compared to the joint approach.

	AUC	95% CI	*p*-value	Sensitivity	95% CI	Specificity	95% CI	TP/all positives	FP/all negatives
Nipple position	0.987	0.002	0.008	0.363	0.064	0.990	0.002	32/88	2/252
Nipple shadow	0.842	0.004	0.004	0.240	0.040	0.991	0	28/116	2/224
Breast contour	0.885	0.003	0.475	0.149	0.025	0.990	9E-14	7/48	3/320
Joint	0.935	0.003	n/a	0.554	0.036	0.990	1E-18	79/143	2/197

a small and skinny breast with a clearly visible nipple shadow close to the breast contour line. However, it was rated as negative by the readers since it is hardly possible to get better images of such a small breast in the present view and a repeated scan probably would not enhance the image. Figure 8(b) shows a false negative case, where the dark region is not directly below the nipple but rather in a half ring around it. In Fig. 8(c), the false negative classification was caused by a relatively bright and fuzzy shadow. However, the algorithm was designed to detect very prominent, low-intensity nipple shadows, as shown in Figs. 3(a) and 3(b). On average, it took 5 s $\pm$ 2 s per ABUS image to compute all possibly relevant features.

4.4 *Breast Contour Shape*

The classification of irregular breast contour shapes achieved an AUC of 0.89 [see Fig. 6(c)]. At a specificity of 0.99, the sensitivity was 0.15. At the best operating point, specificity was 0.82 $\pm$ 0.04 and sensitivity was 0.79 $\pm$ 0.04. Figure 9(a) shows a sample false positive case. The breast as such is imaged correctly, but parts of the axilla and the arm cause atypical contour lines, which are misinterpreted by the classifier. Figures 9(b) and 9(c) show false negative cases, where parts of the breast are not imaged correctly. Nevertheless, the breast mask has smooth contours, obscuring missing parts and misleading the classifier. The average computing time was 6 s $\pm$ 4 s.

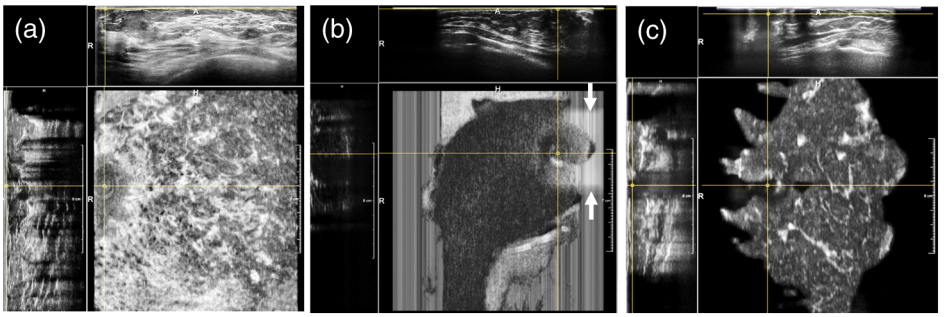

Fig. 7 Examples for outliers of the nipple position classification. (a) A false positive case, where a significant part of the breast is not visible on the scan. (b) A false negative due to an erroneous breast mask caused by intense reflections within the coupling layers of the transducer. (c) A false negative case caused by the irregular breast contour shape.

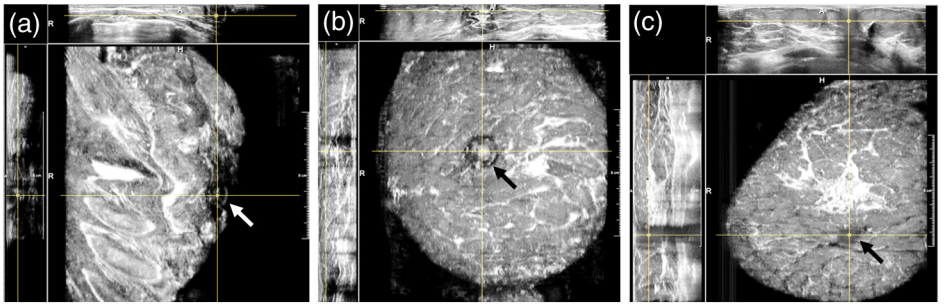

Fig. 8 Incorrectly classified cases of the nipple shadow. (a) A false positive case caused by the nipple being very close to the breast contour line. (b) A false negative with a structured, ring-like nipple shadow. (c) A false negative case with fuzzy and bright nipple shadow.

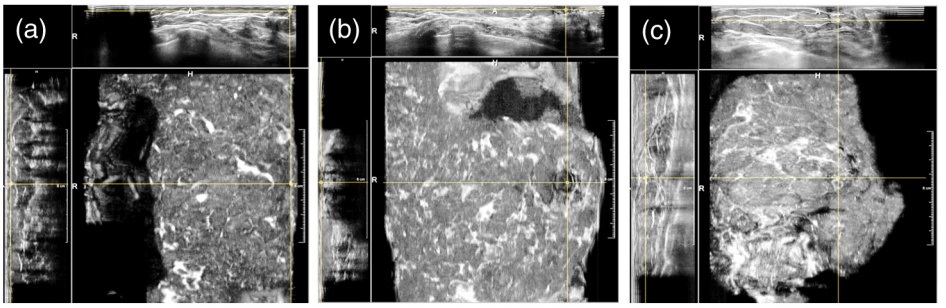

Fig. 9 Examples for outliers of the breast contour classification. (a) A false positive case where parts of the axilla and the arm are visible on the image. The false negative cases in (b) and (c) show relatively smooth contours, obscuring the fact that parts of the breast are not imaged correctly.

4.5 *Joint Classification*

In this approach, the manual ratings for the three distinct artifacts were combined to a joint quality measure. If, according to both readers, at least one artifact was present, an image was assigned the positive class. Random forest classification was based on all computed features at once and achieved an AUC of 0.94 [see Fig. 6(d)]. At a specificity of 0.99, the sensitivity was 0.55. The best operating point had a specificity of 0.91 ± 0.01 and sensitivity of 0.81 ± 0.01. According to the Bonferroni-corrected (multiplied by 3) p-values of Table 3, the AUC of the joint approach was significantly ($p < 0.05$) smaller than that of the nipple position classification and significantly larger than that of the nipple shadow classification. The difference between the AUCs of the joint approach and the breast contour shape classification was not significant.

4.6 *Additional Performance Tests*

Applying classifiers trained on dataset A to the independent test dataset B resulted in the values shown in Table 4. Here we also show the results of a simple combination of classifier outputs (named "combination" in the table) compared to using all features in one single classifier (named "joint"). The "combination" approach combined the outputs of the three distinct classifiers (nipple position, nipple shadow, breast contour) into one rating in the same way as the manual ground truth annotation was combined for the global quality rating: If at least one of the three

Table 4 Rating results retrieved from test dataset. "Combination" refers to the simple combination of the three single rating results, whereas "joint" describes the classifier that was trained on all features at once. R2 and R3 refer to Readers 2 and 3. κ refers to Cohen's inter-rater agreement. Given values are mean (stdev) from 100 times bootstrapping.

	AUC	Specificity	Sensitivity	κ (R2 versus R3)	κ (AQUA versus R2&R3)
Nipple position	0.91 (0.02)	0.91 (0.02)	0.55 (0.08)	0.66 (0.04)	0.41 (0.07)
Nipple shadow	0.86 (0.03)	0.83 (0.02)	0.78 (0.06)	0.75 (0.03)	0.48 (0.05)
Breast contour	0.86 (0.03)	0.94 (0.01)	0.14 (0.10)	0.33 (0.08)	0.06 (0.07)
Combination	n/a	0.78 (0.02)	0.79 (0.05)	0.66 (0.04)	0.44 (0.05)
Joint	0.91 (0.02)	0.81 (0.02)	0.82 (0.04)	n/a	0.50 (0.05)

artifacts was detected by the distinct classifier, the image was rated positive.

5 Discussion

In this work, we presented automatic techniques to assess the quality of ABUS images. The algorithms have short run times and can be applied to the images right after acquisition such that impeded scans could be repeated while the patient is still in the examination room. The focus of the proposed algorithms was on high specificity in order to avoid unnecessary rescans, but depending on the preferred application, other classifier settings could be chosen.

The algorithm to rate the nipple position performed very well. Based on measures such as the distance between nipple and breast contour line, this algorithm had a high specificity and sensitivity at the same time. The good performance may be correlated to the fact that the manual rating of the nipple position was essentially driven by the same parameters as the automatic classification. This means that the clinicians marked the nipple being "too close to the edge of the breast" if the distance between nipple and breast contour line was very small. Exactly the same distance measure, $d_{\min}$, was used as feature for classifier training; i.e., the semantic gap between human perception and computed attributes was very small in this case. Outliers were generally caused by an erroneous breast mask due to irregular breast contour shapes. In some other cases, the algorithm was not able to reproduce the complex decision process that a human reader performs. Even if the above described features were determined as expected, the reader might have anticipated and considered other aspects, e.g., parts of the breast that were not visible in the image, as shown in Fig. 7(a).

The proposed algorithms to detect prominent nipple shadows and irregular breast contour shapes had a similar performance with an AUC slightly smaller than that of the nipple position classification. This might be due to the larger variance in the physical appearance of an acoustic nipple shadow and an irregular breast contour when compared to the relative nipple position. This variance might also be correlated to the larger disagreement between the readers, making this classification an "ill-posed" problem.

Evaluations on independent test data resulted in slightly lower sensitivities and specificities than estimated from repeated cross-validation. Nevertheless, an overall good performance showed that the classifiers were not overfitting in the training step. Note that the images of the test dataset had been acquired in a different clinic and that one of the readers was different than for the training data.

Finally, the joint evaluation of all three artifacts yielded a sensitivity and a specificity of 0.55 and 0.99 in the training dataset as well as 0.82 and 0.81 in the test dataset, respectively, which is promising and justifies the next step toward clinical application. Even if the sensitivity is only moderate, the proposed method has high potential to improve the current standard, as outlined in Sec. 2. As a reasonable number of corrupt images were detected while the specificity of the automatic image quality rating was very high, the technicians could rely on the rating without the risk of producing too many unnecessary rescans. It will be evaluated in clinical practice whether precise information on the kind of the detected artifact is relevant. As shown in Table 4, the inter-rater agreement of R2 versus R3 is in all cases higher than for AQUA versus R2&R3. Nevertheless, the trend of both measures is similar; i.e., if the agreement of both readers is high, the agreement of AQUA versus R2&R3 is also high, showing the classifiers' dependency on the clarity of the manual ground truth annotation. Thus, more readers as well as a clearer definition and separation of the single artifacts might be beneficial. Note that the joint classification approach ($\kappa = 0.58$, AUC $= 0.91$) slightly outperforms the simple combination of single classifiers ($\kappa = 0.44$) as well as the single classification alone (AUC of 0.86 to 0.91). Although the effect is not as pronounced as expected, the joint approach can provide a higher sensitivity and specificity than the single classifiers, however, at the expense of detailed information on a specific quality aspect.

About 28 cases of the training dataset were excluded from the analysis of the first two algorithms because the nipple was not visible at all in the images. In some rare cases, e.g., for very large breasts, this is inevitable. It is, however, unclear how to handle these cases in clinical practice. One possibility is to use an automatic nipple detection algorithm[16] to determine whether the nipple is visible or not. Another option is to find an agreement with the technicians on how to handle the cases where they do not see the nipple in the image. So far, this case is not covered by the standard acquisition protocol.

Apart from the computations described in this work, we also tested the classifier performance by only using those cases that were given the same class by both readers. For all considered artifacts, the sensitivity, specificity, and AUC of the algorithms were slightly higher, showing that the presented approach partly relies on the used image data and the manual annotations. For the nipple position classification, e.g., the AUC was raised to 0.99 leading to a sensitivity of 0.46 at a specificity of more than 0.99. Nevertheless, we decided to include the cases with disagreement as negatives, in order not to bias the data base

by excluding the difficult cases. In another experiment (not explicitly shown in this manuscript), the classifiers were trained only on the clear, i.e., concordantly annotated, cases of dataset A and tested on the unclear cases. The classifier accuracies (sum of true positives and true negatives divided by all cases) when compared to Reader 1 and Reader 2, respectively, were 0.52 and 0.48 for the nipple position, 0.27 and 0.73 for the nipple shadow, as well as 0.45 and 0.55 for the breast contour shape. Thus, for all considered quality aspects, the trained classifiers were consistently more in line with Reader 2 than with Reader 1.

Beneath the random forests, other classifiers like the K* instance-based learner using an entropic distance measure[24] or the J48 decision tree were tested. The latter is an open-source Java implementation of the C4.5 decision tree,[25] which uses normalized information gain (difference in entropy) as a splitting criterion. However, they were outperformed by the random forests yielding the best results in terms of AUC and correlated measures while still being fast enough for the planned application. Furthermore, random forests are robust against overfitting, which was observed in single decision tree classification, and against dependent features.

The average total computing time was determined to be 14 s $\pm$ 5 s per image. Since a typical ABUS examination takes several minutes, the algorithms are fast enough for the planned online feedback application. To our knowledge, there is no previous work that our results could be directly compared to.

According to the manual rating and considering the three discussed image quality aspects, in 40 out of all 83 provided examinations, there was no or only one corrupt image, while in 43 examinations, there were two or more corrupt images. This means that early feedback to the technician after the first scan that showed problems might have helped to avoid another image with incorrect settings. However, throughout this work, the ABUS volumes were considered as independent images. Their correlation to the other images of one examination and the consequences for the usefulness of this examination were not investigated in detail and will be subject to further studies.

Concerning patient age and breast density, no direct influence on the image processing routines or on the image quality rating were detected during this work. Evaluating a first clinical installation of a prototype, it turned out, however, that the breast size has an essential impact on the reliability of the rating of the relative nipple position: in large breasts, transducer positioning often has to be performed such that the nipple is pushed toward the edges of the breast in order to capture the whole breast volume with the available views (AP, MED, LAT, and so on). Therefore, the breast volume is an important additional feature, but computing the actual 3-D volume of the breast based on an ABUS scan is not trivial and, to the authors' knowledge, has not yet been performed completely automatically by any other group. First steps like fully automatic chest wall segmentation have been presented by Tan et al.,[26] who approximated the chest wall by a cylinder. However, computing time was reported to be 6 min and 30 s per breast image, which would be too slow for the application we were aiming at. Thus, extracting information from 3-D images by projecting them to 2-D was more reliable, i.e., reproducible, and reduced complexity and computational costs.

6 Conclusion

In this work, a computerized approach for image quality assessment in ABUS imaging was presented. We have shown that the proposed algorithms have the potential to detect up to 55% of images (at a specificity of 99%) that are currently accepted but present diminished diagnostic value. Apart from the potential to train and support the technicians and to save time and money for patient recalls, the presented algorithms will also help to filter and prepare data for further computer-assisted detection.[26] Although the sensitivity for the single quality aspects is only moderate, the described algorithms are fast and accurate enough to be tested in clinical practice, as the specificity is high, preventing too many false positive cases and unnecessary rescans. In conclusion, by using classifiers, expert knowledge was turned into algorithms that can be used in clinical practice. The contribution of this work is not only to provide a full working application for ABUS but also to test the methodology and the general concept of AQUA software development based on clinical image data. More image quality assessment algorithms for ABUS and other imaging devices such as MRI will be developed in the future in order to complement and upgrade the presented pipeline.

Acknowledgments

The authors would like to thank Jürgen Jenne from Fraunhofer Mevis, Bremen, Germany, for fruitful discussions and support. Parts of this work have been funded by the European Commission's FP7 Cooperation program project ASSURE (project number 306088) as well as by the UdG Grant No. MPC UdG2016/022.

References

1. J. S. Drukteinis et al., "Beyond mammography: new frontiers in breast cancer screening," *Am. J. Med.* **126**(6), 472–479 (2013).
2. M. T. Mandelson et al., "Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers," *J. Natl. Cancer Inst.* **92**(13), 1081–1087 (2000).
3. L. Yaghjyan et al., "Mammographic breast density and subsequent risk of breast cancer in postmenopausal women according to tumor characteristics," *J. Natl. Cancer Inst.* **103**(15), 1179–1189 (2011).
4. K. M. Kelly et al., "Breast cancer detection: radiologist's performance using mammography with and without automated whole-breast ultrasound," *Eur. Radiol.* **20**(11), 2557–2564 (2010).
5. E. K. Arleo et al., "Recall rate of screening ultrasound with automated breast volumetric scanning (ABVS) in women with dense breasts: a first quarter experience," *Clin. Imaging* **38**(4), 439–444 (2014).
6. V. Fiaschetti et al., "Breast MRI artefacts: evaluation and solutions in 630 consecutive patients," *Clin. Radiol.* **68**, e601–e608 (2013).
7. J. A. Harvey et al., "Breast MR imaging artifacts: how to recognize and fix them," *RadioGraphics* **27**(Suppl 1), S131–S145 (2007).
8. T. Kober, R. Gruetter, and G. Krueger, "Prospective and retrospective motion correction in diffusion magnetic resonance imaging of the human brain," *Neuroimage* **59**, 389–398 (2012).
9. N. El-Zehiry et al., "Learning the manifold of quality ultrasound acquisition abstract," in *Medical Image Computing and Computer-Assisted Intervention*, Vol. 13, pp. 122–230, Springer-Verlag, Berlin, Heidelberg (2013).
10. T. Böhler and H. O. Peitgen, "Reducing motion artifacts in 3-D breast ultrasound using non-linear registration abstract," *Lect. Notes Comput. Sci.* **5242**, 998–1005 (2008).
11. N. M. Gibson, N. J. Dudley, and K. Griffith, "A computerised quality control testing system for b-mode ultrasound," *Ultrasound Med. Biol.* **27**(12), 1697–1711 (2001).
12. J. M. Thijssen, M. C. van Wijk, and M. H. M. Cuypers, "Performance testing of medical echo/Doppler equipment," in Y. Lemoigne, A. Caner, and G. Rahal, Eds., *Physics for Medical Imaging Applications*, Vol. **240**, pp. 177–195 (2007).
13. N. Dudley et al., "The BMUS guidelines for regular quality assurance testing of ultrasound scanners," *Ultrasound* **22**(1), 6–7 (2014).

14. J. Schwaab et al., "Image quality in automated breast ultrasound images: a preliminary study for the development of automated image quality assessment," in *MICCAI Workshop: Breast Image Analysis*, 2013, http://www.diku.dk/forskning/Publikationer/tekniske_rapporter/2013/BIA_MICCAI_2013_Workshop_Proceedings.pdf (4 October 2015).
15. J. C. M. van Zelst et al., "Multiplanar reconstructions of 3D automated breast ultrasound improve lesion differentiation by radiologists," *Acad. Radiol.* **22**(12), 1489–1496 (2015).
16. L. Wang et al., "A hybrid method towards automated nipple detection in 3D breast ultrasound images," in *36th Annual Int. Conf. of the IEEE on Engineering in Medicine and Biology Society*, pp. 2869–2872 (2014).
17. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979).
18. L. Breiman, "Random forests," in R. E. Schapire, Ed., *Machine Learning*, pp. 5–32, Kluwer Academic Publishers, The Netherlands (2001).
19. G. Bradski, "The openCV library," *Dr. Dobb's J.* **25**(11), 120–126 (2000).
20. J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**(1), 29–36 (1982).
21. O. J. Dunn, "Multiple comparisons among means," *J. Am. Stat. Assoc.* **56**(293), 52–64 (1961).
22. J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.* **20**(1), 37–46 (1960).
23. C. Keeble et al., "Is there agreement on what makes a good ultrasound image?," *Ultrasound* **21**, 118–123 (2013).
24. J. G. Cleary and L. E. Trigg, "K*: an instance-based learner using an entropic distance measure abstract," in *12th Int. Conf. on Machine Learning*, pp. 108–114 (1995).
25. J. R. Quinlan, *C4.5: Programs for Machine Learning*, in *The Morgan Kaufmann Series in Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California (1993).
26. T. Tan et al., "Chest wall segmentation in automated 3D breast ultrasound scans," *Med. Image Anal.* **17**(8), 1273–1281 (2013).

Julia Schwaab studied physics in Heidelberg, Germany. She did her master's thesis on ultrasound-based motion compensation for imaging and therapy. Lately, she received the Dr.rer.nat. in physics from the University of Bremen, Germany. Her dissertation was on automated image quality assessment in medical images. Currently, she is working in both fields at mediri GmbH, Heidelberg.

Yago Diez, received his PhD in 2008 from Barcelona Tech. Until 2015, he was a Post-doctoral Researcher (VICOROB group) at Girona University (Spain). He is currently an assistant professor at Tohoku University (Sendai, Japan) working on computer vision applications imaging technologies for disaster scenario within the "IMPACT TRC" project. His research interests span from Medical Imaging to Computational Geometry. He has published 10 JCR journal papers and more than 30 articles in peer-reviewed conferences.

Arnau Oliver received the PhD degree in information technology from University of Girona, Girona, Spain, in 2007, where he is currently a lecturer. His research interest is focused on medical image computing, especially on the analysis of diseased brains and the development of automatic tools for early breast and prostate cancer detection. He is author of more than 100 papers in journals and proceedings of international conferences.

Johannes Gregori studied physics in Freiburg, Germany and received the Dr.rer.nat. (equivalent PhD) in physics from Heidelberg University with a dissertation on Magnetic Resonance Imaging techniques in 2010. After several years at Fraunhofer MEVIS, Bremen, and leading positions at mediri, Heidelberg, in the field of image based clinical trial services, he is currently managing director at mediri.

Biographies for the other authors are not available.