Automated segmentation based upon remitted scatter spectra from pathologically distinct tumor regions

P. Beatriz Garcia-Allende*[a], Venkataramanan Krishnaswamy[b], Kimberley S. Samkoe[b], P. Jack Hoopes[b c], Brian W. Pogue[b c], Olga M. Conde[a], Jose M. Lopez-Higuera[a]
[a]Photonics Engineering Group, University of Cantabria, Avda. Los Castros S/N, 39005 Santander, Spain;
[b]Thayer School of Engineering, Dartmouth College, Hanover, NH 03755;
[c]Department of Surgery, Dartmouth Medical School, Hanover, NH 03755

ABSTRACT

Multi-spectral scatter visualization of tissue ultra-structure *in situ* can provide a unique tool for guiding surgical resection, but since changes are subtle and the data is multi-parametric, an automated methodology was sought to interpret these data, in order to classify their tissue sub-type. Tissue types observed across AsPC-1 pancreatic tumor samples were pathologically classified under three major groups (epithelium, fibrosis and necrosis) and the variations in scattering parameters, i.e. scattering power, scattering amplitude and average scattered intensity, across these groups were analyzed. The proposed scheme uses statistical pre-processing of the scattering parameter images to create additional data features followed by a *k*-nearest neighbors (kNN) based algorithm for tissue type classification. The classification accuracy inside some predefined regions of interest was determined and the mean region values of scattering parameters turned out to be stronger data sets for classification, rather than the individual pixel values. This presumably indicates that pixel-to-pixel variations in the remitted spectra need to be minimized for reliable classification approaches. Results show a strong correlation between the automated and expert-based classification within the predefined regions of interest.

Keywords: automatic classification, tumor, necrosis, confocal reflectance imaging, scatter, feature extraction, *k*-nearest neighbors (kNN)

1. INTRODUCTION

Detection of scatter changes associated with tumor boundaries can help a surgeon to decide on surgical margins in near real-time. Therefore, a device able to quantify tumor-associated scatter changes *in situ* would be a great asset in assisted surgery.

In a previous study [1] a raster scanning reflectance imager was developed and its ability to measure tumor associated scatter changes was demonstrated. However, no consistent trend was encountered in the scattering parameters across the different tissue types. Changes are subtle, data is multi-parametric and, as a consequence, an automated kind of interpretation into what the signal means relative to the pathology is required before proceeding to clinical studies. A methodology able to quantify in an automated manner the scattering coefficient heterogeneities and to establish a correlation with their corresponding tissue morphologies is reported here.

The proposed methodology consists of two distinct steps. Initially, statistical pre-processing is performed to create additional data features from the measured scatter parameter images. In this regard, the mean, standard deviation, skewness and kurtosis of each pixel are computed in the immediate spatial vicinity of the latter and then, these data are concatenated with its original scattering parameters. In the second step, the *k*-nearest neighbors (kNN) algorithm [2], whose classification capability has been demonstrated in a wide variety of scenarios ranging from face recognition [3] to food industry [4], is used on the processed dataset to perform the classification. In [1] a veterinary pathologist (P.J.H.) analyzed the H&E stained sections of the measured samples and identified several regions of interest corresponding to the tissue types observed. These tissue types were classified under three major groups: epithelium, fibrosis and necrosis with some constituent subgroups. According to the exhibited nucleus to cytoplasm ratio, epithelial cells were classified in high and low proliferation index tumor cells, while three different fibrosis subgroups (early, intermediate and mature)

Biomedical Applications of Light Scattering III, edited by Adam Wax, Vadim Backman
Proc. of SPIE Vol. 7187, 718717 · © 2009 SPIE · CCC code: 1605-7422/09/$18 · doi: 10.1117/12.808322

were also distinguished. The ability of the methodology to correlate the scatter changes with their corresponding tissue type among these six within the expert predefined regions of interest was determined as a function of the size of the spatial vicinity considered in the statistical pre-processing of the scatter images.

2. MATERIALS AND CLASSIFICATION METHODS

2.1 Scatter imaging system and parameter fitting

The scatter imaging system primarily consists of a confocal spectroscopic system and a raster-scanning sample platform built using translation stages. The optical and electromechanical subsystems are integrated via a custom developed LabVIEW interface. Figure 1 shows a photograph and a schematic (in the inset) of the measurement set-up. The illumination optics train consisted of a 50 µm core fiber (F1) coupled to a 100W tungsten-halogen white light source placed at the front focal point of an achromatic lens (L1). A 10X, long working distance, plan-apochromatic objective (L2) was used to refocus the light on to the sample plane. The illumination optics train was modeled in Zemax software to make sure the illumination spot size was less than one scattering length (typically 100 µm) over the entire wavelength band. This assures directly imaging of the scatter from tissue. The detection optics train used the same microscope objective to pick-up the backscattered light from the sample and a 50/50 beam splitter was used to separate the illumination and detection beam paths. Another achromatic lens (L3) was used to focus the detected photons on to the proximal face of a 100 µm core detection optical fiber which also acted as the confocal pinhole. The size of the detection spot on the target was controlled by the detection fiber's core diameter and the lateral magnification of the optical system. The spectrometer was calibrated to operate in the wavelength range of 510 nm – 785 nm that encompasses the strong hemoglobin absorption peaks. A more detailed description of the system and its calibration and characterization procedures can be found in [1].

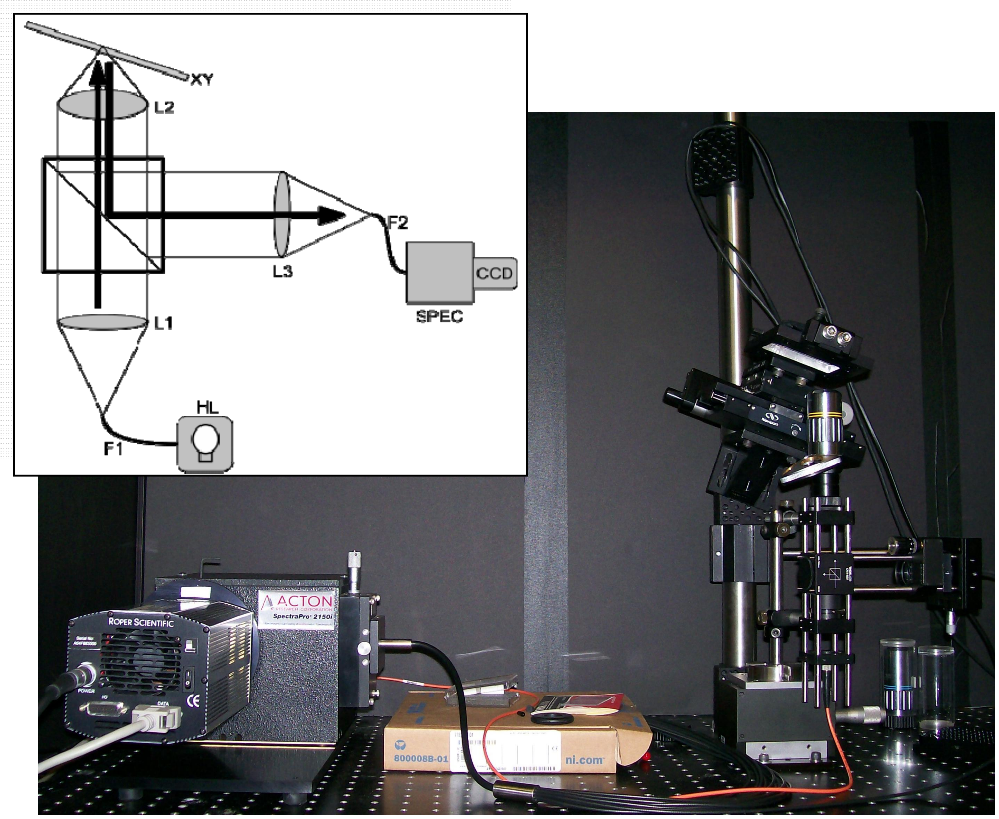


Fig. 1. Photograph and schematic (upper inside) of the scatter imaging system, with inverted microscope design, and light coming from a tungsten lamp into the system, and the detected light coupled through a fiber into the spectrometer-CCD detection system.

The acquired spectral reflectance is fitted by the empirical relationship, which accounts for scatter versus wavelength, and then exponential attenuation to absorption.

$$I_R = A\lambda^{-b} e^{(-k \cdot c \cdot (d \cdot HbO_2(\lambda) + (1-d) Hb(\lambda)))} \tag{1}$$

where A is the scattered amplitude, b the scattering power, c is proportional to the concentration of whole blood and d is the oxygen saturation fraction. This relationship is valid in the presence of significant local absorption, for very small source-detector separation and when the scattering and absorption coefficients are within the typical range found in tissue [5,6]. The extinction spectra of oxygenated and de-oxygenated hemoglobin, $HbO_2(\lambda)$ and $Hb(\lambda)$, were obtained from the Oregon Medical Laser Center data base. Absorption from other chromophores were assumed to be negligible in the wavelengths from 510 to 785 nm as mentioned before, and the path length k is assumed to be a wavelength independent constant.

Along with the scattered amplitude and the scattering power, the third scattering parameter of interest was the average scattered irradiance, I_{avg}, which was obtained by integrating I_R over wavelengths beyond 610 nm to avoid the strong hemoglobin absorption peaks.

2.2 Pancreatic tumors

Data from AsPC-1 pancreatic tumor samples from a previous study [1] was used. In this regard, human pancreatic tumor cells AsPC-1 were grown and injected subcutaneously in the flank region of male mice. Tumors were harvested seven weeks after injection until they measured 6 – 7 mm in diameter and 5 – 6 mm in thickness. Then they were dissected into 4 – 5 mm thick sections and imaged using the system described in Section 2.1. After the measurement, the sample was routinely processed for subsequent histology evaluation by a veterinary pathologist.

2.3 Classification methods

A schematic of the discrimination of pathologically distinct tumor regions by means of the kNN classifier is depicted in Figure 2. For visualization purposes, only those tumor pixels whose scattering parameters are smaller than the mean value of each scattering parameter plus a third of its variance, and greater than the mean value minus a third of the variance are depicted in the $A - b - I_{avg}$ space. For classification, every pixel inside the pre-defined regions of interest is considered as a vector in this 3-dimensional space, the so-called *feature space*. Since kNN is based on the idea that similar data should belong to the same class, the *k* pixels with the most similar scattering parameters to an unclassified pixel are initially determined. This similarity is measured in terms of the Euclidean distance, *D*, expressed as:

$$D(p_1, p_2) = \left((A_1 - A_2)^2 + (b_1 - b_2)^2 + (I_{avg1} - I_{avg2})^2\right)^{1/2} \tag{2}$$

where $p_1(A_1, b_1, I_{avg1})$ and $p_2(A_2, b_2, I_{avg2})$ are the two compared tissue pixel localizations. The unclassified pixel is assigned to the most numerous tissue sub-type (high or low proliferation index epithelium, necrosis, and early, intermediate or mature fibrosis) among the closest *k* neighbors. The kNN is, therefore, able to naturally deal with multiclass data while some of the more advanced classifiers such as Support Vector Machines (SVM) [7], require to combine the results of a combinatorial sets of such classifiers to accurately simulate multiclass results [8].

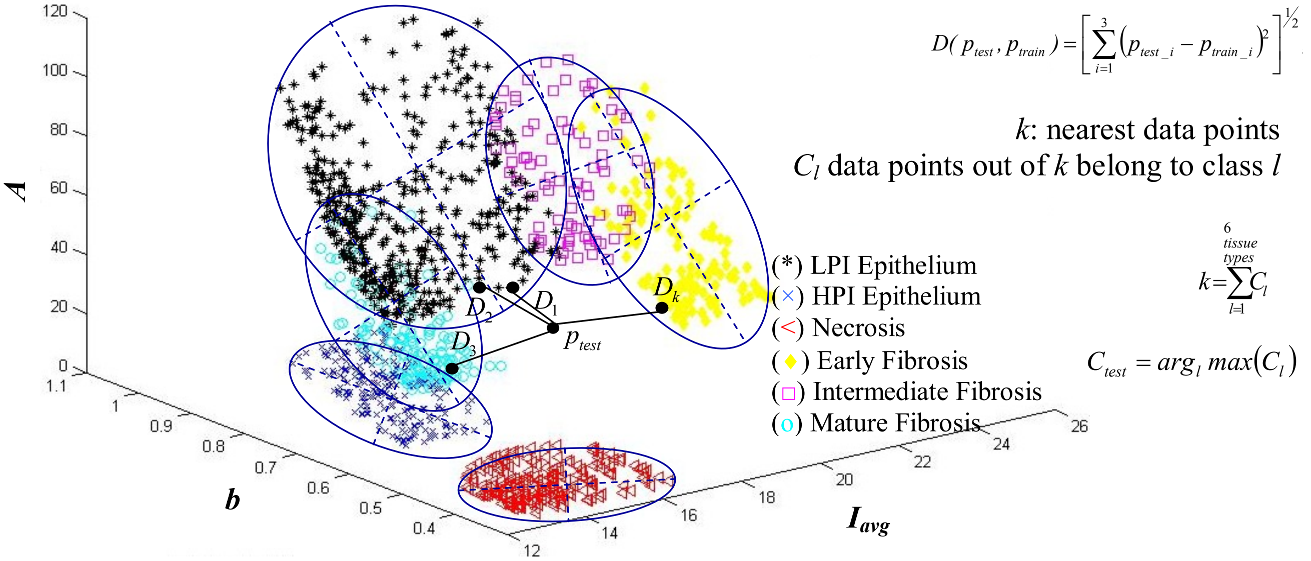


Fig. 2. Schematic of the tumor sub-type discrimination by means of the kNN classifier. The axes on this plot are scattering amplitude (A), scattering power (b) and average total scatter irradiance I_{avg}. Although the centroids are clearly different, only part of the data is plotted to allow better visualization of the results. The full data set is plotted in the next figure.

As stated in (2), kNN is based on distances between sample points in the feature space. Hence, scattering parameters are initially unity normalized to assure that none of them are strongly weighted than others. In addition, to attain generality in the determination of methodology capabilities, i.e. to avoid tissue discrimination results dependent on the training or test sets, a *cross-validation procedure* [9,10] has been implemented. Tumor pixel localizations within the expert-determined regions of interest were divided in 3 non-overlapping sets, containing each one of them the same number of pixel localizations of each tumor sub-type. Two of these sets were employed as the training set, and the other one, the so-called validation set, was used to calculate the error in the segmentation process. This procedure was repeated three times, each time with different training and validation sets. Finally, the estimated performance of the classifier was evaluated by averaging the three resulting errors.

The segmentation capacity of the methodology as described above was expected to be weak, since six different tissue morphologies have to be discriminated from a data set that lies in a 3-dimensional space. The addition of statistics calculation to increase the dimensionality of the data space has been proven successful [8] for breast tissue analysis. The first four statistical moments: mean ($\overline{p}$), standard deviation (σ_p), skewness (S_p) and kurtosis (K_p), of each scattering parameter p were calculated over a square spatial vicinity region centered in each pixel localization. Statistics were then concatenated with the fitted scattering parameters to form a 15-dimensional feature space. In this way, a remarkable improvement was achieved in methodology segmentation capabilities.

3. SEGMENTATION OF PATHOLOGICALLY DISTINCT TUMOR REGIONS

Figure 3 shows the actual 3D map representing the pixel localizations in the $A-b-I_{avg}$ space. The population of the map consists of all tumor pixel localizations within the predefined regions of interest and not only those pixels whose scattering parameters were within the interval mentioned above as in the schematic in Figure 2. As shown, the scattering parameters are not well grouped according to their tissue subtype. This means that no consistent trend is encountered in these parameters across the distinct subtypes, and, an automated interpretation methodology of the changes was therefore required.

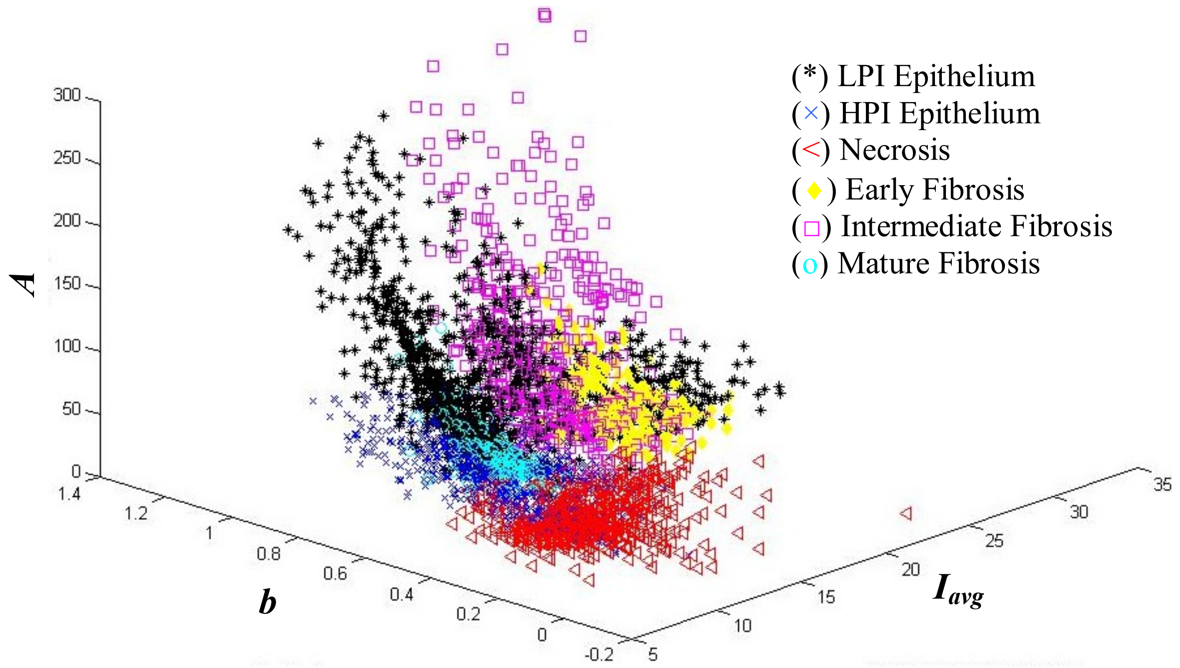


Fig. 3. Grouped scatter plot for tumor pixel localizations in the $A - b - I_{avg}$ space, similar to the previous figure, but including all data points from the images.

Figure 4 depicts the segmentation of a tumor sample into its distinct tumor regions by means of the kNN algorithm as described in the schematic in Figure 2, i.e. without the inclusion of statistics calculation in the processing scheme. Classification accuracy is expected to increase with the number of neighbors, *k*, because this reduces the influence of training data points assigned to a wrong class. Samples were segmented as indicated by the pathologist, and, therefore, a good correlation is achieved between expert-based and automated sample segmentation when each unknown pixel localization is assigned to the same tissue subtype than its closest neighbor, i.e. $k = 1$. However, the limitation of the methodology in sample boundary detection was significant.

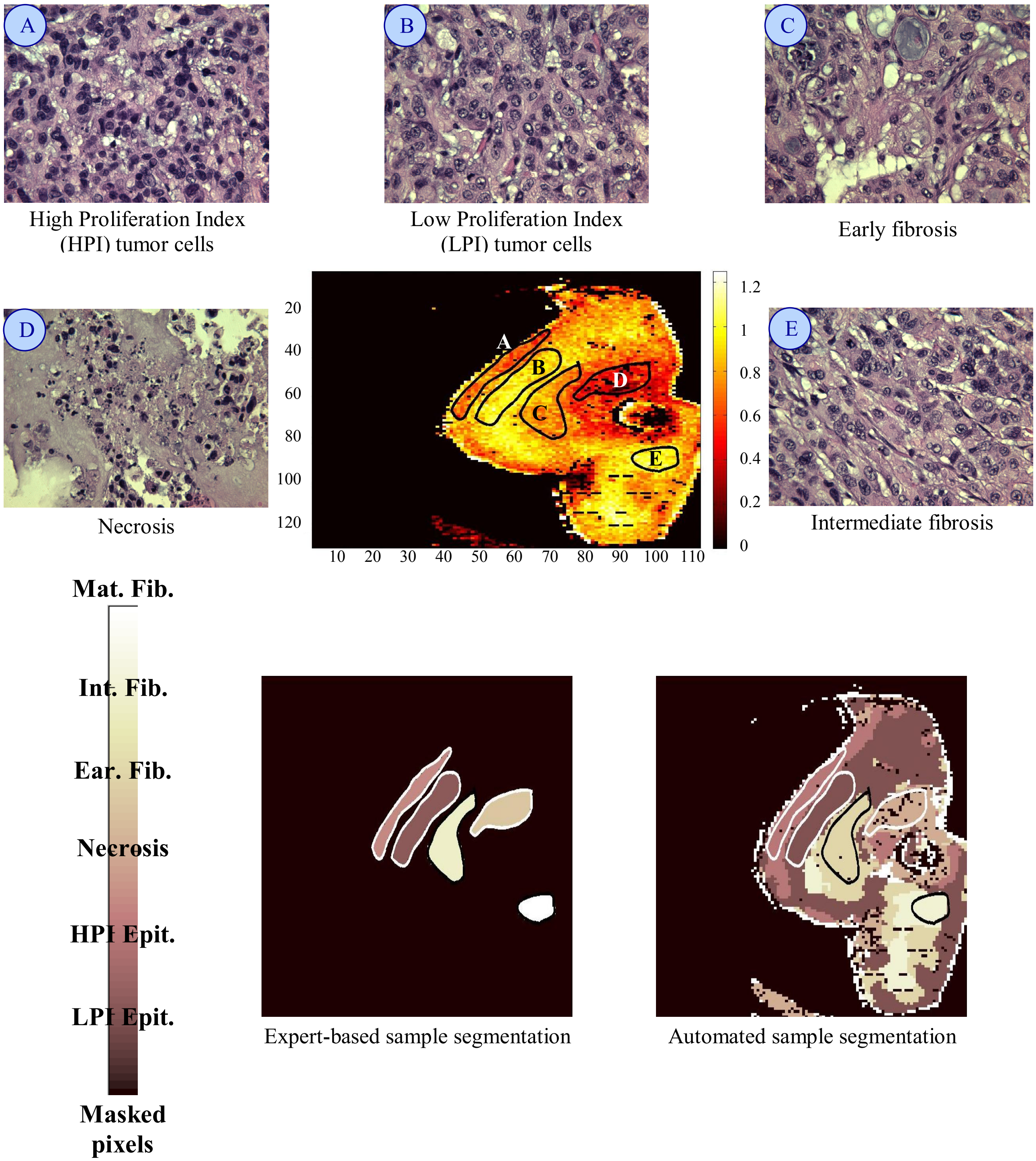


Fig. 4. Tumor sub-type discrimination by means of the kNN classifier (k=1), using a color bar scale to indicate classification into different tissue types. Example histology images of each region are shown above to illustrate the microscopic features which are different in each region.

An extraction procedure of additional features based on statistics calculation is included in methodology in an attempt to achieve a better delineation of the tissue. Figure 5 shows its block diagram. kNN-based tissue classification was performed identically but each pixel was considered now as a vector in a 15-dimensional space, instead of a 3-dimensional space.

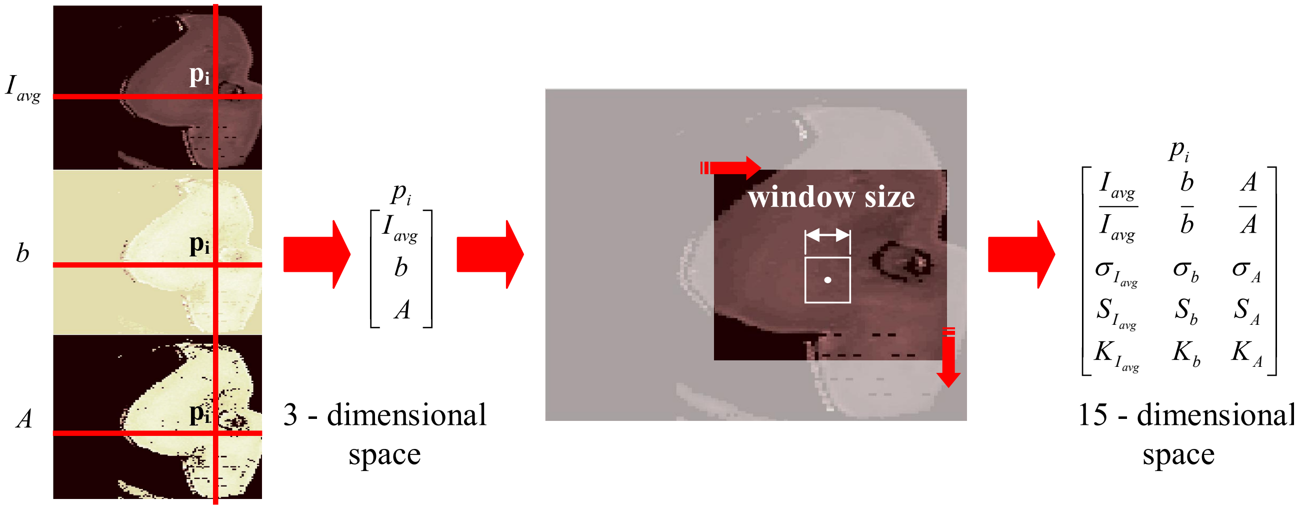
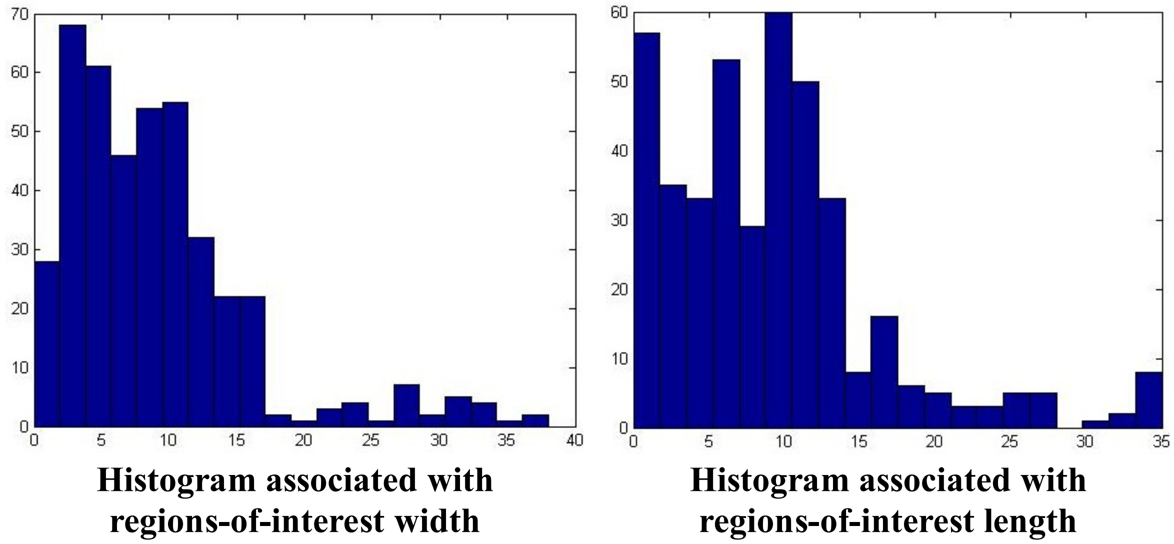

I_{avg}
b
A
pi
pi
pi
p_i
$\begin{bmatrix} I_{avg} \\ b \\ A \end{bmatrix}$
3 - dimensional space
window size
p_i
$\begin{bmatrix} I_{avg} & b & A \\ \overline{I_{avg}} & \overline{b} & \overline{A} \\ \sigma_{I_{avg}} & \sigma_b & \sigma_A \\ S_{I_{avg}} & S_b & S_A \\ K_{I_{avg}} & K_b & K_A \end{bmatrix}$
15 - dimensional space

Fig. 5. Block diagram of the feature extraction procedure, with the three parameter images at left (assembled in to a large 3-dimensional matrix) and when these parameters were spatially processed with the window method for statistical estimates. The resulting data set (including standard deviation, skewness and kurtosis) then occupies a 15 dimensional matrix.

The dimension of the vicinity region employed in statistics calculation, *window size* in Figure 5, was not defined a priori. In preliminary testing, it was apparent that the larger the window size, then the statistics were more accurate, but only if they were predominantly computed for pixel localizations inside the same tissue sub-type, i.e. the window size had to be reasonable from the tissue morphology point of view. To assure this, the histogram of the sizes, width and length, of the regions of interest identified by the pathologist were obtained and shown in Figure 6. Sizes larger than 15 had mostly low occurrence probabilities. Therefore, it was not advisable to attain relevant statistical information. A window size of 12 was employed in the proposed methodology.

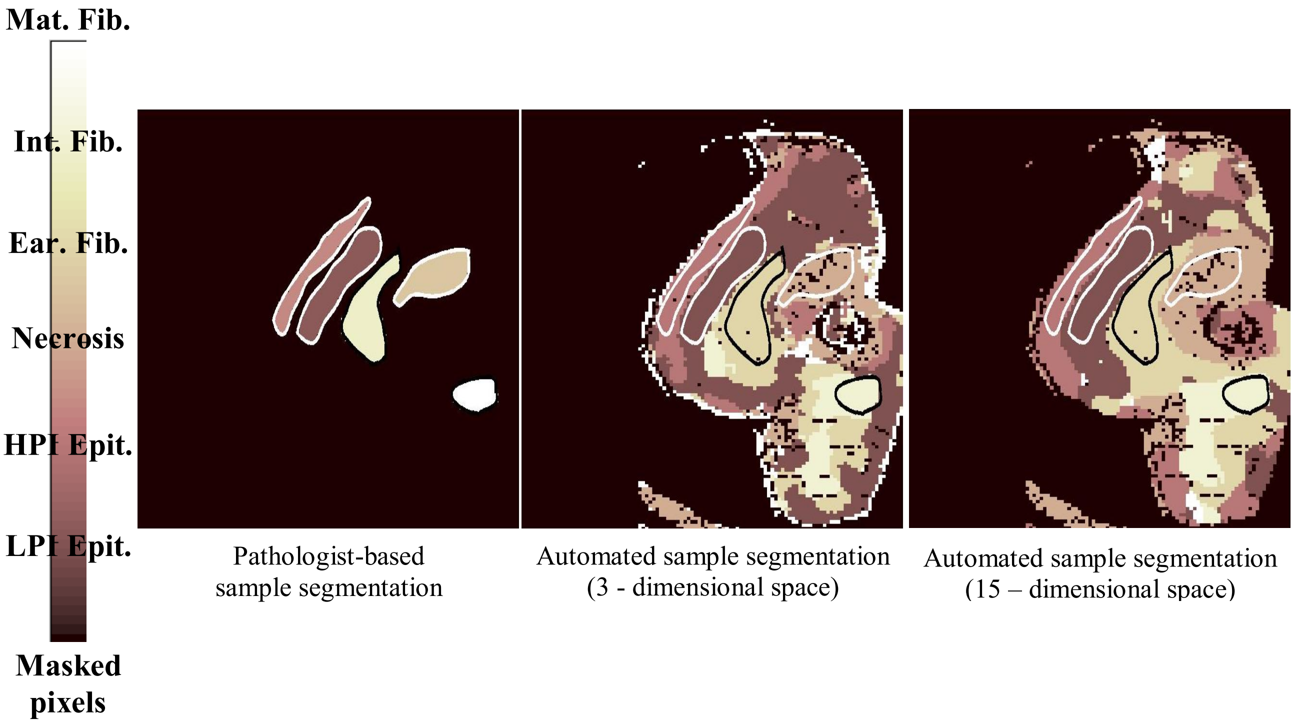


Histogram associated with regions-of-interest width

Histogram associated with regions-of-interest length

Fig. 6. Histogram of the sizes of the regions of interest, showing the change in standard deviation with the size of the region to be used in the window calculation, for both width and length changes.

Figure 7 shows the improvement achieved in sample delineation by the inclusion of spatially-defined statistical estimates of the pixel values. Although it is barely perceptible in the figure, the classification error within the predefined regions of interest with the kNN classifier on its own, was approximately 9 % in the case of using the initial 3 dimensional data set. After using the full data set of 15 dimensional space that included statistical computations of the parameters, the error in classification became lower than 1 %. This improvement in classification is further analyzed in an ongoing study to be published soon, incorporating multiple pathology samples.



Pathologist-based sample segmentation

Automated sample segmentation (3 - dimensional space)

Automated sample segmentation (15 – dimensional space)

Fig. 7. Qualitative comparison among segmentations of a tumor sample in the initial and extended feature spaces are shown, with the pathology (true) regions in (left), and the estimated values in (middle) and (right) for data with only the 3 dimensional space, and then 15 dimensional space, respectively.

4. CONCLUSIONS

An automated methodology into what scatter changes in tissue means relative to pathology was developed. Initially, the segmentation capability when individual pixel-based scattering parameters were processed directly by the kNN classifier was evaluated. Although a good correlation between expert-based and automated region of interest identification was achieved, a weak capacity in tissue delineation was encountered. To improve the latter, a more sophisticated approach was followed. It consisted of inclusion of a pre-processing procedure to generate additional features that were based upon statistical estimates of the standard deviation, skewness and kurtosis between neighboring pixels within a window of interest. That window was scanned around the image to generate statistical values for each pixel. In this way, not only a better delineation of tumor margins was obtained, but also classification error within the predefined regions of interest improved. Mean scattering parameters became stronger data sets for tissue classification than the individual pixel values. This presumably indicates that pixel-to-pixel variations in the remitted spectra need to be minimized for reliable classification approaches. It should be mentioned that, although the feasibility of the proposed technique has been demonstrated, further studies considering different tissue types have to be conducted before it could gain clinical adoption.

ACKNOWLEDGMENTS

P. Beatriz Garcia Allende is grateful for financial support from the Spanish Government's Ministry of Science and Technology through the projects TEC'2005-08218-C02-02 and TEC'2007-67987-C02-01 allowing her to spend three months in the Thayer School of Engineering at Dartmouth College. She also thanks the NIR Imaging Group at the Thayer School of Engineering for their welcoming and their valuable help. Funding at Dartmouth for this work came from NIH research grants PO1CA84203 and PO1CA80139.

REFERENCES

[1] Krishnaswamy V., Hoopes, P. J., Samkoe, K. S. and Pogue, B. W., "A raster scanning reflectance imager for non-model based quantification of tissue scatter", Proc. SPIE 6864, 686402 (2008).

[2] Fukunaga, K., [Introduction to statistical pattern recognition], Academic Press, San Diego, 2nd Edition, 1990.

[3] Conde, C., Ruiz, A. and Cabello, E., "PCA vs. low resolution images in face verification", Proc. of the 12th Int. Conf. on Image Analysis and Processing, 63-7 (2003).

[4] O'Farrell, M., Lewis, E., Flanagan, D., Lyons, W., and Jackman, N., "Comparison of the kNN and neural networks methods in the food industry", *Sensors and Actuators B* 111-112, 354-62 (2005).

[5] Amelink, A. and Sterenborg, H.J., "Measurement of the local optical properties of turbid media by different path-length spectroscopy", *Applied Optics* 43, 3048-54 (2004).

[6] Amelink, A., Sterenborg, H.J., Bard, M.P. and Burgers, S.A., "*In vivo* measurement of the local optical properties of tissue by use of differential path-length spectroscopy", *Optics Letters* 29, 1087-9 (2004).

[7] Vapnik, V., [The nature of statistical learning theory], Springer, New York, 1995.

[8] Oliver, A., Freixenet, J., Martí, R., Pont, J. Pérez, E., Denton, E. R. E., Zwiggelaar, R., "A novel breast tissue density classification methodology", IEEE Transactions on Information Technology in Biomedicine 12 (1), 55-65 (2008).

[9] Goutte, C., "Note on free lunches and cross validation", *Neural Computation* 9, 1211-15 (1997).

[10] Zhu, H. and Rohwer, R., "No free lunch for cross-validation", *Neural Computation* 8, 1421-26 (1996).