Automated breast cancer classification using near-infrared optical tomographic images

James Z. Wang
Clemson University
School of Computing
Clemson, South Carolina 29634

Xiaoping Liang
University of Florida
J. Crayton Pruitt Family Department of Biomedical Engineering
Gainesville, Florida 32611

Qizhi Zhang
University of Florida
J. Crayton Pruitt Family Department of Biomedical Engineering
Gainesville, Florida 32611

Laurie L. Fajardo
University of Iowa
Department of Radiology
Iowa City, Iowa 52242

Huabei Jiang
University of Florida
J. Crayton Pruitt Family Department of Biomedical Engineering
Gainesville, Florida 32611
E-mail: hjiang@bme.ufl.edu

Abstract. An automated procedure for detecting breast cancer using near-infrared (NIR) tomographic images is presented. This classification procedure automatically extracts attributes from three imaging parameters obtained by an NIR imaging system. These parameters include tissue absorption and reduced scattering coefficients, as well as a tissue refractive index obtained by a phase-contrast-based reconstruction approach. A support vector machine (SVM) classifier is utilized to distinguish the malignant from the benign lesions using the automatically extracted attributes. The classification results of *in vivo* tomographic images from 35 breast masses using absorption, scattering, and refractive index attributes demonstrate high sensitivity, specificity, and overall accuracy of 81.8%, 91.7%, and 88.6% respectively, while the classification sensitivity, specificity, and overall accuracy are 63.6%, 83.3%, and 77.1%, respectively, when only the absorption and scattering attributes are used. Furthermore, the automated classification procedure provides significantly improved specificity and overall accuracy for breast cancer detection compared to those by an experienced technician through visual examination. *© 2008 Society of Photo-Optical Instrumentation Engineers.* [DOI: 10.1117/1.2956662]

Keywords: phase-contrast diffuse optical tomography; refractive index image; absorption coefficient image; scattering coefficient image; support vector machine classifier; image segmentation; feature extraction.

Paper 07307RR received Aug. 20, 2007; revised manuscript received Feb. 4, 2008; accepted for publication Mar. 6, 2008; published online Jul. 28, 2008.

1 Introduction

Near-infrared (NIR) diffuse optical tomography (DOT) is emerging as a potential clinical tool for breast cancer detection due to its ability to quantitatively image the high optical contrast that arises intrinsically from molecular and cellular signals generated through the presence of blood, water, and lipid, as well as cellular density, which are the predominate transformations associated with malignancy. Clinical studies conducted at multiple institutions and countries have repeatedly shown that there exist 2:1 and higher absorption contrasts in breast cancers that can be tomographically imaged.[1–10]

Since 1997, our laboratory has developed three complete clinical platforms for NIR optical tomography of the breast, evolving from single-wavelength/2-D to multiwavelength/3-D capabilities. These systems have been used to evaluate the potential of our approach through a series of studies designed to quantify the imaging contrast in the normal and abnormal breast and to provide initial assessments of the operating characteristics of the imaging systems for diagnostic decision-making in the setting of screen-detected breast lesions.[10–14] Specifically, both absorption and scattering properties of breast tissues can be obtained to sensitively distinguish between normal and abnormal breast tissues. Cysts can be clearly differentiated from solid tumors based on these two properties alone. Further, Hb and HbO_2 are two important parameters for enhancing sensitivity, consistent with the finding from Chance et al.[15] However, these imaging parameters available from the current NIR tomography, do not appear to be able to fully characterize breast tissues, resulting in limited sensitivity and specificity. In 2003, it was shown for the first time that refractive index/phase contrast could be used as a new imaging parameter for NIR tomography where refractive index and absorption/scattering parameters were reconstructed using two different algorithms.[16] Our initial clinical results demonstrate that phase-contrast DOT combined with conventional DOT offers considerably improved sensitivity and specificity compared to that by using conventional DOT alone.[13] In addition, our results show that cellular density and size derived from the scattering spectra can characterize the nature of breast lesions more accurately than the scattering property or scattering amplitude/scattering power. Initial results in 14 breast abnormalities show that malignant tumors can be differentiated from benign lesions with high

Address all correspondence to Huabei Jiang, Univ. of Florida, J. Crayton Pruitt Family Dept. of Biomedical Engineering, Gainesville, FL 32611. Tel: 352-392-7943; Fax: 352-392-9791; E-mail: hjiang@bme.ufl.edu.

1083-3668/2008/13(4)/044001/10/$25.00 © 2008 SPIE

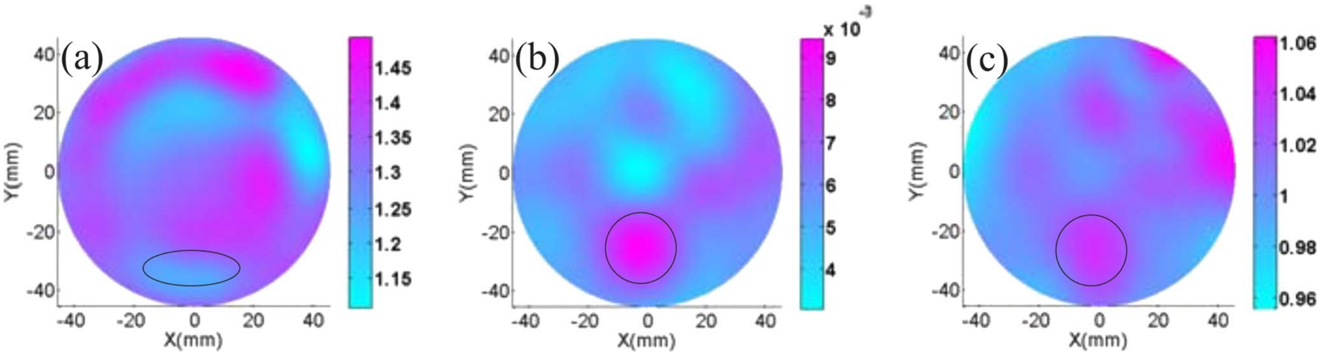


Fig. 1 An infiltrating ductal carcinoma: (a) Refractive index image. (b) and (c) Absorption and scattering coefficient images. The axes (left and bottom) illustrate the spatial scale (mm), whereas the color scale (right) records the refractive index (dimensionless), absorption, or scattering coefficient (mm^{-1}). Circled area indicates the tumor location.

accuracy,[14] meaning that phase contrast and cellular density/size together with the functional parameters can provide a more complete spectrum with much improved sensitivity and specificity for accurate characterization of breast lesions.

Given the relatively large set of imaging parameters now available from our NIR reconstruction approach, it is natural to adapt and develop methods for computer-aided classification of breast lesions. Computer-aided diagnosis has been well studied and widely accepted in the fields of conventional imaging.[17–20] This paper reports our initial effort in applying automatic classification algorithms to the analysis of tissue phase contrast, absorption, and scattering parameters. Our classification results of 35 breast masses using a support vector machine (SVM) classifier demonstrate for the first time that the specificity can be significantly improved from 71% based on the visual examination to 92% when phase contrast, absorption, and scattering parameters are all used.

The rest of this paper is organized as follows. First, the visual examination process for detecting breast cancer is reviewed in Sec. 2. Then, the automated procedure for extracting image features is proposed in Sec. 3. These image features in turn will be used in Sec. 4 to detect breast cancer by an SVM classifier. Last, concluding remarks and future studies are discussed in Sec. 5.

2 Visual Detection of Breast Cancer

2.1 *Image Presentation of Optical Parameters*

Thirty-five breasts from 33 different patients were imaged using a compact, parallel-detection diffuse optical mammography system.[21] The absorption and scattering images were reconstructed using a finite-element-based algorithm,[22–24] while the refractive index images were recovered using a finite-element-based phase–contrast DOT algorithm.[16] All the images were reconstructed using a finite element mesh consisting of 700 nodes, thus giving the refractive indices, absorption, and scattering coefficients at 700 locations evenly distributed across the entire sample area. To enable visual examination of the reconstructed optical parameters, the partial differential equations (PDE) toolbox in MATLAB is used to display the obtained parameters. Figures 1 and 2 demonstrate the coronal refractive index, absorption, and scattering images from two representative patients. The first case (Fig. 1) is a 52-year-old female with a 3-cm infiltrating ductal carcinoma, and the second case (Fig. 2) is a 64-year-old female with biopsy-confirmed benign microcalcifications.

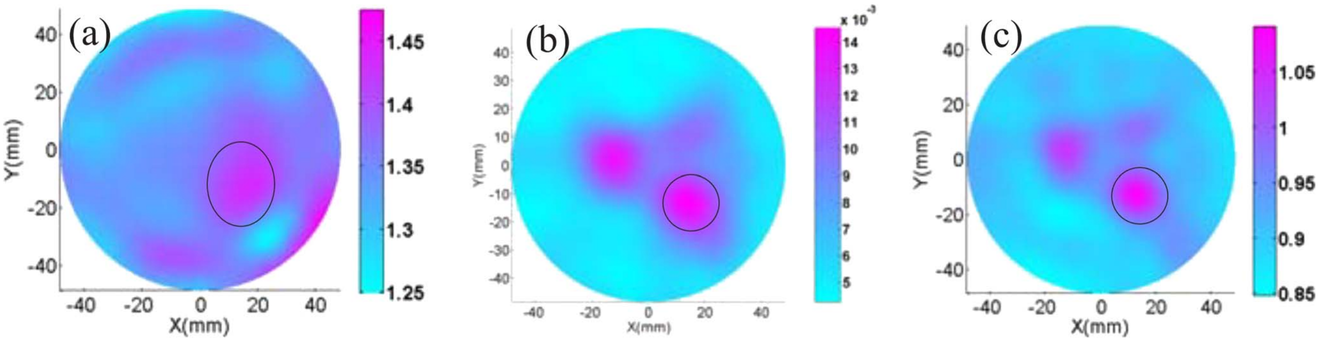


Fig. 2 Benign microcalcifications: (a) Refractive index image. (b) and (c) Absorption and scattering coefficient images. The axes (left and bottom) illustrate the spatial scale (mm), whereas the color scale (right) records the refractive index (dimensionless), absorption, and scattering coefficient, respectively (mm^{-1}). Circled area indicates the lesion location.

Table 1 Statistics of breast cancer detection by visual examination of refractive index, absorption, and scattering images.

True positives	True negatives	False positive	False negatives	Sensitivity	Specificity	Rate of false positive (FPR)	Overall accuracy
9	17	7	2	81.8%	70.8%	29.2%	74.3%

2.2 *Visual Detection*

Using the refractive index, absorption, and scattering images plotted by the MATLAB PDE toolbox, an experienced technician may visually distinguish a malignant tumor from a benign one. For instance, examining the absorption and scattering images shown in Figs. 1(b) and 1(c), the tumor area can be identified at around the coordinate (−1, −28). Inspecting the corresponding area in the refractive index image shown in Fig. 1(a), it is clear that the refractive index in this area is lower than the surrounding area. Thus, this is a cancer case. For the images shown in Fig. 2, the lesion area is identified at the coordinate (10, −15) by checking the absorption and scattering images given in Figs. 2(b) and 2(c). Examining the image shown in Fig. 2(a), the refractive index in the corresponding area is found higher than the surrounding area. Therefore, this is a benign case. However, in these two cases, visually examining only the absorption and scattering images without checking their associated refractive index images cannot distinguish between the malignant case and the benign one. These two examples indicate that it is possible to use the refractive index image in conjunction with absorption and scattering images to differentiate malignant from benign lesions.

To further validate the feasibility of the visual examination method, the absorption, scattering, and refractive index images of 35 breasts were obtained from 33 patients before their biopsy procedures. Using the aforementioned method to classify the images visualized by the MATLAB PDE toolbox, the statistics of the visual examination results over these 35 breast masses [biopsy confirmed 11 invasive carcinomas (malignant group) and 24 benign lesions (benign group)] are presented in Table 1.[13]

While the results shown in Table 1 are promising, the visual examination process has several drawbacks. First, it is time consuming, because the technician has to manually create MATLAB files to visualize the images. Second, the visual identification of malignant lesions depends on the subjective judgment of human beings: physicians or technicians have to be specially trained to make a reliable classification, and human errors may be inevitable. Third, some images, especially the refractive index images, are relatively noisy, and it is difficult to give a reliable classification using visual examination. For instance, the images shown in Fig. 3 can be classified as either a malignant or a benign case using visual examination because half of the lesion area (the corresponding lighter color areas in the absorption and scattering images) has high refractive index while the other half has low refractive index. The technician or physician has to make a best guess based on his/her experience. Therefore, the accuracy of the classification is questionable. Last, only 700 discrete points of the sample area on a triangular mesh were used to obtain the refractive index, absorption, and scattering parameter values (see Fig. 4). These 700 values are then visualized into smooth images, as shown in Figs. 2–4, using the MATLAB PDE toolbox. This visualization process may introduce imprecision to the visual classification of images.

To address these drawbacks, an automatic procedure is proposed in this paper to first directly analyze the image data to extract the classification attributes and then to detect breast cancer using an SVM classifier.

3 Automatic Feature Extraction

The first step of our automated classification procedure is to automatically extract classification attributes from the refrac-

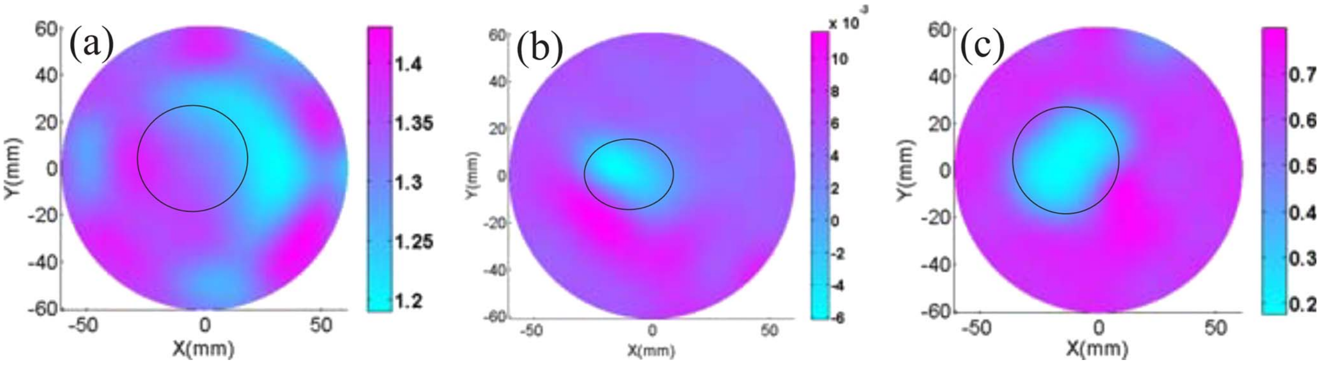


Fig. 3 A benign lesion: (a) Refractive index image. (b) and (c) Absorption and scattering coefficient images. The axes (left and bottom) illustrate the spatial scale (mm), whereas the color scale (right) records the refractive index (dimensionless), absorption, and scattering coefficient, respectively (mm^{-1}).

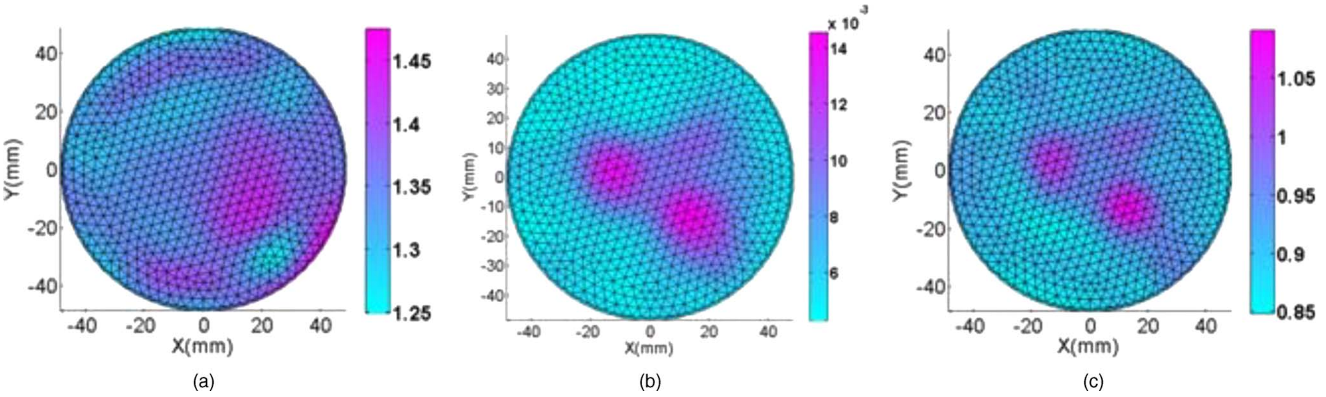


Fig. 4 The finite element mesh fused with the images shown in Fig. 2. (a) Refractive index image. (b) and (c) Absorption and scattering coefficient images. The axes (left and bottom) illustrate the spatial scale (mm), whereas the color scale (right) records the refractive index (dimensionless), absorption, and scattering coefficient, respectively (mm^{-1}).

tive index, absorption, and scattering images. Unlike the manual process, which uses the PDE toolbox in MATLAB to generate color images of the optical parameters, a program is developed using the C programming language to automatically extract interested features from the recovered image data to avoid possible errors introduced by the visualization process. This feature extraction process consists of two phases: image segmentation and parameter extraction.

3.1 *Image Segmentation*

It is important to identify the lesion area before extracting features for classification. The lesion areas will be identified by analyzing only the distribution of absorption and scattering coefficients because the refractive index data are relatively noisy. Due to their cellular morphology and biochemical compositions, some lesions may show higher absorption and scattering coefficients than normal tissues (e.g., Fig. 1), while other lesions may yield lower absorption and scattering coefficients relative to the surroundings (e.g., Fig. 3). This is another reason that it is unreliable to use only the absorption and scattering coefficient values to distinguish cancers from the benign lesions.

To automatically identify the lesion areas, the image segmentation process must first determine whether the areas with high coefficient values or low coefficient values should be selected. Because our automated classification procedure is designed for early noninvasive detection of breast cancer, it is reasonable to assume that a lesion usually exists in a small area of the entire imaging domain. Therefore, applying statistical analysis on the absorption and scattering coefficients can determine the possible background data range and identify the potential lesion areas. At first, the median absorption and scattering coefficient values of the 700 data samples are calculated. Then, the mean values of the upper quartile (upper 25%) sample data and the lower quartile (lower 25%) sample data are computed. If the difference between the mean of the upper quartile sample data and the median is greater than the difference between the median and the mean of the lower quartile sample data, lesion areas should have high coefficient values; otherwise, the lesion area should have low coefficient values. Although the refractive index images are relatively noisy, the same statistical analysis can be used to identify the background to apply the same segmentation process.

After the data range for the background is determined, possible lesion areas are identified through image segmentation. Image segmentation is a process in which regions or features sharing similar characteristics are identified and grouped together. Image segmentation may use statistical classification,[25] thresholding,[26] edge detection,[27] region detection,[28] or any combination of these techniques. The output of the segmentation is usually a set of classified elements, such as tissue regions or tissue boundaries. Most segmentation techniques are either region-based or edge-based. Region-based techniques rely on common patterns of values within a cluster of neighboring pixels or sample points. This cluster is referred to as the region, and the goal of the segmentation algorithm is to group regions according to their anatomical or functional roles. Edge-based techniques rely on discontinuities in image values between distinct regions, and the goal of the segmentation algorithm is to accurately demarcate the boundary separating these regions. A good segmentation procedure is the key to the success of the image processing, while weak or erratic segmentation algorithms almost always guarantee eventual failure.

Because our image data are 700 discrete sample points distributed on a triangular mesh, as shown in Fig. 4, traditional edge-based approaches are not suitable for these data. Therefore, a region-based thresholding segmentation method is used to identify the regions of interest (possible lesion areas). If the high coefficient areas are the targeted areas, the segmentation process can start at any point with a data value above a certain threshold, and the region is expanded by including the points directly connecting to any point in the region with a value above the threshold. This process continues until all points with values above the threshold are examined and included in a region. Conversely, if the low coefficient areas are the targeted areas, the segmentation process can start at any point with a data value below a certain threshold, and the region is expanded by including the points directly connecting to any point in the region with a value below the threshold. Because normal tissue absorption, scattering, and refractive index values vary for different patients, it is unde-

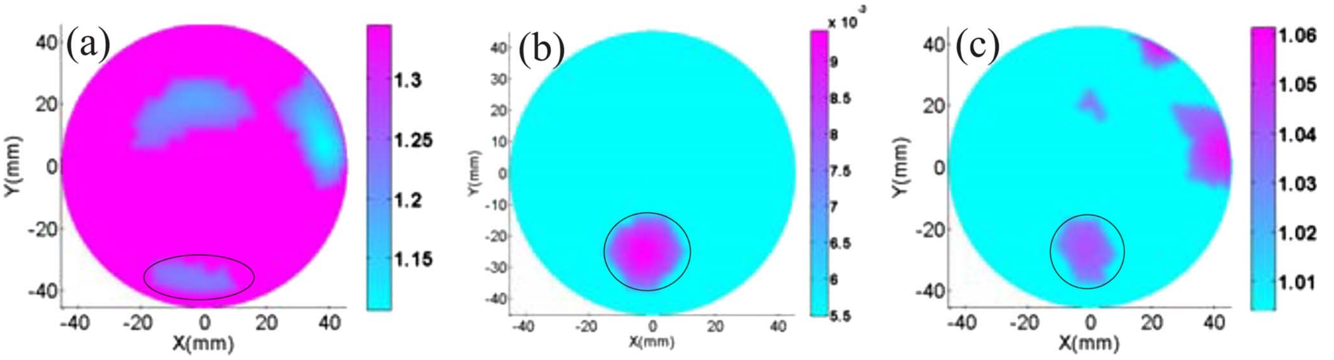


Fig. 5 Segmentation results of the images shown in Fig. 2: (a) Refractive index image. (b) and (c) Absorption and scattering coefficient images. The axes (left and bottom) illustrate the spatial scale (mm), whereas the color scale (right) records the refractive index (dimensionless), absorption, or scattering coefficient (mm^{-1}).

sirable to use an absolute threshold for the image segmentation. Instead, a relative threshold $th(0<th<1)$ is used in our image segmentation procedure. Assuming the maximum and minimum values of sample points to be $v_{\max}$ and $v_{\min}$, respectively, for any sample point with a value v, the sample point belongs to a high-value region of interest if $v-v_{\min}>th\cdot(v_{\max}-v_{\min})$, and conversely, it belongs to a low-value region of interest if $v-v_{\min}>(1-th)\cdot(v_{\max}-v_{\min})$. The threshold values for absorption and scattering coefficient images are set to 0.7, while the threshold for refractive index images is 0.6 based on experiments. The reason that a smaller threshold value is used for the refractive index images is because the variations of the refractive index values at different sample points are much smaller than those of absorption and scattering sample data. Figures 5 and 6 demonstrate the images after the region-based thresholding segmentation was applied on the images presented in Figs. 1 and 2 respectively. We note that the edges of the areas of interest on the images shown in Figs. 5 and 6 are not as smooth as those given in Figs. 1 and 2. This is due to the fact that the segmentation algorithm is directly applied on the 700 discrete sample points. In addition, the background values of these images are the mean value of the sample points in the region of non-interest.

3.2 *Feature Extraction*

After the segmentation, the classification attributes will be extracted from the regions of interest. If only the absorption or scattering images are used to classify the lesions, the region with the largest size or having the largest mean value is selected as the lesion area for each image. Once the lesion area is identified, the mean coefficient, size, length, and width of this area and the mean coefficient of the background are extracted as the attributes for image classification. However, the method of determining the lesion area is different when both the absorption and scattering coefficients are considered. Based on our experiments, a location correlation exists between the absorption and scattering coefficients. Therefore, when both absorption and scattering images are available, a region of interest on the absorption image will be selected as the lesion area only if it has the largest overlap area with any of the regions of interest in the scattering image, or its distance to any of the regions of interest in the scattering image is minimal if there are not any overlapped regions of interest between the absorption and scattering images. Hence, a lesion area identified using correlation between the absorption and scattering coefficients may be different from that identified by using absorption or scattering image alone. Again, the mean

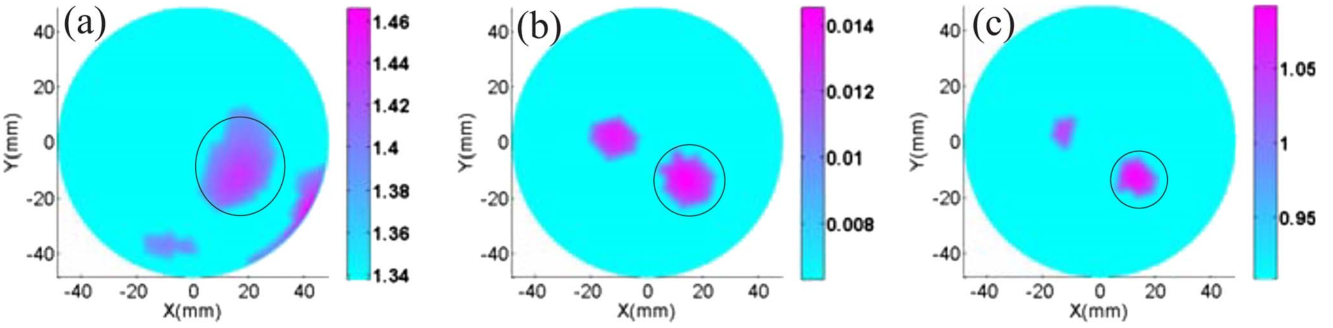


Fig. 6 Segmentation results of the images shown in Fig. 3: (a) Refractive index image. (b) and (c) Absorption and scattering coefficient images. The axes (left and bottom) illustrate the spatial scale (mm), whereas the color scale (right) records the refractive index (dimensionless), absorption, and scattering coefficient, respectively (mm^{-1}).

coefficient of the lesion areas and their size, length, and width and the mean coefficient of the background are extracted for image classification. In addition, the overlap ratio of the regions of interest on the absorption and scattering images is also included as a classification attribute. Because using the absorption and scattering images is sufficient to identify the lesion areas and the refractive index images are relatively noisy, the refractive index images are not used to identify the lesion areas. However, once the lesion area is identified using absorption and scattering images, the mean refractive index value at the lesion area and the refractive index value in the surrounding area are included in the classification attributes for cancer detection.

4 Image Classification

4.1 *Classification Method*

With the extracted diagnostic attributes, a support vector machine (SVM)[29,30] classifier is used to detect the breast cancer. SVMs are a new generation of machine-learning systems based on recent advances in statistical learning theory. SVMs deliver the state-of-the-art performance in real-world applications such as image classification, bio-sequence analysis, etc., and are now considered one of the standard tools for machine learning and data mining.

Given a training data set $D=\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, where $\mathbf{x}_n$ is a data sample and y_n is the associated class label, our breast cancer detection is actually a binary classification problem, i.e., y_n is from a label space $\{\pm 1\}$ where $+1$ denotes the cancer and -1 the *noncancer*. Let $\phi(\mathbf{x})$ be a mapping function that projects data samples from the data space to a feature space. The SVM learning algorithm finds a hyperplane $(\mathbf{w}, b)$ in the feature space to solve the following optimization problem:

$$\min_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N} \varepsilon_n,$$

$$\text{subject to } y_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b) \geqslant 1-\varepsilon_n, \quad n=1,\ldots,N,$$

$$\varepsilon_{\mathrm{n}} \geqslant 0, \quad n=1,\ldots,N, \qquad (1)$$

where $C>0$ is the penalty parameter of the error term. This optimization problem can be solved in the dual domain using quadratic programming:

$$\min_{\alpha} \frac{1}{2}\sum_{n,m}^{N} \alpha_n\alpha_m y_n y_m K(x_n,x_m) - \sum_{n=1}^{N} \alpha_n,$$

$$\text{s.t.} \sum_{n=1}^{N} \alpha_n y_n = 0, \quad 0 \leqslant \alpha_{\mathrm{n}} \leqslant C, \quad n=1,\ldots,N, \qquad (2)$$

where $K(\mathbf{x}_n,\mathbf{x}_m)=\phi^T(\mathbf{x}_n)\phi(\mathbf{x}_m)$ is the kernel. By solving Eq. (3), the decision function, given an unseen test sample $\mathbf{x}$, is expressed as:

$$f(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n y_n K(\mathbf{x}_n,\mathbf{x}) - b. \qquad (3)$$

The following four kernel functions are often used by SVM classification:

- Linear: $K(\mathbf{x}_n,\mathbf{x}_m)=\mathbf{x}_n{}^T\mathbf{x}_m$
- Polynomial: $K(\mathbf{x}_n,\mathbf{x}_m)=(\gamma\mathbf{x}_n{}^T\mathbf{x}_m+r)^d$, $\gamma>0$.
- Radial basis function (RBF):

$K(\mathbf{x}_n,\mathbf{x}_m)=\exp(-\|\mathbf{x}_n-\mathbf{x}_m\|^2/\sigma^2)$.

- Sigmoid: $K(\mathbf{x}_n,\mathbf{x}_m)=\tanh(\gamma\mathbf{x}_n{}^T\mathbf{x}_m+r)$.

Here, γ, r, and d are kernel parameters.

In our lesion image classification, the RBF kernel is used due to some of its advantages. First, the RBF kernel nonlinearly maps samples into a higher dimensional space so that it can handle the cases when the relation between class labels and attributes is nonlinear. Conversely, the linear kernel cannot deal with the nonlinear relationship between class labels and attributes. In fact, the linear kernel can be viewed as a special case of RBF since one can always achieve the same performance using the RBF kernel with some parameters (C,γ) as that using the linear kernel with a penalty parameter $\tilde{C}$.[31] Second, although the sigmoid kernel behaves like RBF for certain parameters, this kernel may be invalid (i.e., not the inner product of two vectors) under certain parameters.[32] Last, the polynomial kernel has more hyperparameters than the RBF kernel and may be more complex in model selection.

Using an RBF kernel, two parameters, C and γ, must be determined through the training data because it is impossible to know beforehand which C and γ are the best for a particular problem. In our automated classification procedure, a computer program is implemented using C programming language and the application programming interface (API) provided by Weka data mining tools[33] to automatically search for the best parameters. Because a high training accuracy (i.e., classifiers accurately predict training data whose class labels are indeed known) may not necessarily result in a high accuracy in prediction of unknown data due to the overfitting problem with many advanced classification algorithms, a tenfold stratified cross-validation is used to evaluate the accuracy of the SVM classifier.

4.2 *Classification Results Based Solely on Absorption Coefficient*

Our first experiment is to evaluate the SVM classifier trained by the attributes extracted from only the absorption coefficient images. Five attributes are extracted from each absorption coefficient image. They are the size of the lesion area in terms of the number of sample points, the mean coefficient of the lesion area, the mean coefficient of the background, and the length and width of the lesion area. Figure 7 shows the absorption attributes obtained by our feature extraction procedure. The confusion matrix of the 10-fold cross-validation results is depicted in Table 2. A confusion matrix is a visualization tool typically used in supervised machine learning. Each column of the confusion matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see whether the system is confusing two classes (i.e., commonly mislabeling one as another). As

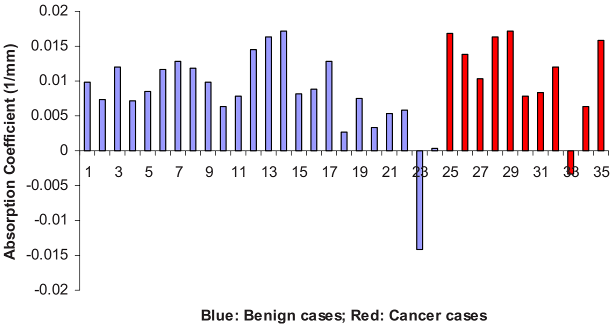


Fig. 7 Absorption attributes obtained by our automated feature extraction procedure. Note that three data sets (extremely small or negative absorption coefficient) were obtained from breasts with a diameter greater than 12 cm. For such large breasts, the coefficient values obtained are not reliable due to the very low signal-to-noise ratio. However, we keep these noisy data here to test whether our automated classification procedure can overcome the inability of our DOT system and correctly classify these samples into the proper classes. (Color online only.)

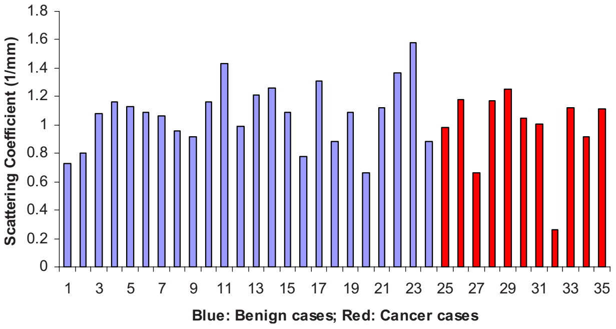


Fig. 8 Scattering attributes obtained by our automated feature extraction procedure. Note that the same explanation as described in the caption of Fig. 7 is applied to the same three data sets. (Color online only.)

shown in Table 2, the first row represents the number of cancer instances, and the second row represents the number of noncancer instances. On the other hand, the first column of Table 2 represents the number of instances that are classified as cancer by our SVM classifier, while the second column represents the number of instances classified as noncancer by our SVM classifier. The data in the first row show that there are 11 actual cancer instances, and 6 of them are identified as cancer by the SVM classifier. Therefore, the sensitivity of the classification is 54.5% (6/11). On the other hand, the data in the second row show that there are 24 actual noncancer instances, and 17 of them are identified as noncancer by the SVM classifier. Thus, the specificity is 70.8% (17/24). Although these results show a specificity of 70.8% on the classification using only absorption coefficient images, the low sensitivity (54.5%) indicates that using the absorption coefficient images alone cannot distinguish the malignant from the benign cases.

4.3 *Classification Results Based Solely on Scattering Coefficient*

Our second experiment is to evaluate the SVM classifier trained by the attributes extracted from the scattering coefficient images. The same five attributes as in the first experiment are extracted from each scattering coefficient image. Figure 8 shows scattering attributes obtained by our feature extraction procedure. The confusion matrix of the 10-fold cross-validation results is depicted in Table 3. Again, the results shown in Table 3 indicate that using the scattering coefficient images alone cannot distinguish the malignant from the benign cases since the sensitivity is only 45.5%.

4.4 *Classification Results Based on Both Absorption and Scattering Coefficients*

Our third experiment is to evaluate the SVM classifier trained by the attributes extracted from both the absorption and the scattering images. As discussed earlier, the lesion areas are identified by considering the co-existence of areas of interest in the same location on both the absorption and the scattering coefficient images. Therefore, some data shown in Fig. 9 are different from those shown in Figs. 7 and 8, which were obtained by analyzing only the absorption and the scattering images, respectively. In addition to the 10 attributes extracted from the absorption and the scattering images, respectively (5 attributes for each image), the overlap ratio of the regions of interest on the absorption and scattering images is also included as a classification attribute. The overlap ratio is calculated as twice the number of overlapped points divided by the total number of points in the corresponding regions of interest on the absorption and scattering images. The confusion matrix of the 10-fold cross-validation results is depicted in Table 4.

The results shown in Fig. 9 indicate that combining the attributes extracted from the absorption images with those obtained from the scattering images improves the classification performance. With the combined attributes, the sensitivity, specificity, and overall accuracy of our classification are 63.6%, 83.3%, and 77.1%, respectively.

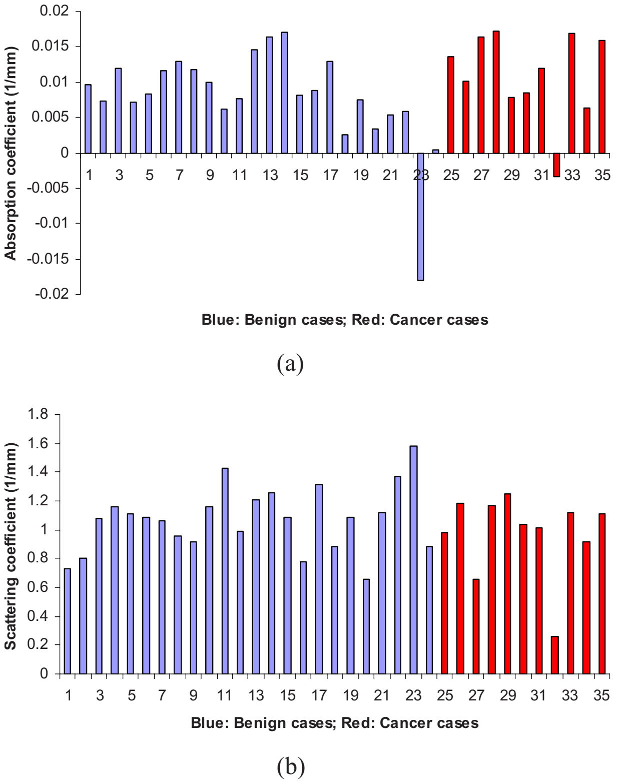
As discussed in Ref. 13, it is impossible to distinguish the malignant from the benign cases by just visually examining both absorption and scattering images. However, our automated classification procedure can achieve reasonable classification results using both absorption and scattering coeffi-

Table 2 Confusion matrix of the SVM classifier using attributes extracted from absorption coefficient images.

	Cancer	Noncancer
Cancer	6	5
Noncancer	7	17

Table 3 Confusion matrix of the SVM classifier using attributes extracted from scattering coefficient images.

	Cancer	Noncancer
Cancer	5	6
Noncancer	8	16



(a)

(b)

Fig. 9 (a) Absorption and (b) scattering attributes obtained by our automated feature extraction procedure considering the co-existence of the interested regions in the same location on both the absorption and scattering coefficient images. Note that the same explanation as described in the caption of Fig. 7 is applied to the same three data sets. (Color online only.)

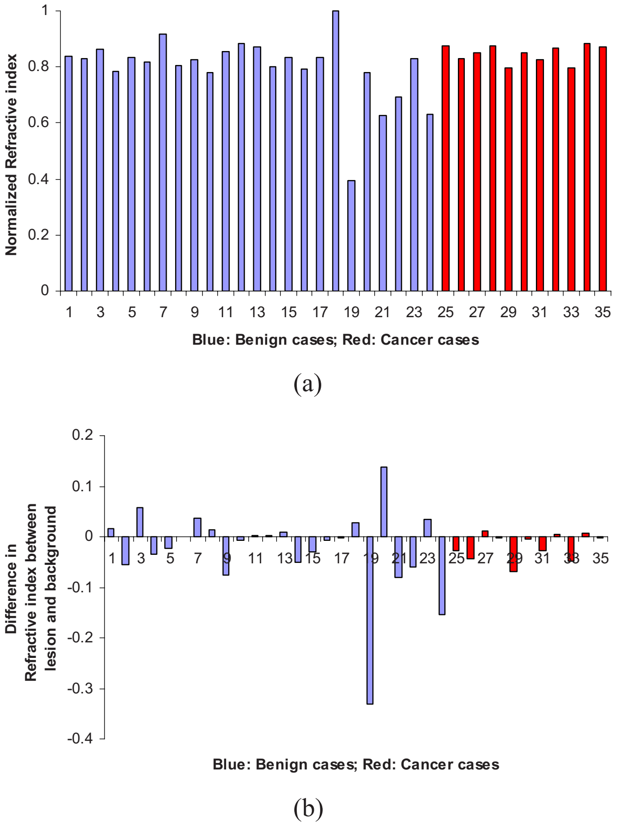
cient images. Especially, the specificity of the results obtained by our automated classification procedure on two parameters (absorption and scattering images) is much higher than the specificity of the visual examination results using three parameters (absorption, scattering, and refractive index images). However, the sensitivity of the automated classification using only absorption and scattering attributes is still low, suggesting that it is necessary to use the refractive index attributes for classification.

4.5 *Classification Results Based on Absorption and Scattering Coefficients and Refractive Index*

Our final experiment is to evaluate the SVM classifier trained by the attributes extracted from the refractive index images combined with the attributes from the corresponding absorption and scattering images. As discussed earlier, the location correlation between the regions of interest on the absorption and scattering images is used to determine the lesion area. In addition to all attributes used in the third experiment, the mean refractive index of the lesion area and the mean refractive index of the area surrounding the lesion area are added into the attribute list. These attributes are listed in Fig. 10. Training the SVM classifier using the attributes obtained by all three kinds of images, the confusion matrix of the 10-fold cross-validation results is presented in Table 5.

Table 4 Confusion matrix of the SVM classifier using attributes extracted from both absorption and scattering coefficient images.

	Cancer	Noncancer
Cancer	7	4
Noncancer	4	20



(a)

(b)

Fig. 10 (a) Refractive index attributes obtained from the lesion areas identified by the associated absorption and scattering coefficient images. (b) Difference in refractive index between lesion and background tissue. Note that the same explanation as described in the caption of Fig. 7 is applied to the same three data sets. (Color online only.)

The results in Table 5 show that the sensitivity, specificity, and overall accuracy of the automated classification procedure are 81.8%, 91.7%, and 88.6%, respectively. Comparing to the classification results using only the attributes extracted from absorption and scattering images, classification using refractive index attributes in conjunction with the absorption and

Table 5 Confusion matrix of the SVM classifier using attributes extracted from absorption, scattering, and refractive index images.

	Cancer	Noncancer
Cancer	9	2
Noncancer	2	22

scattering attributes improves the sensitivity and specificity by 19 and 8 percentage points, respectively. These results are also better than the visual examination results listed in Table 1. In particular, the automated classification procedure improves the specificity of the classification by more than 20 percentage points, comparing to the visual examination method presented in Ref. 13.

5 Conclusions

An automated procedure for detecting breast cancer based on optical tomographic images is developed. This procedure uses a computer program to automatically extract attributes from absorption, scattering, and refractive index images for lesion classification. An SVM classifier is used to distinguish between the malignant and benign lesions based on these automatically extracted attributes. The classification results show that the sensitivity, specificity, and overall accuracy using this automated procedure are 81.8%, 91.7%, and 88.6%, respectively. In contrast, the sensitivity, specificity, and overall accuracy of the classification using attributes extracted from only the absorption and scattering coefficient images are 63.6%, 83.3%, and 77.1%, respectively. These results indicate that combining the refractive index with the absorption and scattering coefficients can achieve significantly improved classification performance over using only absorption and scattering coefficients. Furthermore, these results are also better than the results obtained by visual examination of images, in which the sensitivity, specificity, and overall accuracy are 81.8%, 70.8%, and 74.3% respectively.

It is worth mentioning that it is critical to obtain reliable sample data from the breast masses for accurate image processing and classification. Our experiments show that the automated classification procedure cannot consistently classify the samples obtained from the three large breasts (>12 cm in diameter) due to the low signal-to-noise ratio (SNR) of the hardware system for these cases. If these data samples were removed, the automated classification results should have higher sensitivity, specificity, and overall accuracy.

To achieve better classification results, we are currently developing data collection strategies that can improve the SNR of our imaging system so that it can produce reliable coefficient data for large breasts. In addition, our automated procedure used a predetermined threshold for image segmentation. We are currently investigating the entropy-based and iterative selection methods to automatically determine an optimal segmentation threshold for a particular image.

Last, there was possibly cross talk between the refractive index and the absorption/scattering parameters and so the recovered refractive index was just an estimation. However, the cross talk was reduced to a certain extent via a two-step strategy where the refractive index and absorption/scattering parameters were reconstructed using two different algorithms.[16] Importantly, the estimation or semiquantitative nature of the recovered refractive index, although limited in accuracy, is sufficient for us to classify the cancer and benign groups effectively using the automated classification algorithms described in this paper. A more accurate estimate of the refractive index will likely improve the accuracy for cancer classification—we are currently developing schemes that can enhance the separation of refractive index from absorption/scattering parameters. In fact, we have recently reported a region-based reconstruction approach that has shown promising results in this regard using phantom studies.[34] We plan to evaluate this and upcoming new methods for better refractive index reconstruction on clinical data in the near future.

Acknowledgments

This work was supported in part by a grant from the National Institutes of Health (NIH), Grant No. R01 CA090533.

References

1. A. E. Cerussi, A. J. Berger, F. Bevilacqua, N. Shah, D. Jakubowski, J. Butler, R. F. Holcombe, and B. J. Tromberg, "Sources of absorption and scattering contrast for near-infrared optical mammography," *Acad. Radiol.* **8**(3), 211–218 (2001).
2. A. Cerussi, D. Hsiang, N. Shah, R. Mehta, A. Durkin, J. Butler, and B. J. Tromberg, "Predicting response to breast cancer neoadjuvant chemotherapy using diffuse optical spectroscopy," *Proc. Natl. Acad. Sci. U.S.A.* **104**(10), 4014–4019 (2007).
3. T. Durduran, R. Choe, J. P. Culver L. Zubkov, M. J. Holboke, J. Giammarco, B. Chance, and A. G. Yodh, "Bulk optical properties of healthy female breast tissue," *Phys. Med. Biol.* **47**(16), 2847–2861 (2002).
4. V. Ntziachristos, A. G. Yodh, M. D. Schnall, and B. Chance, "MRI-guided diffuse optical spectroscopy of malignant and benign breast lesions," *Neoplasia* **4**(4), 347–354 (2002).
5. Q. Zhu, E. B. Cronin, A. A. Currier, H. S. Vine, M. Huang, N. Chen, and C. Xu, "Benign versus malignant breast masses: optical differentiation with US-guided optical imaging reconstruction," *Radiology* **237**(1), 57–66 (2005).
6. B. Brooksby, S. Jiang, H. Dehghani, B. W. Pogue, and K. D. Paulsen, "Combining near-infrared tomography and magnetic resonance imaging to study *in vivo* breast tissues: implementation of a Laplacian-type regularization to incorporate magnetic resonance structure," *J. Biomed. Opt.* **10**, 011504 (2005).
7. L. Spinelli, A. Torricelli, A. Pifferi P. Taroni, G. M. Danesini, and R. Cubeddu, "Bulk optical properties and tissue components in the female breast from multiwavelength time-resolved optical mammography," *J. Biomed. Opt.* **9**(6), 1137–1142 (2004).
8. M. A. Franceschini et al., "Frequency-domain techniques enhance optical mammography: initial clinical results," *Proc. Natl. Acad. Sci. U.S.A.* **94**(12), 6468–6473 (1997).
9. D. Grosenick, K. T. Moesta, H. Wabnitz, J. Mucke, C. Stroszczynski, R. Macdonald, P. M. Schlag, and H. Rinneberg, "Time-domain optical mammography: initial clinical results on detection and characterization of breast tumors," *Appl. Opt.* **42**(16), 3170–3186 (2003).
10. H. Jiang, Y. Xu, N. Iftimia, J. Eggert, K. Klove, L. Baron, and L. Fajardo, "Three-dimensional optical tomographic imaging of breast in a human subject," *IEEE Trans. Med. Imaging* **20**, 1334–1340 (2001).
11. H. Jiang, N. Iftimia, Y. Xu, J. Eggert, L. Fajardo, and K. Klove, "Near-infrared optical imaging of the breast with model-based reconstruction," *Acad. Radiol.* **9**, 186–194 (2002).
12. X. Gu, Q. Zhang, M. Bartlett, L. Schutz, L. L. Fajardo, and H. Jiang, "Differentiation of cysts from solid tumors in the breast with diffuse optical tomography," *Acad. Radiol.* **11**(1), 53–60 (2004).
13. X. Liang, Q. Zhang, C. Li, S. R. Grobmyer, L. L. Fajardo, and H. Jiang, "Breast cancer detection using phase contrast diffuse optical tomography," *Proc. SPIE* **6434**, 643425 (2007).
14. C. Li, S. R. Grobmyer, N. Massol, X. Liang, Q. Zhang, L. Chen, L. Fajardo, and H. Jiang, "Noninvasive *in vivo* tomographic optical imaging of cellular morphology in the breast: Possible convergence of microscopic pathology and macroscopic radiology," *Med. Phys.* **35**(6), 2493–2501 (2007).
15. B. Chance, S. Nioka, J. Zhang, E. Conant, E. Hwang, S. Briest, S. Orel, M. Schnall, and B. Czerniecki, "Breast cancer detection based on incremental biochemical and physiological properties of breast cancers: a six-year, two-site study," *Acad. Radiol.* **12**(8), 925–933 (2005).
16. H. Jiang and Y. Xu, "Phase-contrast imaging of tissue using near-infrared diffusing light," *Med. Phys.* **30**, 1048–1051 (2003).
17. K. Doi, "Current status and future potential of computer-aided diag-

nosis in medical imaging," *Br. J. Radiol.* **78**, S3–S19 (2005).

18. M. K. Markey, J. Y. Lo, and C. E. Floyd Jr., "Differences between computer-aided diagnosis of Breast Masses and That of Calcifications," *Radiology* **223**, 489–493 (2002).
19. M. Chen, Y.-H. Chou, K.-C. Han, G.-S. Hung, C.-M. Tiu, H.-J. Chiou, and S.-Y. Chiou, "Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks," *Radiology* **226**(2), 504–514 (2003).
20. C. E. Floyd Jr., J. Y. Lo, and G. D. Tourassi, "Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions," *Am. J. Roentgenol.* **175**(5). 1347–1352 (2000).
21. N. Iftimia, X. Gu, Y. Xu, and H. Jiang, "A compact, parallel-detection diffuse optical mammography system," *Rev. Sci. Instrum.* **74**, 2836–2842 (2003).
22. N. Iftimia and H. Jiang, "Quantitative optical image reconstruction of turbid media using dc measurements," *Appl. Opt.* **39**, 5256–5261 (2000).
23. Y. Xu, X. Gu, T. Khan, and H. Jiang, "Absorption and scattering images of heterogeneous scatteringmedia can be simultaneously reconstructed by use of dc data," *Appl. Opt.* **41**, 5427–5437 (2002).
24. H. Jiang, N. Iftimia, Y. Xu, J. Eggert, L. Fajardo, and K. Klove, "Near-infrared optical imaging of the breast with model-based reconstruction," *Acad. Radiol.* **9**, 186–194 (2002).
25. N. Vandenbroucke, L. Macaire, and J. Postaire, "Color image segmentation by pixel classification in an adapted hybrid color space: application to soccer image analysis," *Comput. Vis. Image Underst.* **90**(2), 190–216 (2003).
26. Y. Zhang H. Qu, and Y. Wang, "Adaptive image segmentation based on fast thresholding and image merging," in *Proc. 16th International Conference on Artificial Reality and Telexistence—Workshops (ICAT06)*, 308–311 (2006).
27. D. Ziou and S. Tabbone, "Edge detection techniques an overview," *Appl. Opt.* **8**(4), 537–559 (1998).
28. X. Song, B. W. Pogue, S. Jiang, M. M. Doyley, H. Dehghani, T. D. Tosteson, and K. D. Paulsen, "Automated region detection based on the contrast-to-noise ratio in near-infrared tomography," *Appl. Opt.* **43**(5), 1053–1062 (2004).
29. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*, Cambridge University Press, Cambridge, UK (2000).
30. C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learn.* **20**(3), 273–297 (1995).
31. S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Comput.* **15**(7), 1667–1689 (2003).
32. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York (1995).
33. I. H. Witten, and E. Frank, "Data Mining: Practical machine learning Tools and techniques," 2nd ed., Morgan Kaufmann, San Francisco, (2005).
34. X. Liang, Q. Zhang, and H. Jiang, "Quantitative reconstruction of refractive index distribution and imaging of glucose concentration by using diffusing light," *Appl. Opt.* **45**(32), 8360–8365 (2006).