Biomedical Optics EXPRESS

Direct identification of breast cancer pathologies using blind separation of label-free localized reflectance measurements

Alma Eguizabal,[1] Ashley M. Laughney,[2] Pilar Beatriz García-Allende,[3] Venkataramanan Krishnaswamy,[2] Wendy A. Wells,[4] Keith D. Paulsen,[2] Brian W. Pogue,[2] Jose M. Lopez-Higuera,[1] and Olga M. Conde[1,*]

[1]Photonics Engineering Group, Dep. TEISA, University of Cantabria, Plaza de la Ciencia sn, 39005 Santander, Spain
[2]Thayer School of Engineering, 8000 Cummings Hall, Dartmouth College, Hanover, New Hampshire 03755, USA
[3]Helmholtz Zentrum München GmbH, Germany
[4]Dartmouth Hitchcock Medical Ctr., USA
**olga.conde@unican.es*

Abstract: Breast tumors are blindly identified using Principal (PCA) and Independent Component Analysis (ICA) of localized reflectance measurements. No assumption of a particular theoretical model for the reflectance needs to be made, while the resulting features are proven to have discriminative power of breast pathologies. Normal, benign and malignant breast tissue types in lumpectomy specimens were imaged *ex vivo* and a surgeon-guided calibration of the system is proposed to overcome the limitations of the blind analysis. A simple, fast and linear classifier has been proposed where no training information is required for the diagnosis. A set of 29 breast tissue specimens have been diagnosed with a sensitivity of 96% and specificity of 95% when discriminating benign from malignant pathologies. The proposed hybrid combination PCA-ICA enhanced diagnostic discrimination, providing tumor probability maps, and intermediate PCA parameters reflected tissue optical properties.

OCIS codes: (170.6935) Tissue characterization; (170.6510) Spectroscopy, tissue diagnostics.

References and links

1. World Health Organization (2008), http://www.who.int/en/
2. F. Fitzal, O. Riedl, and R. Jakesz, “Recent developments in breast-conserving surgery for breast cancer patients,” Langenbecks Arch. Surg. **394**(4), 591–609 (2009).
3. R. G. Pleijhuis, M. Graafland, J. de Vries, J. Bart, J. S. de Jong, and G. M. van Dam, “Obtaining adequate surgical margins in breast-conserving therapy for patients with early-stage breast cancer: current modalities and future directions,” Ann. Surg. Oncol. **16**(10), 2717–2730 (2009).
4. S. Srinivasan, B. W. Pogue, S. Jiang, H. Dehghani, C. Kogel, S. Soho, J. J. Gibson, T. D. Tosteson, S. P. Poplack, and K. D. Paulsen, “Interpreting hemoglobin and water concentration, oxygen saturation, and scattering measured in vivo by near-infrared breast tomography,” Proc. Natl. Acad. Sci. U.S.A. **100**(21), 12349–12354 (2003).
5. A. M. Laughney, V. Krishnaswamy, E. J. Rizzo, M. C. Schwab, R. J. Barth, B. W. Pogue, K. D. Paulsen, and W. A. Wells, “Scatter spectroscopic imaging distinguishes between breast pathologies in tissues relevant to surgical margin assessment,” Clin. Cancer Res. **18**(22), 6315–6325 (2012).
6. V. Krishnaswamy, P. J. Hoopes, K. S. Samkoe, J. A. O’Hara, T. Hasan, and B. W. Pogue, “Quantitative imaging of scattering changes associated with epithelial proliferation, necrosis, and fibrosis in tumors using microsampling reflectance spectroscopy,” J. Biomed. Opt. **14**(1), 014004 (2009).
7. S. C. Kanick, H. J. C. M. Sterenborg, and A. Amelink, “Empirical model of the photon path length for a single fiber reflectance spectroscopy device,” Opt. Express **17**(2), 860–871 (2009).
8. G. Zonios and A. Dimou, “Modeling diffuse reflectance from homogeneous semi-infinite turbid media for biological tissue applications: a Monte Carlo study,” Biomed. Opt. Express **2**(12), 3284–3294 (2011).
9. S. L. Jacques and S. Prahl, Oregon Medical Laser Center (2010).
10. J. Glatz, N. C. Deliolanis, A. Buehler, D. Razansky, and V. Ntziachristos, “Blind source unmixing in multi-spectral optoacoustic tomography,” Opt. Express **19**(4), 3175–3184 (2011).

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

11. I. Schelkanova and V. Toronov, “Independent component analysis of broadband near-infrared spectroscopy data acquired on adult human head,” Biomed. Opt. Express **3**(1), 64–74 (2012).
12. S. Kohno, I. Miyai, A. Seiyama, I. Oda, A. Ishikawa, S. Tsuneishi, T. Amita, and K. Shimizu, “Removal of the skin blood flow artifact in functional near-infrared spectroscopic imaging data through independent component analysis,” J. Biomed. Opt. **12**(6), 062111 (2007).
13. J. Virtanen, T. Noponen, and P. Meriläinen, “Comparison of principal and independent component analysis in removing extracerebral interference from near-infrared spectroscopy signals,” J. Biomed. Opt. **14**(5), 054032 (2009).
14. J. L. Semmlow, *Biosignal and Biomedical Image Processing: MATLAB-Based Applications* (CRC Press, 2004), Chap. 9.
15. R. Gallardo-Caballero, C. J. García-Orellana, H. M. González-Velasco, and M. Macías-Macías, “Independent component analysis applied to detection of early breast cancer signs,” in *Proceeding of 9th International Work-Conference on Artificial Neural Networks* (IWANN, San Sebastian, Spain, 2007), pp. 988–995.
16. I. Kopriva and A. Peršin, “Unsupervised decomposition of low-intensity low-dimensional multi-spectral fluorescent images for tumour demarcation,” Med. Image Anal. **13**(3), 507–518 (2009).
17. F. Abu-Amara and I. Abdel-Qader, “Detection of breast cancer using independent component analysis,” in *Proceedings of IEEE International Conference on Electro/Information Technology* (Institute of Electrical and Electronics Engineers, New York, 2007), pp.428–431.
18. A. M. Laughney, V. Krishnaswamy, P. B. Garcia-Allende, O. M. Conde, W. A. Wells, K. D. Paulsen, and B. W. Pogue, “Automated classification of breast pathology using local measures of broadband reflectance,” J. Biomed. Opt. **15**(6), 066019 (2010).
19. A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” Neural Netw. **13**(4-5), 411–430 (2000).
20. A. Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” IEEE Trans. Neural Netw. **10**(3), 626–634 (1999).
21. R. Bro, E. Acar, and T. G. Kolda, “Resolving the sign ambiguity in the singular value decomposition,” J. Chemometr. **22**(2), 135–140 (2008).

1. Introduction

Breast cancer continues to be the most diagnosed cancer among women, comprising 23% of all female cancers. Non-invasive small lesions, detected at an early-stage, however, can be treated successfully with breast conserving therapy (BCT), which includes local tumor excision followed by moderate-dose radiation therapy [1]. Early invasive breast cancers (stage I and stage II) have, however, a high risk of reoccurrence when residual disease is left at or near the cut edge. In fact, BCT has been demonstrated to be equally effective as mastectomy only when no residual disease is left on margins [2], thereby minimizing the need for more a more radical therapy like mastectomy. Despite its therapeutic predictive value, most studies report high variability in the number of patients treated with BCT with residual disease, demonstrating a lack of standardization for margin delineation [3].

Light scattering spectroscopy has been applied broadly to identify residual disease in resected breast tissues by detecting changes in the scattering spectrum induced by morphological variations in the size and number of density of cells and the tissue extra-cellular matrix [4]. Natural heterogeneity in light scattering from tissue morphology has been observed and its spatial distribution can be used to improve discrimination between tissue subtypes [5]. Consequently, the scanning spectroscopy system demonstrated in [6] has been designed to be maximally sensitive to elastic scattering, although some partial coupling with hemoglobin absorption has been observed. Signal localization is employed in the illumination and detection paths to preserve the weakly scattered spectrum. Optical properties, namely the reduced scattering and absorption coefficients, are traditionally parameterized according to theoretical models of light scattering in turbid media. These models are valid when specific, physical conditions are met in the data acquisition geometry [7]. The full accomplishment of these conditions is sometimes impossible to fulfill, yielding uncertainty in the separation of absorption and scattering signatures. Typically, analytical solution for the problem of diffuse reflectance from turbid media such as biological tissues only exists for idealized systems, like a point source in a semi-infinite medium [8]. Models assume light incidence on an optically homogenous medium, which is also only approximate for biological tissues. Furthermore, single-fiber reflectance measurements do not accurately recover the photon pathlength,

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

limiting absolute quantification of optical parameters [9]. Consequently, the existence of a model-free approach would be a great asset.

Blind Signal Separation (BSS) is a set of signal processing techniques able to decouple information arising from multiple sources. Consequently, they can be employed to decouple the information generated by absorption and scattering in tissue for unique acquisition geometries. These methods have been extensively used for removing interference and noise or for feature extraction from optical signals [10–13]. Principal Component Analysis (PCA) performs a change of basis and finds new uncorrelated projections, while Independent Component Analysis (ICA) creates a new independent feature space, with more statistical separation than uncorrelation [14]. The validity of ICA has been previously demonstrated in diverse scenarios. It has been proven to enhance classification features from mammograms for breast cancer detection [15] and also for other cancer types: maps of tumor probability have been extracted from the ICA of RGB fluorescence images taken from the skin [16]. PCA is typically employed to reduce data dimensionality and to enhance performance of the independent ICA algorithms [17]. PCA and ICA have been effectively applied to un-mix distinct exogenous fluorophores in multispectral opto-acoustic tomography data [10] and also for removing or studying blood absorbance from NIRS signals [11,12]. Nevertheless, the overall outperformance of ICA over PCA is still to be proven [13].

Here, a feature-extraction method is presented to discriminate benign from malignant pathologies in resected breast tissue and its diagnostic performance is validated according to histology, the diagnostic gold standard. No analytical models are performed to extracted diagnostic components from the scattering spectrum. Instead, PCA is used to transform spectral data into an uncorrelated feature space that reduces data dimensionality and eliminates cross talk between hemoglobin absorption and scattering signatures. Then, ICA optimized by PCA is used to predict the breast tissue subtype. Finally, sign ambiguities associated with the BSS algorithm are solved by a user-guided, soft calibration process.

2. Materials and methods

2.1 Optical imaging data from breast tissues and the modeling of reflectance

Localized measures of broadband reflectance from resected breast tissues were obtained from previous work [18], using a custom-built, quasi-confocal acquisition geometry [6]. This system separates weakly scattered from multiply scattered light by spatial confinement of the illumination and detection spot sizes (~100µm). The system employs a broadband fiber-coupled tungsten-halogen light source, operating in the (510 – 785 nm) spectral waveband. An optical-fiber, coupled to a CCD-based spectrometer, was used for confocal spectroscopic detection. The spectral resolution of the system provides 512 spectral images for each sample.

Samples of freshly resected breast tissues acquired during breast conserving surgery, were obtained directly from the Department of Pathology at Dartmouth-Hitchcock Medical Center, when there was tissue in excess of that required to make a clinical diagnosis. Tissues were 1-2cm^2 with a thickness of 3-5mm. Immediately after each imaging procedure, each sample with formalin-fixed and paraffin embedded, then stained with Hematoxylin and Eosin (H&E) for pathology correlation. 29 resected tissues were imaged and, on each specimen, several regions of interest (ROIs) were further evaluated by a pathologist for precise co-registration with optical maps. In total, 48 different ROIs were identified that were not uniform in size, having diameters from 500 um to 0.2 cm. Tissues were characterized as benign, malignant or adipose, as summarized in Table 1.

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

Table 1. Regions of Interest (ROI) Diagnosed by the Pathologist on the 29 Specimens

Tissue type	ROI number	Pixel number
Benign	25	36979
Malignant	14	23220
Adipose	9	7021
Total	48	67220

An analytical solution that accurate describes the diffuse reflectance arising from turbid media such as biological tissues has not yet been demonstrated. In spite of this, under the spatial constraints it is possible to model the measured backscattered reflectance with the aid of an empirical approximation. To compare reflectance modeling parameters and PCA-ICA analysis, an empirical approximation validated on previous study [18] was considered to contrast blindly obtained results. This model is shown in Eq. (1):

$$R(\lambda) = A\lambda^{-b} \quad \exp\left[-\rho C_{HbT}\left(f_{O_2}\varepsilon_{HbO_2}(\lambda) + \left(1 - f_{O_2}\right)\varepsilon_{Hb}(\lambda)\right)\right] \tag{1}$$

Here, A is the scattering amplitude, b is the scattering power, ρ the pathlength, C_{HbT} the concentration of hemoglobin, and f_{O_2} the fraction of oxygenated hemoglobin, $\varepsilon_{HbO_2}(\lambda)$ and $\varepsilon_{Hb}(\lambda)$ are the molar extinction coefficients of oxygenated and deoxygenated hemoglobin respectively, obtained from Oregon Medical Laser Center Database [19].

2.2 Multivariate linear analysis and BSS

Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are linear processing techniques characterized by their simplicity and low computational load. Equation (2) shows the linear transformation that describes both processes.

$$\mathbf{s}_{\mathbf{M\times N}} = \mathbf{W}_{\mathbf{M\times M}}\,\mathbf{x}_{\mathbf{M\times N}} \tag{2}$$

where **x** comprises the reflectance information per tissue sample measured at N locations (approximately 4000 pixels or observations per sample in this case) and M different spectral bands (being M = 512); **W** is the mixing matrix that represents the linear operation to be applied on the original data **x** to provide **s,** that contains the multivariate data result with the decoupled mixed signals or scores i.e., the representation of the raw data **x** in the new component space.

This also can be described in the opposite way, i.e. the measured data as a linear mix of the components as indicated in Eq. (3)

$$\mathbf{x}_{\mathbf{M\times N}} = \mathbf{W}^{-1}_{\mathbf{M\times M}}\,\mathbf{s}_{\mathbf{M\times N}} = \mathbf{A}_{\mathbf{M\times M}}\,\mathbf{s}_{\mathbf{M\times N}} \tag{3}$$

where **A** is the matrix of coefficients or loadings. PCA and ICA algorithms do not require any training, modeling, supervision or previous signal information and they are considered accordingly Blind Signal Separation techniques (BSS).

2.2.1 Linear mixture of components

PCA and ICA assume linear mixtures. If neperian logarithm is applied on empirical expression in Eq. (1) a linear sum of the reflectance spectra parameters can be defined and then compared with PCA-ICA results, as shown on Eq. (4):

$$\begin{aligned} X(\lambda) &= \ln\left(R(\lambda)\right) \\ &= \ln(A) - b\ \ln(\lambda)\ -\rho C_{HbT}[f_{O_2}\varepsilon_{HbO_2}(\lambda) - \left(1 - f_{O_2}\right)\varepsilon_{Hb}(\lambda)] \\ &= S_1\sigma_1(\lambda) + S_2\sigma_2(\lambda) + S_3\sigma_3(\lambda) \end{aligned} \tag{4}$$

#187362 - \$15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

where $X(\lambda)$ is the logarithm of reflectance $R(\lambda)$ where the contributions to the spectra are linearly mixed. Interpreting the general expression, S_1, S_2 and S_3 would be the linear weights that modulate the contribution of each spectral component, whereas each spectral behavior is represented by the $\sigma_1(\lambda), \sigma_2(\lambda), \sigma_3(\lambda)$ functions that could be directly associated with absorption and scattering features. The sum of all contributions would result into the initial reflectance spectrum $X(\lambda)$.

Extrapolating the analysis pixel by pixel, a mixing matrix solution of BSS can be associated with the feature variation across wavelength, $\sigma_n(\lambda)$, and according to Eqs. (3) and (4) the expression for the spectrum becomes as shown in Eq. (5):

$$X(\lambda, N) = \mathbf{x}_{M\times N} = \boldsymbol{\sigma}_{M\times M}\; s_{M\times N} = \mathbf{W}^{-1}_{M\times M}\; \mathbf{s}_{M\times N} = \mathbf{A}_{M\times M}\mathbf{s}_{M\times N} \qquad (5)$$

where the columns of matrix **A** would become directly the spectral features of the spectral components of tissue $\sigma_n(\lambda)$ and they would be related with the properties of its components; the sources **s** are the blindly extracted parameters, which might be related to the contribution of tissue to the scattering and absorption phenomena.

2.2.2 PCA to uncorrelate components and compress spectral data

Principal Component Analysis (PCA) is usually employed as a technique to reduce the number of variables in a data set with a minimal loss of information and to search for a more significant data representation. However, the physical meaning of these new variables is not always straightforward.

PCA assumes a linear approximation of the problem, as the one described in 2.4.1. The covariance matrix **C** from input data **x** must be calculated, assuming that **x** is a mean-centered version of the initial reflectance data. Since the covariance matrix is symmetric, calculation can be described as in Eq. (6):

$$\mathbf{C} = \mathbf{x}\mathbf{x}^T = \mathbf{E}\mathbf{D}\mathbf{E}^T \qquad (6)$$

where **D** is a diagonal matrix containing the eigenvalues of **C**, **E** are the eigenvectors of the covariance matrix **C**. The mixing matrix in Eq. (5), **W** is defined as its Hermitian $\mathbf{W} = \mathbf{E}^{\mathrm{H}}$. This matrix **W** is the one that transforms the input data **x** into the uncorrelated components in vector **s**, being the components ordered according to the contribution of their eigenvalues to the total variance of the data set. Focusing on Eq. (4), components in vector **s** could represent the contributions to variation of spectra S_1, and $\sigma_n(\lambda)$ would be the normalized spectral variation of tissue components. The first few columns of matrix **W** could extract those tissue properties, being the rest components with small associated eigenvalues related to noise.

A criterion must then be established to decide these few number of maintained uncorrelated components from the initial $M = 512$ to L. The chosen criterion is to maintain $L<M$ eigenvalues, with a joint variance above a specific threshold, as shown in Eq. (7).

$$\text{Kept variance} = \frac{\sum_{q=1}^{L} D(q,q)}{\sum_{q=1}^{M} D(q,q)} 100(\%) \geq \text{threshold} \qquad (7)$$

where $D(q,q)$ is the q^{th} eigenvalue of the covariance matrix **C**.

2.2.3 ICA to identify independent latent factors

Independent Component Analysis (ICA) is also a multivariate linear blind separator that uses higher order statistics, instead of covariance, to extract the new set of linearly unmixed

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

components. Since statistical independence is a stronger condition than uncorrelation, more accurate maps of diagnosis can be obtained. This work assumes that exist malignancy tissue properties that are statistically independent from other tissue types such as normal or adipose. This hypothesis is based on the differences in absorbance and scattering generated by each tissue condition.

All measures of spectral reflectance could be used to discriminate between tissue types, but this is frequently not optimal and always computationally demanding [20]. Here, PCA is proposed to reduce the data dimensionality [21]. Consequently, the data arising from this pre-processing step can be analyzed with ICA. This is the reason why ICA results cannot be spectrally interpreted as in Eq. (4): the dimension of data now is not the spectral 512 components but the very few PCA pre-processed components. In fact, a similar situation as in Eq. (3) is faced, but now $\mathbf{x}$, i.e. the detectors, are the uncorrelated components maintained after the PCA analysis, and $\mathbf{s}$, i.e. the sources, will be the IC components, which are supposed to be more diagnostically discriminating.

Figure 1 summarizes the whole analysis procedure to obtain the maps of tumor probability: PCA is first applied to the logarithm of the initial reflectance data set, containing 512 images, one per wavelength Fig. 1(a). Then, a few uncorrelated components are maintained Fig. 1(b) and are input into the ICA algorithm. A tumor map probability Fig. 1(c) is computed from the resulting independent components, which, because of the more stronger condition mentioned above, are expected to be more diagnostically relevant and unmixed than the principal components attained in the immediate prior analysis stage. To obtain the independent components, a FastICA algorithm was employed that is based on maximization of the fourth statistical moment, i.e. kurtosis. It is computationally simple, fast and requires little memory space [21].

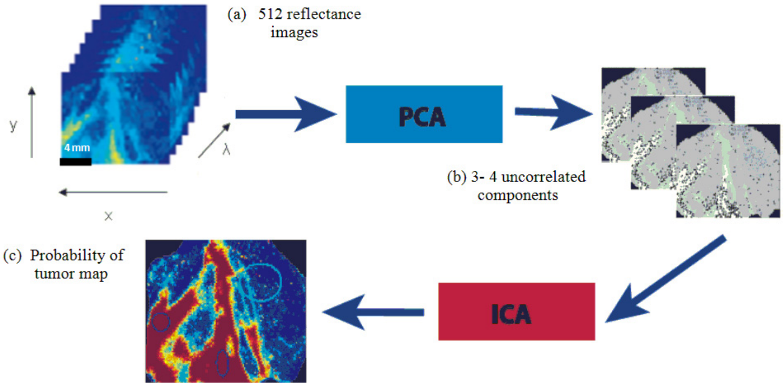

(a) 512 reflectance images
y
4 mm
λ
x
PCA
(b) 3- 4 uncorrelated components
(c) Probability of tumor map
ICA

Fig. 1. ICA-PCA analysis process on the reflectance images.

2.2.4 Solving ambiguity problems from PCA and ICA

The two linear analyses, PCA and ICA, jointly described as in Eq. (3), have another point in common that needs to be emphasized: sign ambiguity. The resulting component $\mathbf{s}$ could be multiplied by -1 and the solution of Eq. (3) would be identical and correct. Moreover, ICA has another ambiguity on the order of result components.

PCA analysis is based on the extraction of the singular value decomposition matrix (SVD) and there is mathematically no way to avoid this sign ambiguity arising from a multiplicative term such as the pair of singular vectors [21]. In ICA, the variance of the independent components cannot be determined [20], so the magnitudes of the independent components may be fixed, but this still leaves the ambiguity of the sign. Several strategies have been tested to deal with the ambiguity problems of BSS analysis. The sign ambiguity of PCA can be solved with the spectral interpretation of the unmixed signals, $\sigma_n(\lambda)$. However, ICA in this case does not permit a spectral interpretation. Concerning order ambiguity, while PCA components are ordered by variance, the intrinsic order ambiguity of ICA impedes a discriminating rank of independent components [20].

FastICA is a recursive algorithm that starts with an initial guess. If this initial guess is not fixed, the algorithm begins with a random matrix resulting in different signs and orders of the output signals, even if it initiates from the same set of measurements. The **W** matrix resulting from PCA, i.e. the uncorrelated coefficients, is proposed as the initial seed to limit this ambiguity effect. Additionally, surgeon-guidance is proposed to compensate for this ambiguity. Visual inspection of results reveals that one significant independent component is sufficient to distinguish benign from malignant pathologies. The surgeon could guide selection of the significant component or alternatively, the significant component could be identified by cross-correlating a digital photograph of the sample with its spectrally-derived ICA parameters, mimicking the surgeon's viewpoint. Even though sign ambiguities in the magnitude of the selected independent component still would induce an error in the tissue category assignment. To this end, a calibration method is employed in which the user, ultimately the surgeon, specifies a set of known pixels, i.e. obviously malignant tissue at the center of the lesion. Informed with this initial information, ICA then provides a map of tumor extent.

2.3 Well-known point's strategy

The requirement for the sign ambiguity in ICA to be corrected and the approach to implement this correction vary among applications [20], although a majority of them are based on the employment of supervised classifiers after the ICA process. In the present tissue diagnosis application, different pairs of tissue regions (adipose-benign, adipose-malignant, benign-malignant) become well separated by the PCA-ICA combination but the sign ambiguity introduces a constraint in the tissue category assignment to perform an absolutely blind selection. In the validation against the pathologist-based diagnosis, it is precisely the sign of the magnitude of the selected independent component the one that differentiates between tissue diagnoses. A pair of two possibilities of diagnosis can be considered corresponding to positive sign and negative sign regions. However, the same sign is not always associated to the same pathology for different patients.

The proposed procedure however needs some *a priori* knowledge about the sign that is associated per tissue type, since this information is required to specify the associated tissue category (malignant or non-malignant) in the final guidance map. Taking advance of both their experience and the information provided by pre-interventional techniques, surgeons are able to clearly identify the tumor center and healthy tissue. The main difficulty they face is the accurate delineation of the malignant area far from the center. This is the point where the proposed guidance map would be of great interest. Once surgeons are asked to locate malignant and non-malignant centers, these points will work as calibration points for the algorithm identifying the actual sign for malignancy regions. In order to emulate this surgeon selection, 25 pixels on each ROI have been selected as "well-known", contrasted points to be certainly diagnosed. Then a detection mask can be easily created. This process is summarized in Fig. 2.

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

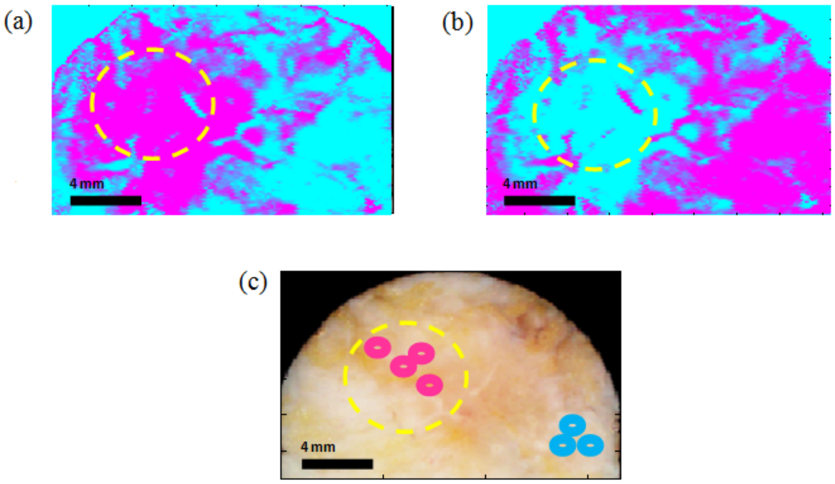


Fig. 2. ICA discriminant results with approximate scale bar: (a) and (b) are both correct solutions due to sign ambiguity, where the magenta (strongest tone) are values with positive signs. When surgeon selects some points of the image where the pathology is sure (circles on (c)) the region of malignancy (striped ROI) can be defined as a positive (magenta) region, avoiding the sign ambiguity.

3. Results and discussion

The goal of this paper is to design a blind data analysis, i.e. without model fitting, to segment tumor from normal tissues in lumpectomy specimens using localized, measured broadband reflectance. This blind analysis is designed to discriminate between areas in a single tissue sample and not between samples. This is acceptable for margin detection because the success of BCT is measured by accurate tumor delineation within each patient. Results will compare the performance obtained by PCA, PCA-ICA and the extraction of optical parameters according to an empirical approximation to Mie theory [18]. The metrics considered to address the performance are the probability of detection and false alarm (sensitivity and specificity) of PCA-ICA and PCA itself.

3.1 PCA results: uncorrelation has less strong diagnostic ability than independency

For blind signal separation, PCA is applied to reduce the dimensionality of broadband reflectance data, to estimate the number of components used to inform a diagnosis, and to analyze if their diagnostic relevance.

Kept variance presents different slopes depending on each tissue sample. A dynamic threshold based on the derivative of the kept variance curve has been empirically selected. To this end, the L maintained components will be those whose kept variance varies more than 0.2% from the previous set of $L\text{-}1$. The resulting number of maintained uncorrelated components varies from 2 to 7 in the data set, being usually 3. Figure 3 shows two different cumulative variance plots corresponding to two different samples: normal-adipose and malignant-adipose. The first few components correspond to the large eigenvalues, while the components on the right part of the graph have small eigenvalues and are presumed to be related to noise. The reflectance spectral map of sample 1 (normal-adipose) is more uniform than the one of sample 2 (malignant-adipose) due to their different tissue composition. This spectral fact makes that the proposed dynamic criteria would select 5 components for sample 1 (black) instead of the 3 components of sample 2 (red) to account for significant reflectance content.

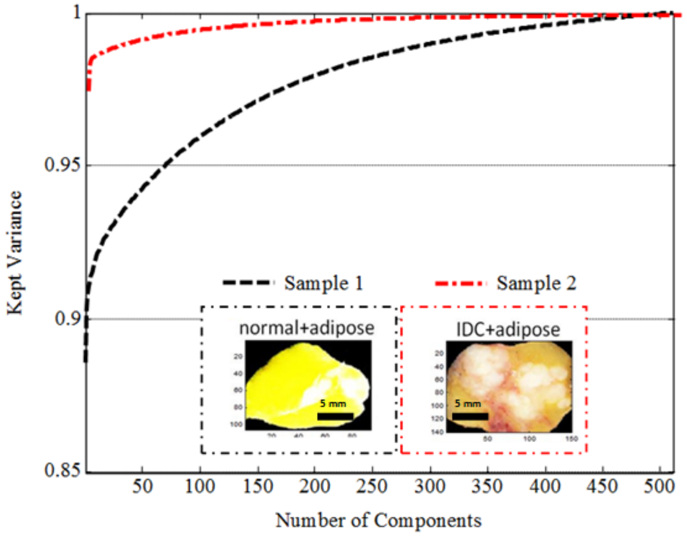


Fig. 3. Example of kept variance for different tissue samples: sample 1 (black) consists of normal-adipose tissue combination; sample 2 (red) consists of malignant-adipose tissue.

When the **A** matrix of PCA coefficients of Eq. (3) is qualitatively observed, the first principal coefficient displays a constant spectral tendency while the second shows exponential or negative logarithm behavior. Figure 4 represents the mean of the first three principal coefficients along the 29 samples. The optical system is optimized not to detect absorption, but just scattering [18]. Nonetheless the third principal coefficient exhibits high correlation with absorption by hemoglobin.

Considering the Mie linear approximation of reflectance as noted in Eq. (4), the similarity with PCA results is found as stated by Eq. (8):

$$\ln\left(R\left(\lambda\right)\right)=pc_1-pc_2\ \ln\left(\lambda\right)-pc_2\,K\left(\lambda\right) \tag{8}$$

Being $K\left(\lambda\right)$ the exponent of hemoglobin absorbance on the model and pc_n would be the PCA scores, as defined in Eq. (4).

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

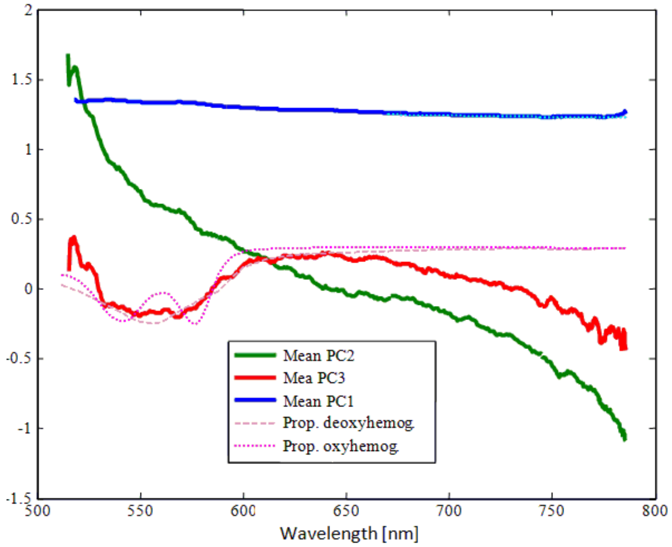


Fig. 4. Mean spectral variations of PC1, PC2 and PC3 coefficients on matrix (**A**) for the whole number of breast samples, where "prop." means proportional.

If this supposition of likeness was right, similarity between PCA scores and model-based parameters should be found. Table 2 shows the correlation between the mean PCA scores and optical parameters extracted from the empirical approximation fitting given by Eq. (1), to assess the contribution of scattering power and hemoglobin absorption to each score.

Table 2. Correlation Study To Determine the Hemoglobin Presence on Each Principal Component*

Principal Component	Mean correlation with scattering from Mie Fitting	Mean correlation with oxyhemoglobin	Mean correlation with desoxyhemoglobin
First	0.2403	0.1394	0.1362
Second	0.7924	0.3439	0.3226
Third	0.2959	0.4850	0.5049

*As described, they are ordered by its contribution to variance (eigenvalues).

Although a high correlation with Mie power scattering is found on PC2 (Fig. 4), and hemoglobin absorption is usually collected on PC3 (Fig. 4), this relationship does not necessarily define PC2 and PC3 as scattering power and hemoglobin, like in a conventional model fitting extraction. However, some similarities are found on the behavior of the statistical features (PCA scores) and the optical features (scattering and absorption from model) which may suggest that BSS analysis accounts for physical variation of parameters of the tissue. Figure 5 shows the maps of the PC2 scores (Fig. 5(a)) for a specific tissue sample when compared with the scattering power map (Fig. 5(b)) obtained from Eq. (1). High correlation between both maps can be observed that is shown also in the associated scatter plot (Fig. 5(c)). Figure 6 represents the influence of the hemoglobin in the scores of PC3. In the digital photograph of the sample (Fig. 6(a)) some blood pools can be observed. The spectral variation of PC3 (Fig. 6(b)) shows similarities with the hemoglobin spectrum and the map of scores of the PC3 exhibits high values in the areas where the blood pools are located (Fig. 6(c)).

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

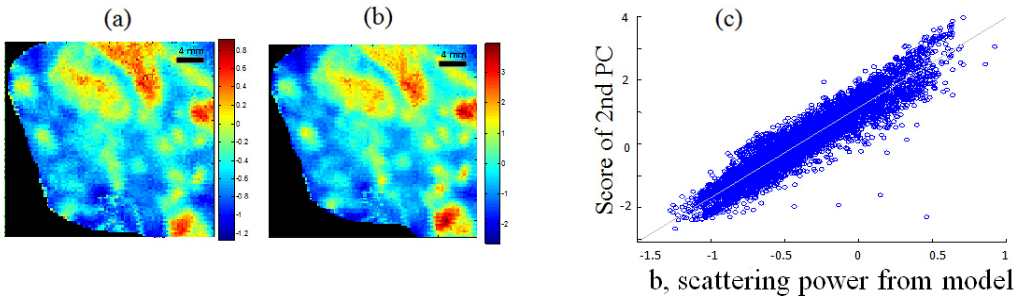


Fig. 5. Blind analysis and physical properties of scattering with an approximation of the scale bar: (a) scattering power map from model fitting; (b) score of PC2; (c) scatter plot showing their correlation.

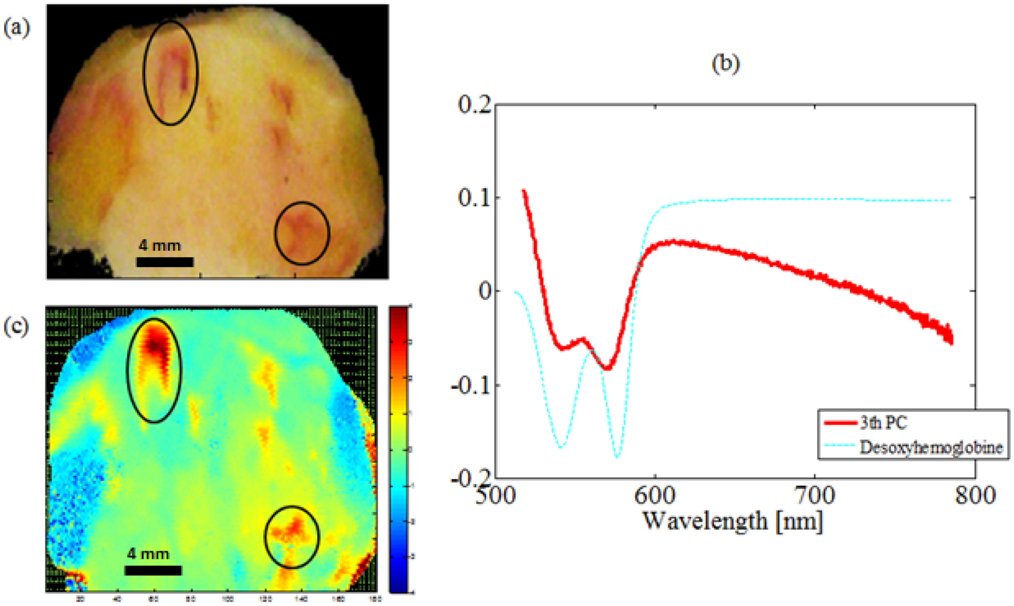


Fig. 6. Blind analysis and physical properties of absorption: (a) digital photograph with blood points in black circles and approximate scale bars, (b) Coefficient, (eigenvector) of the 3th principal component, with its spectral variation, (c) PC3 score.

ICA was then applied to the three most significant PCA components, determined by kept variance, to extract the independent maps of each sample, for improved discrimination. These independent features were used to classify the tissues, and results were compared with the uncorrelated features to check if the independency stronger statistical condition translates into better classification accuracy.

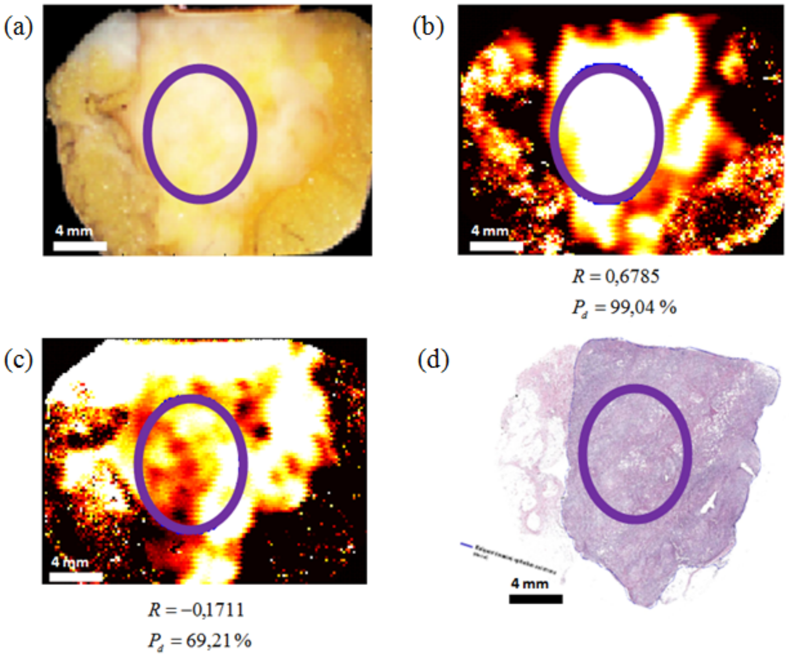
3.2 ICA for the extraction of independent maps

While PCA ambiguity is easy to solve through a correlation with the spectral signatures of tissue chromophores, FastICA is an iterative algorithm that causes two types of ambiguities and such a study is not so straightforward. Because of this, and as mentioned above, PCA matrix (coefficients) is used as the initial seed for the ICA algorithm. Even under this premise, it was not possible to deal with these ambiguities analytically.

By visual inspection it seems easy to determine which component is most discriminating, but automation of this analysis is desirable. The digital photograph of each tissue sample also provides useful information as it mimics surgeons vision. Correlation with the digital photograph is proposed as a fast solution to determine which ICA score is more interesting for

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

diagnostic purposes. This procedure could emulate the surgeon behavior. Figure 7 shows one of the tissue samples and the probability of detection (P_d) and degree of correlation (R) when compared with the digital photograph of the tissue (Fig. 7(a)) in case of selection of the last IC (Fig. 7(b)) or the penultimate IC (Fig. 7(c)). The H&E section is also shown (Fig. 7(d)) for visualization purposes. In this sample, the last IC exhibits the highest correlation with the digital photograph and also achieves the highest probability of detection when results are validated against the ROI's information provided by the pathologist.



$R = 0{,}6785$
$P_d = 99{,}04\,\%$

$R = -0{,}1711$
$P_d = 69{,}21\,\%$

Fig. 7. IC selection based on correlation with the digital photograph of a malignant sample. Circles correspond to the areas identified as malignant by the pathologist, scale bars are approximated. (a) Digital photograph of the breast specimen; (b) last IC with its correlation and probability of detection; (c) penultimate IC with its correlation and probability of detection; (d) H&E section of the sample.

As explained in the previous section, after choosing the most appropriate independent score, the "well-known points" calibration is performed to solve the sign ambiguity. The selected ICA score is expected to be the best for classification purposes and it is also supposed to be more related with single scattering feature, as interferences from absorption and other attenuation contributions are supposed to be minimized by the optical set-up and then is expected to be associated with the discarded independent component maps. Table 3 shows the mean discrimination and standard deviation of each cluster of PCA scores when compared with the attained with the ICA. Figure 8 shows the maps of probability of tumor for different samples compared with the digital photograph, the H&E section and the pathologist diagnosis. The PCA-ICA process provides the highest separation between the different pathologies with very similar standard deviation, achieving a high degree of accuracy with the pathologist decision. This makes easier to implement a linear classifier, which is fast and computationally simple.

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

Table 3. Separation Between Normalized Mean Data from Different Pathologies Depending on the Score Chosen for the Diagnosis

	PC1	PC2	PC3	IC
Mean pathologies separation	0.15	0.17	0.05	0.44
Pathology cluster standard deviation	0.002	0.005	0.002	0.009

Then, after the best ICA score for classification purposes is selected, and according with to the values in the ICA space, a probability of tumor can be calculated. Figure 8 shows the results for 4 different samples with diverse pathologies malignant (purple), non-malignant (white) and adipose (cyan). The last column shows the images of the map for tumor probability and good agreement can be found by visual inspection when compared with the H&E section and the pathologist ROIs. For a more quantitative assessment, Table 4 shows sensitivity and specificity outcomes detecting malignant points and a comparison of strategies to fix and select the diagnostic map. The best results depend on the chosen ICA score, being better when it becomes fixed by visual inspection. The correlation with the digital photograph of the specimen helps in the determination of the best score but its performance is a bit lower than the provided by PCA first score.

Table 4. Comparison Between the Outcomes of Sensitivity and Specificity of a Classifier of Malignancy for the Different BSS Strategies and a Supervised Technique as KNN

Strategy	Sensitivity	Specificity	Advantage	Disadvantages
Best ICA score selected by surgeon	0.963	0.958	Far best in sensitivity and specificity	Chosen by visual inspection
Digital photograph correlation for selection of ICA score	0.932	0.810	Order ambiguity solved automatically, better specificity than PCA itself	Slower
2nd PCA score (which presents high correlation with model-based scattering parameter)	0.861	0.743	No need of ICA so slightly faster	On average the worst option
Optical parameter from model + KNN classifier [18]	0.91	0.77		Inherent problems of analytical modeling and supervision and complexity of KNN classifier

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

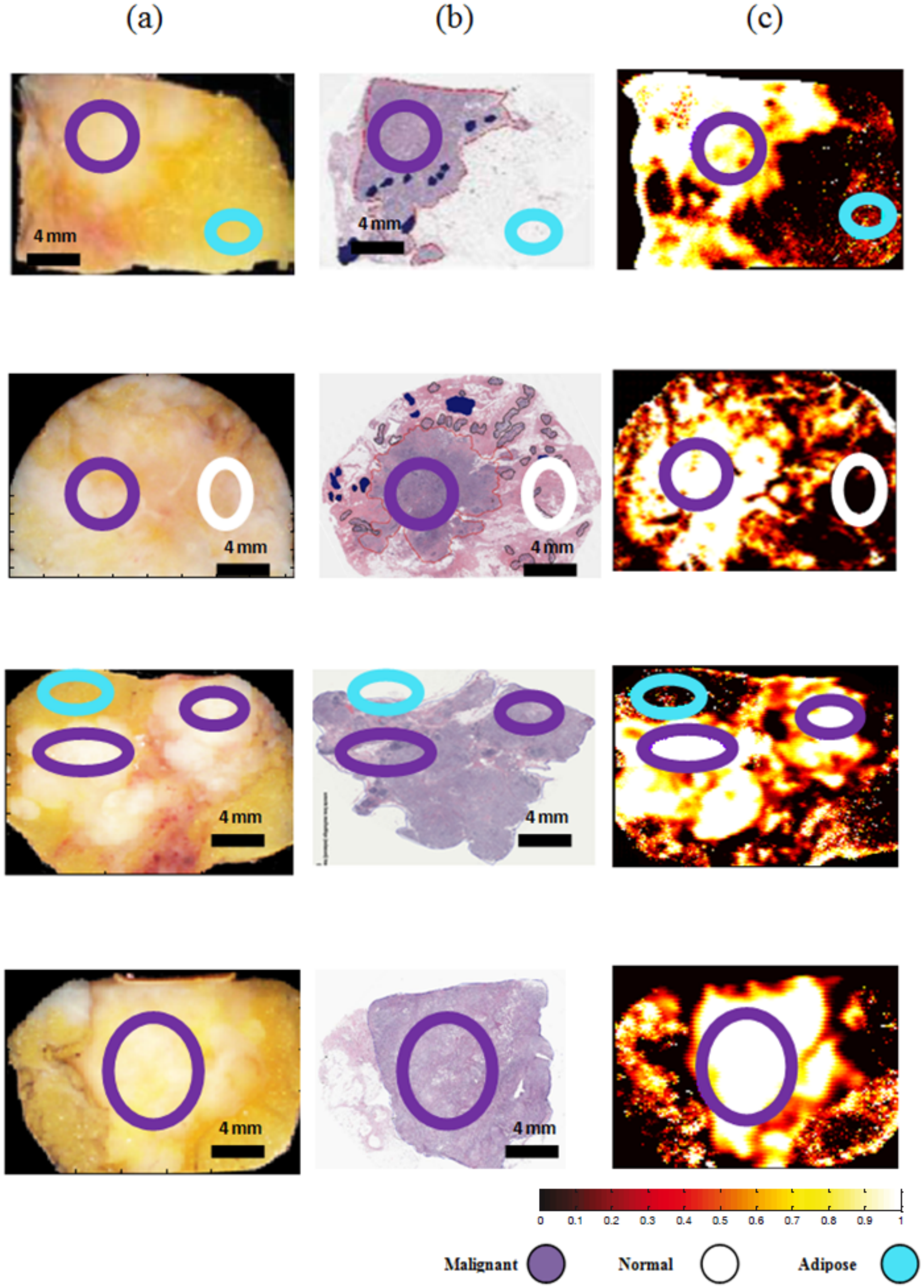


Fig. 8. Tumor diagnosis results, where the pathologist evaluation is shown with different colored circles and the scale bar is an approximation: (a) photograph image; (b) co-registered H&E section; (c) probability of tumor map.

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013

4. Conclusions

The feasibility of PCA and ICA to blindly detect and localize breast tissue pathologies is proposed and successfully checked in this paper. To preserve the tissue properties, elastic scattering that requires low optical power is used. The analysis here proposed is designed to discriminate between tissue areas within a single sample, and not between samples, with the ultimate goal of surgeon guidance for Breast Conserving Therapy purposes.

Computationally efficient BSS analysis has been directly applied to 512 optical localized reflectance breast measurements, instead of reflectance model fitting, to readily identify their corresponding cancer pathologies. Reflectance is directly obtained from endogenous tissue properties, mainly scattering from tissue morphology, without injection of contrast agents that require expensive biocompatibility studies and regulatory approval for clinical use.

PCA reduces the dimension of the data set, from the initial 512 spectral bands to just 3-5 uncorrelated components. The latter exhibit significant similarities with the parameters extracted based on an empirical model based on the Mie theory, specifically scattering power and the hemoglobin absorption spectrum. They can additionally be used as classification features by applying a linear threshold. However, the statistical feature of uncorrelation is softer and less significant than statistical independence so ICA has been employed to compare results of classification. Combined PCA-ICA analysis has provided the best significant diagnosis maps with probability of tumor information. Discriminating spectral information, sometimes lost in empirical approximations of light scattering, contributes here to a better tissue type separation.

Sign ambiguities limiting discrimination by ICA have been resolved by selecting some "well-known" points that the surgeon can provide in a real scenario to determine a calibration environment. However, ambiguity arising from the order in which the scores are generated has been a challenge. The selected criterion to confront this ambiguity is to correlate the ICA results with the corresponding digital photograph of the tissue. The best sensitivity-specificity possible attained with ICA is 96%-95% while "photograph correlation for selection" solving proposal yielded 93%-81%. Therefore a loss in the classification is induced if the selection of best score is not optimized. However, both ICA solutions are still better choice than the selection of the second PCA score, which presents 86%-74%.

Furthermore, important correlation between tumor probability and H&E maps is also obtained, which suggest that a future application of the system could be margin delimitation. The goal of this approach is not to diagnose malignancy but to map its extent. During surgery the tumor is already localized, so a seeding of the algorithm by the surgeon is feasible.

To conclude, this contribution validates and optimizes the ability of PCA and ICA to blindly detect breast tissue pathologies. Tissue features related to elastic scattering and blood absorption have been extracted from label-free localized reflectance measurements, using no training information nor empirical models, although further contrast of this aim needs to be proven based on tissue simulating phantoms of known optical properties. Even though, PCA and ICA extract significant features to provide a map of tumor probability to be used in an intraoperative context.

Acknowledgments

This work has been supported by the Spanish Government through CYCIT projects DA2TOI (FIS2010-19860), TFS (TEC2010-20224-C02-02) and Alma's Eguizabal PhD Grant (FPU12/04130) and by Dartmouth College.

#187362 - $15.00 USD Received 27 Mar 2013; revised 10 May 2013; accepted 21 May 2013; published 12 Jun 2013