Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Three-dimensional segmentation of breast masses from digital breast tomosynthesis images

Stefanie T. L. Pöhlmann
Yit Y. Lim
Elaine Harkness
Susan Pritchard
Christopher J. Taylor
Susan M. Astley

Three-dimensional segmentation of breast masses from digital breast tomosynthesis images

Stefanie T. L. Pöhlmann,^a Yit Y. Lim,^b Elaine Harkness,^a Susan Pritchard,^b Christopher J. Taylor,^a and Susan M. Astley^{a,*}

^aUniversity of Manchester, Division of Informatics Imaging and Data Sciences, School of Health Sciences, Faculty of Biology Medicine and Health, Manchester, United Kingdom

^bUniversity Hospital of South Manchester, Nightingale Breast Centre, Manchester, United Kingdom

Abstract. Assessment of three-dimensional (3-D) morphology and volume of breast masses is important for cancer diagnosis, staging, and treatment but cannot be derived from conventional mammography. Digital breast tomosynthesis (DBT) provides data from which 3-D mass segmentation could be obtained. Our method combined Gaussian mixture models based on intensity and a texture measure indicative of in-focus structure, gray-level variance. Thresholding these voxel probabilities, weighted by distance to the estimated mass center, gave the final 3-D segmentation. Evaluation used 40 masses annotated twice by a consultant radiologist on in-focus slices in two diagnostic views. Human intraobserver variability was assessed as the overlap between repeated annotations (median 77% and range 25% to 91%). Comparing the segmented mass outline with probability-weighted ground truth from these annotations, median agreement was 68%, and range was 7% to 88%. Annotated and segmented diameters correlated well with histological mass size (both Spearman's rank correlations $\rho = 0.69$). The volumetric segmentation demonstrated better agreement with tumor volumes estimated from pathology than volume derived from radiological annotations (95% limits of agreement –16 to 11 ml and –23 to 41 ml, respectively). We conclude that it is feasible to assess 3-D mass morphology and volume from DBT, and the method has the potential to aid breast cancer management. © 2017 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.4.3.034007](https://doi.org/10.1117/1.JMI.4.3.034007)]

Keywords: digital breast tomosynthesis; mass segmentation; tumor size; tumor volume; Gaussian mixture modeling; texture.

Paper 17023RR received Jan. 31, 2017; accepted for publication Aug. 16, 2017; published online Sep. 19, 2017.

1 Background

Lesion size and morphology provide important information for breast cancer diagnosis,^{1,2} disease staging,³ and treatment.⁴ These features help to distinguish between normal radiographically dense breast tissue and benign and malignant lesions, both visually^{1,2} and for computer-based methods,⁵ such as computer-aided detection (CADe), which is designed to detect and draw attention to suspicious image regions.

Accurate assessment of tumor size is also crucial in breast cancer management and helps to stage the disease accurately. Extent of disease has long been acknowledged as an important predictor of patient outcome. In general, the larger the size of the tumor, the higher the likelihood of nodal involvement and, consequently, the worse the prognosis.^{6,7} In staging systems for carcinoma of the breast, such as the tumor–node–metastases system, size of the primary tumor is one of the staging criteria.³

Accurate knowledge of tumor size can help to individualize treatment of women with early breast cancer⁸ and facilitate surgical management, local radiotherapy, or monitoring of neoadjuvant chemotherapy.⁴ It has been shown that wide local excision, where the tumor and a small margin of normal breast tissue are resected, is not disadvantageous for later survival compared to mastectomy.^{9,10} However, inadequate resection margins can make repeat surgery necessary, and this affects 20% to 30% of women undergoing breast conserving treatment in the UK,

with 1.5% to 2.5% of women requiring a third surgical episode to establish clear margins.^{11–13} Local radiotherapy, either as a boost dose to the tumor bed or radiation of the affected area only, has the potential to reduce recurrence rates and spare healthy tissue.⁴ Treatment can be preceded by neoadjuvant chemotherapy to decrease the size of a primary tumor allowing more conservative surgery and limiting reoccurrence risk.¹⁴ Careful monitoring of the tumor extent before, during, and following treatment is essential.¹⁵

1.1 Measurement of Maximum Tumor Diameter and Tumor Volume

It has been reported that manual assessment of maximum tumor diameter from magnetic resonance imaging (MRI) results in better agreement (Pearson's correlation $r = 0.92^{16}$ and 0.80^{17}) with histological diameter measurements of the excised tumor than measurements from two-dimensional (2-D) images, including mammograms ($r = 0.83^{16}$ and 0.26^{17}) or ultrasound ($r = 0.77^{16}$ and 0.57^{17}), where measurements depend on the orientation of the imaging plane. MRI also enables measurement of tumor volume either manually or automatically,¹⁸ which can help to plan treatment¹⁹ or predict patient outcome,¹⁴ but MRI assessment is not routinely used for reasons of cost, availability, and acquisition time.^{4,15}

Digital breast tomosynthesis (DBT) makes use of an x-ray machine similar to that used in mammography but creates

*Address all correspondence to: Susan M. Astley, E-mail: susan.astley@manchester.ac.uk

volumetric instead of 2-D images. Based on a small number (<25) of low-dose exposures (projections) taken while the x-ray tube moves around the breast in an arc of 15 deg to 50 deg (vendor dependent), an image volume is reconstructed. The three-dimensional (3-D) image is displayed in a series of 20 to 100 thin slices of mammogram-like appearance. Due to the low radiation dose, the individual projections show lower contrast and higher radiographic noise than a conventional mammogram. However, as a result of the limited angle over which projections are taken, information for 3-D image reconstruction is incomplete, resulting in anisotropic image resolution.²⁰ Resolution in the plane of the DBT slices is comparable to mammography but is very low in the perpendicular direction. Breast masses appear sharp in the slice in which they are in focus (in-focus slice) [Figs. 1(a) and 1(b)] but produce blurry, fainter repetitions in adjacent slices [Figs. 1(c) and 1(d)] and throughout the image stack [Fig. 1(e)]. Masses look stretched out in the vertical direction [Fig. 1(f)], and in this direction, the extent of tumors cannot be measured directly for narrow sampling angles.²¹

Studies comparing maximum tumor diameter manually measured from DBT slices and mammograms with histology show superior accuracy for DBT (Pearson's correlation with histological measurements $r = 0.85$ to 0.93) compared to mammography ($r = 0.70$ to 0.83).^{17,22,23} However, in a study conducted by Mun et al.,²⁴ a discrepancy of more than 10 mm was found in 19% of 173 breast tumors when comparing maximum tumor diameter measured during a histology assessment with the measurement from DBT images (29% of diameter measurements from mammograms showed more than 10 mm discrepancy).

1.2 Automated Mass Segmentation from Digital Breast Tomosynthesis

Whereas automatic 2-D segmentation of masses has been widely investigated for mammography,^{5,25} little has been published on automatically segmenting breast lesions from DBT images. A CADe system for DBT produced fewer false positive prompts if both 2-D projections and the reconstructed 3-D volume were analyzed.²⁶ Few publications have presented quantitative assessment of the accuracy of mass segmentation from DBT images using only a single representative 2-D slice^{27,28} with the exception of the study by Reiser et al.,²⁹ which maximized the radial gradient index in three dimensions.

Due to the complex 3-D image properties of DBT, it is challenging to extract realistic 3-D mass morphology as confirmed by histological assessment. It is necessary to distinguish between reconstruction artifacts and in-focus mass structure to obtain a 3-D mass segmentation.

2 Aims

The aim of this study is to segment 3-D morphology of masses from DBT images. Segmented masses are evaluated against expert annotations on in-focus slices and 2-D and 3-D histological measurements.

3 Methods

3.1 Image Acquisition

DBT data from 40 breasts with biopsy-proven malignant soft tissue masses were obtained from the Nightingale Breast Centre and Genesis Prevention Centre at the University Hospital of South Manchester. Each DBT dataset comprised of a cranio-caudal (CC) view and a medio-lateral oblique (MLO) view of the affected breast from a Selenia Dimensions Breast Tomosynthesis System (Hologic Inc., Bedford). Fifteen projection images were acquired over an angle of $\pm 7^\circ$ deg, and images were reconstructed as image stacks of 47 to 100 slices (median 67), with each slice image containing 2457×1890 pixels, using standard vendor-provided reconstruction software (version 1.8.3.4). Image resolution is highly anisotropic, with each pixel representing $140 \mu\text{m} \times 140 \mu\text{m}$ planar to the image detector and 1 mm in vertical direction.

3.2 Specimen Handling

Thirty-seven of the 40 masses were surgically excised following imaging. Specimens were marked to indicate original specimen orientation and stored in 10% buffered formalin. After delivery to the pathology department of the hospital, a trained pathologist processed them according to guidelines from the Royal College of Pathologists,³⁰ and masses were sectioned serially in a parallel fashion into 3-mm-thick slices, where possible, along the largest tumor mass cross section. Histology slices were prepared and sent to a consultant breast histopathologist for reporting.

Fig. 1 Appearance of a 10-mm mass in the DBT image stack: (a, b) mass is sharp on central in-focus slices, (c, d) blurred repetitions appear on neighboring slices and (e) fainter repetitions in distant slices, and (f) out-of plane extent is unsharp on a cross section through the DBT image stack.

3.3 Study Dataset

For all DBT views with visible masses, an experienced breast radiologist annotated each mass tumor boundary twice (at least 2 weeks apart) on a slice centrally intersecting the mass. Annotations were drawn with the help of a stylus on a tablet PC, using customized software to automatically save the manual annotations. An approximate indication of mass location and size was selected on the whole DBT slice, and a more accurate annotation of the mass boundary was drawn on a cropped version with a diameter of the mass (as approximately indicated) plus 150 pixels (~16.8 mm) in each direction. This region of interest size was chosen as it allows accurate annotation, when zoomed to full screen size while sufficient surrounding tissue is included. The radiologist classified each DBT view according to mammographic appearance as spiculated mass (SM), ill-defined mass (IM), circumscribed mass (CM), or architectural distortion (AD). Spiculation is a key feature to determine malignant from benign masses³¹ and were included in the annotation, but the long thin wispy spicules were excluded in the analysis. These thin radiating structures are due to fibrosis and AD and can extend over large distances within the breast as seen on mammography. These are not normally included in the measurement for tumor size on mammography and DBT images.^{22,32}

Breast density was estimated from the mammograms, which were available for all patients and taken shortly before the acquisition of the DBT images. A visual estimate of percentage mammographic density was marked on a 10-cm visual analog scale marked 0% at one end and 100% at the other.³³ According to the mean breast density from all four mammographic views (both breasts, MLO, and CC), women were grouped into four classes of breast density according to the American College of Radiology Breast Imaging Reporting and Data System fourth edition (BI-RADS classification system): (a) <25%, (b) 26% to 50%, (c) 51% to 75%, and (d) >75%.³⁴

For 27 masses, 3-D pathological measurements were made from the excised mass specimens, describing the mass diameter in anterior–posterior (d_{AP}), superior–inferior (d_{SI}), and medio-lateral (d_{ML}) directions according to guidelines from the Royal College of Pathologists guidance.³⁰

A consultant breast histopathologist recorded the maximum diameter for each mass measured from the histology slices microscopically. The histological carcinoma subtype was documented as invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), and ductal cancer *in situ* (DCIS).

3.4 Generating 3-D Segmentations

A 3-D segmentation was created from two separate Gaussian mixture models with the help of location information (Fig. 2). After selection of a region of interest, voxels were filtered using a 2-D Gaussian kernel and resulting gray-level intensity used as feature. In addition, gray-level variance was calculated as texture feature for each voxel. Separate probability sets based on both features were built using Gaussian mixture modeling. The sum of the probabilities from the intensity and texture Gaussian mixture model, weighted by the estimated distance from the mass center, constituted the final mass segmentation. All voxels with >50% confidence of being part of the mass structure were selected, and the largest interconnected blob of voxels was considered to represent the breast mass. The segmentation can be initialized using an approximate location supplied by a user (semiautomatic segmentation) or the output of a CAde

Fig. 2 Schematic overview of 3-D segmentation; both intensity and texture features are used to generate Gaussian mixture models that are combined in a confidence-weighted fashion with a location weighting scheme; during the first iteration (run 1) an approximate 2-D segmentation of the central in-focus slice is generated, the diameter of which is used during the second iteration to create the final 3-D segmentation.

algorithm. Following this, steps for generating 3-D segmentations of masses from DBT images are described in detail.

3.4.1 Volume of interest

A DBT volume of interest was selected for each mass based on the approximate size and location as identified by the radiologist on the whole image slice (Fig. 3). In the imaging plane of the DBT slices, a rectangular region was selected, including the mass and at least 150 pixels (~16.8 mm) of breast tissue to all sides; in the vertical direction, all voxels of all available DBT slices were included to constitute the volume of interest. Hundred and fifty pixels outside the tumor boundary allowed good visualization of the tumor margins for annotation but also showed sufficient breast tissue surrounding the tumor for fitting the Gaussian mixture model to the varied breast structures. The breast edge was automatically identified during the reconstruction process by the vendor-provided reconstruction software and background voxels set to a zero value. If the volume of interest contained voxels outside the breast volume, these were excluded. In this case, the breast edge was identified on each slice, and a band of 13 pixels (~1.5 mm) along the breast edge was jointly excluded, to ensure subsequently calculated features are not influenced by the background.

3.4.2 Feature extraction

The gray-level values were filtered slice by slice using a 2-D Gaussian blurring kernel, size 13×13 pixels, to derive an intensity feature. A Gaussian filter was chosen, as it cannot only remove radiographic noise from the image but also smoothes inhomogeneous areas within the tumor and the surrounding breast tissue appropriately. In the dataset investigated, linear structures, such as wispy spicules, blood vessels, and curvilinear structures (e.g., fibrous tissue), which were not included in the mass annotation, measured between 5 and 10 pixels (0.7 to 1.4 mm) in diameter on an in-focus slice. The chosen Gaussian filter was slightly larger than this and reduced the impact of irrelevant linear structures without obscuring the mass boundary [Fig. 3(a)].

Fig. 3 DBT slice with red and green contours indicating the approximate location annotated by an expert radiologist and the outline of the volume of interest (red box): (a) intensity feature and (b) texture feature on a central in-focus slice and more distant slices.

Gray-level variance has previously been used to estimate the focus of an image region in the presence of high noise levels.³⁵ The gray-level variance σ^2 within a region (filter size $m \times m$) is calculated by comparing each gray value g at location i, j with the mean gray value in the neighborhood (size $m \times m$)

$$\sigma^2 = \sum_{i,j=1}^m \left[g_{i,j} - \frac{1}{m \cdot m} \sum_{i,j=1}^m g_{i,j} \right]^2.$$

For DBT stacks, using texture features indicative of an in-focus structure has been used to suppress blurred artifacts in out-of-focus slices.²⁰ Here, a filter size of 13×13 pixels was used, preserving the mass boundary and reducing irrelevant structures, such as out-of focus artifacts [Fig. 3(b)].

3.4.3 Generation of Gaussian mixture models

For the intensity-based segmentation, intensity feature values from the whole volume of interest were collated and a Gaussian mixture model^{36,37} (inspired by Refs. 38 and 39) computed. Three Gaussian distributions were fitted to the data, representing (i) the mass, (ii) potential artifacts from the mass and other radiographically dense tissue, and (iii) fat (Fig. 4). The distributions were found using expectation maximization (100 iterations)⁴⁰ after initialization with k -means clustering.⁴¹ The probability of each voxel within the volume of interest being part of the tumor (p_{int}) was calculated from the mixture model with values between 0 and 1, where 1 is extremely likely to be part of the tumor.

The texture Gaussian mixture model was built in a similar way using the gray-level variance feature. Three Gaussian distributions were found to distinguish between regions with highly pronounced edges, such as the mass margin and spicules, regions where less sharp edges were created by artifacts or fibro-glandular breast structures, and regions that show a homogeneous texture, such as the breast fat. A probability map for each voxel (p_{tex}) was created in the same fashion as for the intensity values.

To generate the location weighting scheme, the estimated location and diameter of the mass as indicated by the radiologist's approximate selection (on the full DBT slice) were used as initialization (not the accurate annotation) making this a semi-automated method. For a fully automated method, the output from a CAD algorithm could also be utilized for this purpose. A 2-D Gaussian filter with sigma half of the approximated diameter was located at the estimated mass center to provide an approximate location weighting. A more accurate location weighting scheme was generated following the evaluation of a 2-D mass segmentation based on this initialization and information from the intensity and texture Gaussian mixture models (Fig. 5). The 2-D segmentation was created using the calculated probabilities based on the in-focus slice of the intensity and texture models. After discarding all pixels with $<50\%$ likelihood being part of the mass, all 2-D blobs were measured, and the largest blob was used to calculate the initial diameter estimation. This iterative process overcomes potential initialization errors. The final location weighting consisted of a 3-D Gaussian with the diameter from the 2-D segmentation plus 15% to ensure inclusion of the whole mass.

Fig. 4 Histogram of the gray-level intensity feature of a representative volume of interest containing a breast mass (blue): fitted distributions representing the mass (red), potential artifacts from the mass and other radiographically dense tissue (green) and fat (black), and the sum of the fitted Gaussians is shown as indication of the fit to the histogram (magenta).

Fig. 5 Confidence-weighted combination of Gaussian mixture models to generate the final 3-D segmentation: the weight of the intensity sub-based segmentation is maximal centrally and declines toward the periphery of the stack, the weight for the texture-based segmentation reduces toward the center, and the location weighting scheme applies to all slices.

3.4.4 Weighted combination

Probabilities from intensity and texture Gaussian mixture models (p_{int} and p_{tex}) were added in a confidence-weighted fashion (using w_{int} and w_{tex}) and further multiplied by the established location weighting scheme (w_{loc}). This ensures segmentation of high intensity areas (intensity-based segmentation), which are in focus (texture-based segmentation) and compact (location weighting scheme) (Fig. 5)

$$p_{3-D} = (p_{int} \cdot w_{int} + p_{tex} \cdot w_{tex}) \cdot w_{loc}.$$

The intensity-based segmentation found mass voxels with high gray levels confidently. This is the case in the center of the mass and on slices where it is in focus, particularly when masses show a dense center (e.g., masses that show a fibrotic center⁴²). However, as each mass was visible to some extent in slices that were out of focus, the accuracy of the intensity Gaussian mixture model declined with a vertical distance from the center of the mass. The intensity-based segmentation alone produced a mass elongated in the vertical direction, with almost constant cross section over the entire image stack (Fig. 5). To account for this, weights were assigned to the intensity Gaussian mixture model such that weighting was maximum in a central in-focus slice; weights declined at increasing vertical distances from this, reducing to 5% of the maximum weight at the slices corresponding to the 2-D estimate of mass diameter (Fig. 5).

In contrast, the texture Gaussian mixture model identified regions of high gray-level variance, such as the boundary of the tumor, where gray levels change abruptly. This texture-based segmentation did not identify blurred artifacts where there is less image texture (Fig. 3). The mass boundary was usually detected, but the texture-based segmentation often showed a hollow tumor center where the breast mass is potentially fibrotic and relatively homogeneous. Therefore, the weight of this Gaussian mixture model was assigned to be low for the central in-focus slice, increasing with a vertical distance to 95% at the slices corresponding to the initial 2-D estimate of mass diameter (Fig. 5).

Both weights w_{int} and w_{tex} are contrary to each other, and the Gaussians have the same width with respect to the vertical direction. However, we noted that allowing the weights to add up to one, which is intuitive here, does not produce ideal results for masses, which lack a solid center, such as ADs. Therefore, the

weighting of the texture-based Gaussian mixture model remained at least 50% in the center (Fig. 5). This allows probability values over 1 and theoretically up to 1.5 to occur in the very center of the estimated mass location; those values are cropped and set to 1.

The use of location information assured compactness and was applied to the whole image stack to conclude the confidence-weighting scheme (Fig. 5).

3.4.5 Thresholding and evaluation of connectivity

To extract the final mass structure, a hard threshold was applied to the confidence-weighted combination of probabilities. All voxels with <50% confidence, represented by a value <0.5, where 0 to 1 is possible, were discarded. The largest connected structure within the volume of interest was deemed to be the mass.

3.5 Evaluation of Results

All correlations were calculated using Spearman's correlation coefficients (ρ) to account for the nonnormal distribution of the data; in particular, the histology and pathology ground truth showed skewed distributions. Subgroup data were analyzed by carcinoma type, mammographic appearance, and breast density group using the Kruskal–Wallis test.

3.5.1 Comparison of the two expert annotations

To compare the two expert annotations (A_1 and A_2) for each DBT dataset, percentage area overlap (PAO) was computed as the intersection of the area enclosed by the first annotation A_1 and the second annotation A_2 , divided by their union

$$PAO = \frac{A_1 \cap A_2}{A_1 \cup A_2}.$$

For clinical decision-making, such as staging or treatment planning, the maximum mass diameter is used, so correlation and 95% limits of agreement were calculated to compare annotation diameters.⁴³

3.5.2 Comparison of the model outline on the central in-focus slice with expert annotations

A probability-weighted PAO (PAO_W) was used to compare the segmentation cross-section area (S) on the annotated slice with the area enclosed by each annotation (A_1 and A_2) (Fig. 6). Areas where both annotations agree were double weighted compared to areas where annotations disagree

Fig. 6 PAO_W compares the intersection of the segmentation cross section (S) and the area of both annotations (A_1 and A_2) with the union of all three regions. Regions where annotations agree were double weighted.

$$PAO_W = \frac{2 \cdot S \cap A_1 \cap A_2 + S \cap A_1 + S \cap A_2}{S \cup A_1 \cup A_2 + 2 \cdot S \cap A_1 \cap A_2 + 2 \cdot A_1 \cap A_2 + 2 \cdot S \cap \overline{A_1} \cap \overline{A_2}}.$$

The correlation of annotation and model diameter was assessed, and 95% limits of agreement were calculated.

3.5.3 Comparison of annotations and 3-D model with histology

To assess 2-D accuracy, we compared the annotations and our segmentation with the largest mass diameter measured from histology. Measurements from the histology specimen were considered as the ground truth. Therefore, to capture the overall maximum diameter based on the annotations and our segmentation, the maximum diameter as either seen on CC and MLO view was used. To measure the maximum diameter from our segmentation, the diameter of the smallest ellipsoid to enclose the 3-D segmentation was calculated. Correlation and 95% limits of agreement were calculated. Pearson's correlation r was calculated for comparability to previous studies, such as Refs. 16, 17, and 22–24.

3.5.4 Comparison of annotations and 3-D model with pathology

To access 3-D accuracy, the volume of the 3-D model and the volume estimated from the annotations were compared with the ground truth volume as derived from pathological 3-D measurement.

The mass diameters d_{AP} (diameter in anterior–posterior direction), d_{SI} (diameter in superior–inferior direction), and d_{ML} (diameter in medio-lateral direction) are found from examination of the excised tumor, allowing calculation of a 3-D

volume V_{3-D} (assuming that the masses were approximately elliptical)

$$V_{3-D} = \frac{4}{3} \cdot \pi \cdot d_{AP} \cdot d_{IS} \cdot d_{ML}.$$

From the 3-D model, volume can be derived either by counting voxels (V_p) or by calculating the minimum ellipsoid, which encloses the model completely (V_e).

Volume was also derived from the annotations that were drawn on an in-focus slice of the CC and MLO view (V_a), neglecting the 3-D information that DBT provides. The largest measured diameter of the CC and MLO annotations ($d_{CC \max}$ and $d_{MLO \max}$) was used to describe tumor breadth and width; the largest of both minor axis lengths ($d_{\min}$) was included to estimate mass volume

$$V_a = \frac{4}{3} \cdot \pi \cdot d_{CC \max} \cdot d_{MLO \max} \cdot d_{\min}.$$

V_a , V_p , and V_e were compared by calculating correlation coefficients and examining 95% limits of agreement.

4 Results

4.1 Mass Characteristics

The histological subtype and mammographic appearance of the masses are summarized in Table 1. Most masses (73%) were histologically confirmed as IDC, 15% were ILC, or mixed

Table 1 Mass characteristics: carcinoma type as defined by histological assessment (of biopsy or where available surgical specimen); visibility, either on one or both diagnostic views; mammographic appearance of each mammographic image, classified by consultant radiologist; category of % breast density as recorded on a visual analog scale.

	Number (%)	IDC ^a	IDC/DCIS ^b	ILC ^c	IDC/IDL ^d	TUB ^e
Carcinoma type (masses)	40	29 (72.5%)	4 (10%)	3 (7.5%)	3 (7.5%)	1 (2.5%)
Visible in both projections	38	27 (93.1%)	4 (100%)	3 (100%)	3 (100%)	1 (100%)
Mammographic appearance (views)	78					
Spiculated	66 (84.6%)	47 (83.9%)	6 (75%)	6 (100%)	5 (83.3%)	2 (100%)
Ill defined	8 (10.3%)	5 (8.9%)	2 (25%)	0 (0%)	1 (16.6%)	0 (0%)
AD	4 (5.1%)	4 (7.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Breast density (women)	40					
<25% (a)	16 (40.0%)	11 (5.5%)	2 (50%)	1 (66.6%)	2 (66.6%)	0 (0%)
26% to 49% (b)	20 (50.0%)	14 (4.8%)	2 (50%)	2 (33.3%)	1 (33.3%)	1 (100%)
51% to 75% (c)	1 (2.5%)	1 (3.4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
>75% (d)	3 (7.5%)	3 (10.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

^aInvasive ductal cancer.

^bInvasive ductal cancer with DCIS.

^cInvasive lobular cancer.

^dMixed invasive ductal and invasive cancers.

^eTubular cancer.

cancers (IDC/ILC), which are known to pose challenges when measuring size.^{22,44} All but two masses, which were close to the pectoralis muscle, were visible in both mammographic views. The majority of breast masses (85%) were described radiologically as speculated, few presented as IMs or ADs, but none were classified as CM. Median mass diameter measured from histology, which is regarded as ground truth^{17,22–24} was 15 mm and ranged from 5 to 45 mm.

4.2 Comparison of the Two Expert Annotations

Representative examples of annotations (A_1 and A_2) are shown in Fig. 7 in green and red, respectively. Only two annotation pairs (2.6%) showed less than 50% PAO, and 49 of 78 (63.8%) of annotations showed at least 75% PAO. Median PAO for the 78 pairs of annotations was 77% but varied widely from 91% (Fig. 7, mass F) to 25% (Fig. 7, mass A).

Median diameter derived from all annotations was 15.9 mm and ranged from 6.8 to 41.1 mm. For the CC and MLO views, comparing the maximum mass diameter showed good correlation for the two sets of annotations (CC: $\rho = 0.95$, $p < 0.001$ and MLO: $\rho = 0.91$, $p < 0.001$) [Figs. 8(a)–8(c)]. The 95% limits of agreement were -4.9 and 4.8 mm (CC) and -6.5 and 5.5 mm (MLO) [Figs. 8(d) and 8(e)]. For the CC images, maximum diameter of the two annotations differed more in larger masses ($\rho = 0.55$, $p < 0.001$) [Fig. 8(d)]. This was not observed for the MLO view images ($\rho = 0.23$, $p = 0.19$) [Fig. 8(e)]. Diameter discrepancy did not appear to be influenced by mammographic appearance (CC: $p = 0.18$ and MLO: $p = 0.78$) [Fig. 8(a)], carcinoma type (CC: $p = 0.42$ and MLO: $p = 0.72$) [Fig. 8(b)], breast density group (CC: $p = 0.33$ and MLO: $p = 0.61$) [Fig. 8(c)], or percent density (CC: $p = 0.20$ and MLO: $p = 0.86$, not shown).

4.3 Comparison of the Segmentation Outline on the Central in-Focus Slice versus Expert Annotations

Representative examples of the cross section of the segmentation on the central in-focus slice (shown in white) in relation to

the expert annotations (green and red outline) are shown in Fig. 9. Only 16 segmentation cross sections (20.5%) showed <50% weighted PAO with the annotations, and 30 of 78 (38.4%) of annotations showed at least 70% PAO.

Median weighted PAO was 68% but varied widely from 88% (Fig. 9, mass F) to 7% (Fig. 9, mass A).

PAO, as calculated from the annotations (Fig. 7), did not predict the concordance (PAO_W) of the segmentation (Fig. 9). One mass (G, Fig. 9) gave a particularly poor overlap of the segmentation cross section with the annotations. This was an IM characterized neither by high gray levels nor by texture, and large macrocalcifications were present in the image volume. In other cases, the presence of macrocalcification did not influence the segmentation. The best overlap was achieved for a mass where the pectoralis muscle edge was included in the volume of interest without adverse influence on segmentation quality (H, Fig. 9).

The maximum diameter measured from the annotations correlates well with the maximum diameter of the model measured on the same slice for both CC and MLO view (CC: $\rho = 0.80$, $p < 0.001$ and MLO: $\rho = 0.91$, $p < 0.001$) [Figs. 10(a) and 10(b)]. The 95% limits of agreement are -8.3 and 7.8 mm for CC views and -8.4 and 8.8 mm for MLO views [Figs. 10(c) and 10(d)]. Similarity of segmentation diameter and expert annotation was not significantly different for any studied subgroup [p -values between 0.07 (carcinoma types, CC view) and 0.90 (mammographic appearance, MLO view)].

4.4 Comparison of Annotations and 3-D Model with Histology

Maximum annotation diameter and maximum segmentation diameter as measured by an enclosing ellipse demonstrated good correlation with the histological measurement (both: $\rho = 0.69$ and $p < 0.001$) [Figs. 11(a) and 11(b)]. The 95% limits of agreement based on the annotations included masses underestimated by 13.9 mm and overestimated by 17.5 mm [Fig. 11(c)]. For the 3-D segmentation, the 95% limits of agreement included masses underestimated by 10.3 mm and overestimated by 27.1 mm compared to histological ground truth [Fig. 11(d)].

Fig. 7 Percentage of annotation pairs overlapping more than a given PAO threshold and representative examples (A to F) of masses with varying PAO, annotation A_1 and A_2 are shown in green and red, respectively.

Fig. 8 Scatter plot of measured annotation diameters comparing first and second annotation, MLO, and CC data plotted into the same chart with line of equality (gray dashed lines), color coded to show (a) mammographic appearance, (b) carcinoma type, and (c) breast density group. (d and e) Bland-Altman plots show the difference in diameter between both annotations of the same tumor versus the average diameter of both, mean difference (dashed gray line) and 95% limits of agreement (solid gray lines), separately for CC and MLO views.

Fig. 9 Comparing the segmentation cross section S with the area of both annotations A_1 and A_2 : percentage of DBT datasets with PAO_W larger than a threshold and representative examples (A to H) of breast mass segmentation. The segmentation outline is shown (white) together with both annotations (red and green), color-coded plots show pixels where segmentation and both annotations overlap (dark red), pixels where one annotation and the segmentation overlap (red), pixels covered by the segmentation but outside both annotations (orange), pixels covered by both annotations but outside the segmentation (green), pixels covered by one annotation but not by the segmentation (blue), and pixels outside the mass (dark blue).

Fig. 10 (a, b) Scatter plots of the maximum annotation diameter and segmentation diameter on the same slice with line of equality (dashed gray line) for the CC and MLO views and (c, d) Bland–Altman style plot showing the maximum annotation diameter (here seen as ground truth) and the difference between ground truth and central in-focus segmentation diameter for the CC and MLO views, mean difference (dashed gray line) and 95% limits of agreement (solid gray lines).

Fig. 11 (a) Scatter plot of the maximum annotation diameter versus histology diameter and (b) the maximum annotation diameter versus segmentation diameter, both with line of equality (dashed gray line). Bland–Altman style plots showing the diameter measured from histology slices (here seen as ground truth) and the difference between ground truth and (c) maximum annotated diameter or (d) maximum 3-D segmentation diameter, for both the mean difference (dashed gray line) and 95% limits of agreement (solid gray lines) are plotted.

Fig. 12 Representative masses and 3-D segmentations from the DBT images.

Discrepancies between annotation or segmentation and histological diameter were not significantly different for any of the evaluated subgroups (carcinoma type, mammographic appearance, and breast density). However, all four masses in women with dense breasts (BI-RADS c and d) were overestimated by >5 mm by both annotations and the segmentation. Only two masses were visible as ADs in both views. Two of the four IMs were underestimated using the annotations (-11.4 and -5.0 mm), but all were sized correctly by the segmentation ($+1.0$ and $+0.33$ mm).

4.5 Comparison of Annotations and 3-D Segmentation with Pathology

Representative examples of 3-D segmentation from the DBT images are shown in Fig. 12. Correlation of the pathology volume with V_a from the annotations, V_p , the voxel-based volume from the 3-D segmentation and V_e , the ellipsoidal estimate, are similar (all: $\rho = 0.55$ and $p = 0.004$ to 0.005) (Fig. 13). However, V_a tends to overestimate pathology volume, 95% limits of agreement include -23 and $+41$ ml; V_p underestimates volumes with 95% limits of agreement between -23 and $+8$ ml. The most reliable volume estimation is V_e with 95% limits of agreement between -16 and $+11$ ml (Fig. 14).

Fig. 13 Scatterplot of the mass volume as measured by the pathologist (V_{3D}) against volumes derived from the DBT images (V_a , V_p , and V_e).

4.6 Analysis of Individual Gaussian Mixture Models

The individual Gaussian mixture models are less accurate than the presented combined segmentation, when assessed on the in-focus slice. Median weighted PAO of the intensity-based segmentation with the annotations is 57% (1% to 81%), whereas it is 47% ($<1\%$ to 73%) for the texture-based segmentation alone (Fig. 15). Often the intensity-based segmentation misses the mass boundaries where intensity decreases. The texture-based segmentation includes the edges, but in 64% of lesions, some parts of the center of the mass where the tumor is more compact and homogeneous are excluded (Fig. 15).

5 Discussion

A method to extract 3-D breast mass segmentation from DBT images, which allows assessment of mass size and morphology, has been developed. Gaussian mixture models based on intensity and texture were found to segment complementary voxels and, in combination with location information, can generate 3-D segmentation of masses. On a dataset of 40 masses, the

Fig. 14 Bland-Altman style plot showing the volume derived from pathology (here seen as ground truth) and the difference between ground truth and image-based volume measurement, mean difference (dashed lines) and 95% limits of agreement (solid lines) are plotted, color codes indicate image-based measurement method.

Fig. 15 Comparing the final confidence-weighted combination segmentation, the intensity-based segmentation only and the texture-based segmentation only with both annotations A_1 and A_2 . Percentage of DBT datasets with PAO_W larger than a threshold and two representative examples of breast mass segmentation.

segmentation outline on a central, in-focus slice was compared to annotations drawn on two occasions by a consultant breast radiologist. Our method outperformed Peters et al.,²⁷ who employed a hybrid active contour model to segment 10 lesions and achieved 52% mean PAO between masses and annotations; whereas ours showed an average of 62% PAO_W (60% and 59% with each annotation separately). Reiser et al.²⁹ reported that 81% of their datasets showed a PAO of more than 40%, we achieved this for 87% of our datasets.

Comparing mass size from our segmentation to the ground truth diameter provided by histological assessment, 2-D accuracy was not significantly different for the computed model than for manual annotations, but human input was limited to an initial approximate indication of the mass location. The achieved Pearson's correlation of histology with our segmentation and annotations ($r = 0.69$ and 0.68) lies below that reported previously for DBT ($r = 0.83$ to 0.86),^{17,22,23} but with 15% mixed and lobular carcinoma and >15% ill-defined breast masses or ADs, our dataset is challenging.

Three-dimensional segmentations demonstrate the complex, potentially lobulated architecture of breast masses, which could potentially be verified in a follow-up study using imaging modalities with full 3-D capabilities, such as MRI or CT. Here, better agreement was noted comparing the segmentation volume with pathology measurements (95% limits of agreement -16 to 11 ml) than for annotation-based volume estimation (95% limits of agreement -23 to 41 ml).

Mammographic appearance of breast masses is very varied, ranging from dense clearly defined or SMs to subtle ADs. Therefore, supervised model-based segmentation or classification algorithms require large training databases to learn adequate lesion representations.^{45,46} Gaussian mixture modeling does not compare feature values with a learned template but separates pixels based on dissimilarity to the surrounding breast tissue in the same image. Although the presented weighting scheme was able to generate 3-D segmentations reliably, transferability to analyze data from the other DBT systems reconstructed using different algorithms may be limited, and independent data to establish ideal weights in an analytical way were lacking. In our study, we used DBT images acquired using the Hologic DBT system, as this features the narrowest acquisition angle on the market, vertical resolution is very limited, and data are most challenging.⁴⁷ In the future, machine learning algorithms, such as deep convolutional networks, could potentially

provide more adaptive ways of analyzing 3-D feature values and generating 3-D segmentations.^{48,49}

We believe accurate 3-D mass segmentation has the potential to improve automatic disease detection and diagnosis. Knowledge of mass volume and 3-D morphology may improve presurgical disease staging and can provide valuable input for personalized treatment planning and monitoring especially in patients where MRI is contraindicated.

Disclosures

The authors state no conflict of interest and have nothing to disclose. The study was approved by the National Health Service Health Research Authority under the Integrated Research Application System project ID 203227 and performed in accordance with the ethical standards of the 1964 Declaration of Helsinki and its subsequent amendments.

Acknowledgments

We would like to thank the charity Breast Cancer Now, which funded this study as part of a doctoral research studentship. The sponsors had no involvement in planning, conducting, or publishing this study.

References

1. C. D'Orsi et al., *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*, American College of Radiology, Reston, Virginia (2013).
2. A. J. Maxwell et al., "The Royal College of Radiologists Breast Group breast imaging classification," *Clin. Radiol.* **64**(6), 624–627 (2009).
3. L. Sobin, M. Gospodarowicz, and C. Wittekind, *TMN Classification of Malignant Tumours*, 7th ed., Union for International Cancer Control (UICC) and Wiley-Blackwell, Oxford (2009).
4. E. Senkus et al., "Primary breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Ann. Oncol.* **24**(Suppl. 6), vi7–vi23 (2013).
5. A. Oliver et al., "A review of automatic mass detection and segmentation in mammographic images," *Med. Image Anal.* **14**(2), 87–110 (2010).
6. B. Fisher, N. H. Slack, and I. D. Bross, "Cancer of the breast: size of neoplasm and prognosis," *Cancer* **24**(5), 1071–1080 (1969).
7. C. L. Carter, C. Allen, and D. E. Henson, "Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases," *Cancer* **63**(1), 181–187 (1989).

8. A. Goldhirsch et al., "Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2013," *Ann. Oncol.* **24**(9), 2206–2223 (2013).
9. U. Veronesi et al., "Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer," *N. Engl. J. Med.* **347**(16), 1227–1232 (2002).
10. J. A. van Dongen et al., "Long-term results of a randomized trial comparing breast-conserving therapy with mastectomy: European organization for research and treatment of cancer 10801 trial," *J. Natl. Cancer Inst.* **92**(14), 1143–1150 (2000).
11. R. Jeevan et al., "Reoperation rates after breast conserving surgery for breast cancer among women in England: retrospective study of hospital episode statistics," *BMJ* **345**, e4505 (2012).
12. L. E. McCahill et al., "Variability in reexcision following breast conservation surgery," *J. Am. Med. Assoc.* **307**(5), 467–475 (2012).
13. M. Dieterich et al., "Re-excision rates and local recurrence in breast cancer patients undergoing breast conserving therapy," *Geburthshilfe Frauenheilkd.* **72**(11), 1018–1023 (2012).
14. S. C. Partridge et al., "Breast MRI measurements of tumor volume predict response to neoadjuvant chemotherapy and recurrence-free survival," *Am. J. Radiol.* **184**, 1774–1781 (2005).
15. National Institute for Health and Care Excellence (NICE), "Early and locally advanced breast cancer: diagnosis and treatment clinical guideline CG80," <https://www.nice.org.uk/guidance/cg80> (February 2009).
16. N. Wasif et al., "MRI versus ultrasonography and mammography for preoperative assessment of breast cancer," *Am. Surg.* **75**(10), 970–975 (2009).
17. A. Luparia et al., "Accuracy of tumour size assessment in the preoperative staging of breast cancer: comparison of digital mammography, tomosynthesis, ultrasound and MRI," *Radiol. Med.* **118**(7), 1119–1136 (2013).
18. M. B. Rominger et al., "Accuracy of MRI volume measurements of breast lesions: comparison between automated, semiautomated and manual assessment," *Eur. Radiol.* **19**(5), 1097–1107 (2009).
19. R. Faermann et al., "Tumor-to-breast volume ratio as measured on MRI: a possible predictor of breast-conserving surgery versus mastectomy," *Isr. Med. Assoc. J.* **16**(2), 101–105 (2014).
20. S. T. L. Pöhlmann et al., "Appearance and understanding of digital breast tomosynthesis images," in *Medical Image Understanding and Analysis Conf. Proc.*, pp. 177–182 (2015).
21. S. Richard and E. Samei, "Quantitative breast tomosynthesis: from detectability to estimability," *Med. Phys.* **37**(12), 6157–6165 (2010).
22. D. Förmvik et al., "Breast tomosynthesis: accuracy of tumor measurement compared with digital mammography and ultrasonography," *Acta Radiol.* **51**(3), 240–247 (2010).
23. L. Meacock et al., "The accuracy of breast cancer size measurement: digital breast tomosynthesis (DBT) vs. 2D digital mammography (DM)," in *Insights Imaging ECR 2010 Book of Abstracts*, p. S306 (2010).
24. H. S. Mun et al., "Assessment of extent of breast cancer: comparison between digital breast tomosynthesis and full-field digital mammography," *Clin. Radiol.* **68**(12), 1254–1259 (2013).
25. A. R. Domínguez and A. K. Nandi, "Toward breast cancer diagnosis based on automated segmentation of masses in mammograms," *Pattern Recognit.* **42**(6), 1138–1148 (2009).
26. H.-P. Chan et al., "Computer-aided detection of masses in digital tomosynthesis mammography: comparison of three approaches," *Med. Phys.* **35**(9), 4087–4095 (2008).
27. G. Peters et al., "A hybrid active contour model for mass detection in digital breast tomosynthesis," *Proc. SPIE* **6514**, 65141V (2007).
28. L. Apffel et al., "Fuzzy segmentation of masses in digital breast tomosynthesis images based on dynamic programming," in *Proc. of the Int. Conf. on Imaging Theory and Applications and Proc. of the Int. Conf. on Information Visualization Theory and Applications (IVAPP)*, pp. 7–13 (2010).
29. I. Reiser et al., "Evaluation of a 3D lesion segmentation algorithm on DBT and breast CT images," *Proc. SPIE* **7624**, 76242N (2010).
30. I. Ellis et al., *Pathology Reporting of Breast Disease in Surgical Excision Specimens Incorporating the Dataset for Histological Reporting of Breast Cancer*, The Royal College of Pathologists, London (2016).
31. E. Provenzano, "Invasive breast carcinoma," in *Early Breast Cancer: From Screening to Multidisciplinary Management*, J. R. Benson, G. P. H. Gui, and T. Tuttle, Eds., 3rd ed., pp. 266–285, CRC Press, Boca Raton, Florida (2013).
32. F. L. Flanagan et al., "Invasive breast cancer: mammographic measurement," *Radiology* **199**, 819–823 (1996).
33. S. W. Duffy et al., "Visually assessed breast density, breast cancer risk and the importance of the craniocaudal view," *Breast Cancer Res.* **10**(4), R64 (2008).
34. American College of Radiology, *Breast Imaging Reporting and Data System (BI-RADS)*, 4th ed., American College of Radiology, Reston, Virginia (2013).
35. S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognit.* **46**(5), 1415–1432 (2013).
36. B. G. Lindsay, "Mixture models: theory, geometry and applications," in *NSF-CBMS Regional Conf. Series in Probability and Statistics*, Vol. 5, Institute of Mathematical Statistics (1995).
37. G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*, M. Dekker, New York (1988).
38. R. J. Ferrari et al., "Segmentation of the fibro-glandular disc in mammograms using Gaussian mixture modelling," *Med. Biol. Eng. Comput.* **42**, 378–387 (2004).
39. S. Aylward, B. Hemminger, and E. Pisano, "Mixture modeling for digital mammogram display and analysis," *Digital Mammogr.* **13**, 305–312 (1998).
40. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B* **39**(1), 1–38 (1977).
41. J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297 (1967).
42. T. Hasebe et al., "Fibrotic focus in infiltrating ductal carcinoma of the breast: a significant histopathological prognostic parameter for predicting the long-term survival of the patients," *Breast Cancer Res. Treat.* **49**(3), 195–208 (1998).
43. M. J. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet* **327**(8476), 307–310 (1986).
44. I. V. Gruber et al., "Measurement of tumour size with mammography, sonography and magnetic resonance imaging as compared to histological tumour size in primary breast cancer," *BMC Cancer* **13**, 328 (2013).
45. J. Dheeba, N. A. Singh, and S. T. Selvi, "Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach," *J. Biomed. Inf.* **49**, 45–52 (2014).
46. G. D. Tourassi et al., "Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information," *Med. Phys.* **30**, 2123–2130 (2003).
47. N. W. Marshall and H. Bosmans, "Measurements of system sharpness for two digital breast tomosynthesis systems," *Phys. Med. Biol.* **57**(22), 7629–7650 (2012).
48. Z. Wu et al., "3D ShapeNets: a deep representation for volumetric shapes," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1912–1920 (2015).
49. A. Prasoon et al., "Deep feature learning for knee cartilage segmentation using a triplanar convolution neural network," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Vol. 8150, pp. 246–253 (2013).

Stefanie T. L. Pöhlmann is an engineer and researcher in the field of medical technology. Focus of her research is 3-D breast imaging, in particular developing advanced clinical applications for digital breast tomosynthesis and surface imaging. She has conducted doctoral research at the University of Manchester and has also received a degree in mechanical engineering (diplom) from Ilmenau Technical University, Germany. She is currently working as a project engineer developing endoscopic instruments.

Yit Y. Lim is a consultant radiologist of 10 years specializing in breast and genitourinary radiology at the University Hospital of South Manchester, UK. His qualifications include MB BCH, MRCS (Ed), FRCR, and EDBI.

Elaine Harkness works in the division of informatics, imaging, and data sciences at the University of Manchester. She has an MSc and PhD degrees in epidemiology and works with a multidisciplinary team investigating the relationship between measures of breast density and the risk of developing breast cancer.

Susan Pritchard is a consultant histopathologist for 12 years specializing in breast, gastrointestinal and head and neck pathology working at the University Hospital South Manchester, UK. Her qualifications

include a BSc (Hons) degree in biomedical science and MBChB (Hons) from the University of Manchester UK, FRCPath (2003), and CertMgmt (HSC) (Open 2007). She is currently working on projects examining stromal response in breast neoplasia and translation studies following on from the OEO2, OEO5, and STO3 gastro-oesophageal cancer trials.

Biographies for the other authors are not available.