Journal of Biomedical Optics

SPIEDigitalLibrary.org/jbo

On-the-spot lung cancer differential diagnosis by label-free, molecular vibrational imaging and knowledge-based classification

Liang Gao
Fuhai Li
Michael J. Thrall
Yaliang Yang
Jiong Xing
Ahmad A. Hammoudi
Hong Zhao
Yehia Massoud
Philip T. Cagle
Yubo Fan
Kelvin K. Wong
Zhiyong Wang
Stephen T. C. Wong

SPIE

On-the-spot lung cancer differential diagnosis by label-free, molecular vibrational imaging and knowledge-based classification

Liang Gao,[a,b,*] Fuhai Li,[a,*] Michael J. Thrall,[c] Yaliang Yang,[a] Jiong Xing,[a] Ahmad A. Hammoudi,[a,d] Hong Zhao,[a] Yehia Massoud,[d] Philip T. Cagle,[c] Yubo Fan,[a] Kelvin K. Wong,[a] Zhiyong Wang,[a] and Stephen T. C. Wong[a,b,c,d]

[a]Weill Cornell Medical College, The Methodist Hospital Research Institute, Department of Systems Medicine and Bioengineering, Houston, Texas 77030

[b]Rice University, Department of Bioengineering, Houston, Texas, 77005

[c]The Methodist Hospital and Weill Cornell Medical College, Department of Pathology and Laboratory Medicine, Houston, Texas 77030

[d]Rice University, Department of Electrical and Computer Engineering, Houston, Texas, 77005

Abstract. We report the development and application of a knowledge-based coherent anti-Stokes Raman scattering (CARS) microscopy system for label-free imaging, pattern recognition, and classification of cells and tissue structures for differentiating lung cancer from non-neoplastic lung tissues and identifying lung cancer subtypes. A total of 1014 CARS images were acquired from 92 fresh frozen lung tissue samples. The established pathological workup and diagnostic cellular were used as prior knowledge for establishment of a knowledge-based CARS system using a machine learning approach. This system functions to separate normal, non-neoplastic, and subtypes of lung cancer tissues based on extracted quantitative features describing fibrils and cell morphology. The knowledge-based CARS system showed the ability to distinguish lung cancer from normal and non-neoplastic lung tissue with 91% sensitivity and 92% specificity. Small cell carcinomas were distinguished from nonsmall cell carcinomas with 100% sensitivity and specificity. As an adjunct to submitting tissue samples to routine pathology, our novel system recognizes the patterns of fibril and cell morphology, enabling medical practitioners to perform differential diagnosis of lung lesions in mere minutes. The demonstration of the strategy is also a necessary step toward *in vivo* point-of-care diagnosis of precancerous and cancerous lung lesions with a fiber-based CARS microendoscope. [DOI: 10.1117/1.3619294]

Keywords: lung cancer; differential diagnosis; coherent anti-Stokes Raman scattering microscopy; machine learning.

Paper 11221R received May 5, 2011; revised manuscript received Jul. 1, 2011; accepted for publication Jul. 6, 2011; published online Sep. 1, 2011.

1 Introduction

Lung cancer among both sexes is the primary cause of cancer deaths in the United States with 222,500 new cases and 157,300 lung cancer related deaths projected for 2010.[1] Worldwide, the five-year survival for lung cancer patients ranges from 6% to 14% for men and 7% to 18% for women, a very dismal prognosis that has not substantially changed in decades.[2,3] Even though early detection of lung cancer has attracted major research interest,[4,5] less than 1% of patients with early-stage lung cancer can be diagnosed.[6] Pulmonary examination using computed tomography (CT) and magnetic resonance imaging does highlight abnormalities. However, these technologies are not often able to distinguish lung carcinoma from benign lesions, such as organizing pneumonia. As a result, a tissue biopsy is still needed as a follow-up test after the detection of a nodule.

Traditional open lung biopsy requires general anesthesia and an invasive surgical procedure. CT-guided percutaneous core biopsy or fine-needle aspiration reduces the amount of tissue taken and complications, though pneumothorax and hemorrhage remain significant concerns. In addition, because of the respiratory motion of patients, it remains difficult to obtain samples precisely at the site of small lesions.[6] Therefore, some patients will need to undergo re-biopsy, resulting in increased costs and delay in diagnosis and treatment. Given the risks and cost of lung biopsy, it would be beneficial to develop techniques that limit damage to lung tissue, diagnose lung cancer in real time, and provide equal or greater diagnostic yield than existing biopsy methods. Consequently, several new technologies have been developed in the past few decades.

Bronchoscopy, for example, has been widely explored for early lung cancer detection. While conventional white light bronchoscopy is based on the detection of alterations in tissue surface structure, autofluorescence bronchoscopy aims at exploiting the spectral difference between normal and pre-/early cancerous tissues.[7,8] Size and specificity are the major limiting factors of this technique since a small fiber optic probe (<1 mm) is needed for diagnosis of peripheral lesions.[6,9] Optical coherence tomography (OCT) is another imaging modality that is compatible with the design of a conventional bronchoscope.[10] Micrometer-level resolution allows *in vivo* investigation and screening for possible lung lesions using light reflected from

[*]Authors contributed equally to this article.

Address all correspondence to: Stephen Wong, The Methodist Hospital Research Institute, Systems Medicine and Bioengineering, 6565 Fannin Street, B5-022, Houston, Texas 77030. Tel: 1-713-441-5884; Fax: 1-713-441-8696; E-mail: STWong@tmhs.org.

1083-3668/2011/16(9)/096004/10/$25.00

within the tissue to generate cross-sectional images.[11] However, OCT only generates contrast using changes of refraction indexes between tissue layers, limiting its specificity and accuracy.

In view of the collective limitations of the techniques discussed above, there is a great demand to develop a real-time imaging tool to increase the biopsy yield and potentially provide diagnostic information to facilitate definitive treatment. This tool would need to offer cellular resolution, fast imaging rate, and molecular specificity, but without the use of exogenous contrast agents or probes, such as fluorescent dyes, since few of these agents or probes have been approved for human use. Current techniques cannot meet one or more of these criteria and thus fall short of their full potential as effective diagnostic tools.

Coherent anti-Stokes Raman scattering (CARS) imaging technique,[12] on the other hand, satisfies all the above parameters and therefore holds great promise for this diagnostic application. It captures intrinsic biomolecular vibrations to create optical contrast with submicrometer level spatial resolution, as well as video-speed imaging rate.[13] In the CARS process, a pump field (ω_p), a Stokes field (ω_s) and a probe field ($\omega_{p'}$) interact with the samples through a four-wave mixing process.[14] When the frequency difference, $\omega_p - \omega_s$ (beating frequency), is in resonance with a molecular eigenvibration, an enhanced signal at the anti-Stokes frequency, $\omega_{as} = \omega_p - \omega_s + \omega_{p'}$, is generated.[15] The major advantage of CARS is that the signal yield is much higher, typically several orders of magnitude, than the signal yield obtained through the conventional spontaneous Raman scattering process.[16]

Because of these advantages, CARS microscopy has been used to visualize various tissue structures, such as skin,[16] lung, and kidney.[13] In the field of cancer imaging, a recent study showed the use of multiplex CARS for interferometric imaging of breast cancer for identification of cancer boundaries.[17] However, differential diagnosis of cancer using CARS microscopy has, to the best of our knowledge, not been attempted. Currently, in order to accurately delineate the type of lesions for definitive treatment, pathologists routinely stain lung biopsy tissue to examine changes in such cellular and histologic features as cell size, cell-cell distance, and formation of fibrous structures.[18] However, while this method is subject to interobserver variations, the CARS technique already provides high-resolution images which can clearly detect these features, without tissue staining with exogenous agents. Therefore, we hypothesize that the development of a label-free imaging and pattern recognition method, whereby such images could be used as a basis for the quantitative classification of these cellular features in a way that would lead to a differential analysis of lung cancer. This hypothesis was supported by our recent publication[19] that studied the differentiation of cancer from normal prostatic glands in order to aid surgical decision on margin status using calculated cellular features from CARS images. One of the cellular parameters (average cell neighbor distance) was determined to be a good candidate for cancer differentiation using principle components analysis. Inspired by the research findings in Ref. 19, this study aims to perform in-depth classification analysis and calculate the accuracy of the classifier using a leave-one-out training and testing design. In other words, potential cellular parameters are not only selected, but also tested, for constructing a classifier to separate different types of lung lesions. The current study will provide comprehensive and robust results with regard to cancer differential diagnosis using the CARS-based technology *ex vivo*, and will bring the proposed approach of coupling label-free imaging with pattern recognition for cancer diagnosis closer to clinical applications. Accordingly, the established pathological workup and diagnostic features were used as prior knowledge for establishment of a knowledge-based CARS classification module using a machine learning approach. This module was integrated with the CARS microscopy system to provide real-time differential diagnosis of lung lesions using quantitative measurements taken from the visualized cellular features and patterns. To the best of our knowledge, this is the first label-free and knowledge-based differential diagnostic platform to discriminate cancer from normal tissue or benign lesions, as well as cancer subtypes.

2 Materials and Methods

2.1 *Tissue Preparation and Imaging*

Lung tissues were obtained from patients undergoing surgical biopsy and surgery at The Methodist Hospital (TMH), Houston, Texas. Upon removal, the samples were snap-frozen in liquid nitrogen for storage. They were passively thawed at room temperature and kept moist with phosphate buffered saline before imaging. A total of 75 cases were acquired from TMH, including 19 normal cases, 20 adenocarcinoma cases, 25 squamous cell carcinoma cases, 3 small cell carcinoma cases, 6 organizing pneumonia cases, and 2 interstitial fibrosis cases. Seventeen additional frozen samples were purchased from the Cooperative Human Tissue Network, including 2 small cell carcinoma cases and 15 interstitial fibrosis cases. Because resection is usually not clinically indicated, small cell carcinoma cases are only infrequently made available for scientific research, resulting in a lower number of this type of lesion. Tissue samples were imaged on a glass slide using the CARS microscope, *ex vivo*.

The schematic of the setup was previously described.[20] The optical source system is composed of an optical parametric oscillator (OPO) and an Nd:YVO4 laser. The Nd:YVO4 laser delivers 7 ps, 76-MHz pulse trains at both 532 and 1064 nm wavelengths. The Stokes wave is 1064 nm, while 532 nm is used to pump the OPO, which generates a tunable 5 ps output from 670 to 980 nm. A pump wavelength of 816.8 was used. A bandpass filter (hq660/40m-2P, 25 mm diameter, Chroma, Inc.) is placed before the detectors to collect CARS signals and block unwanted backgrounds. Three to four sampling points were imaged for each specimen, and a total of 338 sampling points were examined. At each sampling point, three images were acquired from different imaging depths, resulting in a total of 1014 images. The beating frequency was tuned to 2845 cm^{-1} to probe the CARS signals that originated from symmetric CH_2 stretching bonds. After CARS imaging, all specimens were marked to indicate the sampled locations, sectioned through marked locations, and finally stained with hematoxylin and eosin (H&E). Bright-field images of these H&E slides were captured and examined to determine the type of lesion as a standard control.

2.2 *Data Analysis*

While the front-end CARS microscopy system acquires the initial images, the back-end knowledge-based classification

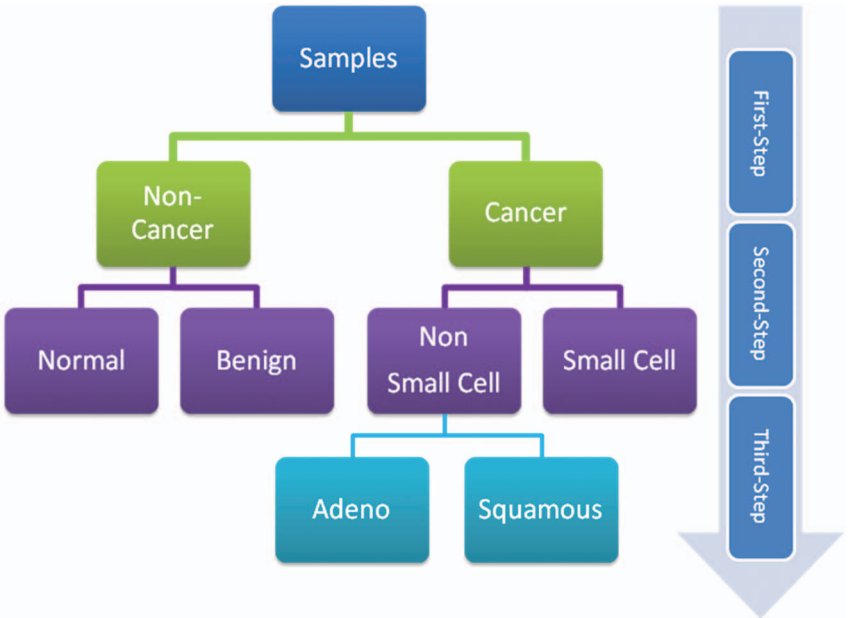

Fig. 1 Overview of the three-step differential process.

module, consisting of nuclei segmentation, feature extraction, and classification analysis functions, is built on identifying cellular and fibril structural features in order to separate different kinds of lesions. The goal is to classify tissue samples into individual subtypes through a three-level process (Fig. 1), which simulates clinical diagnostic workup. In the first level, a lesion is identified as cancerous or noncancerous (normal and benign). The cancerous group includes all three subtypes of lung cancers (adenocarcinoma, squamous cell carcinoma, and small cell carcinoma), while the benign group includes organizing pneumonia and interstitial fibrosis. A total of 145 cell/fibril features are calculated directly without segmentation from the CARS images. The same features were further used for the separation of normal and benign cases in part of the second level. In a clinical setting, the practitioner must initially characterize a lung nodule lesion as normal, cancerous, or benign, and these two steps mimic this diagnostic process. The other part of the second level and the entire third level of our scheme focuses on separation of subtypes of cancers, which includes segmentation of the cell nucleus and measurement of pathologically related features. Specifically, after segmentation, we use parameters, such as cell volume, nuclear size, and cell-cell distance, to measure a total of 35 features, and thereby separate cancer subtypes.

2.2.1 *Separation of cancerous, benign, and normal samples*

Fibril and cell structures can be used to separate cancerous from benign and normal samples because changes in these structures are closely related to different types of lung lesions, including cancer, pneumonia, and interstitial fibrosis.[18] Therefore, in the first step of the differential design, we extracted a set of 145 informative features and built a classifier to characterize fibril and cell features in order to separate fibril-dominant normal and benign lesions from cell-dominant cancerous lesions. Although both normal and benign lesions are fibril-dominant, they appear differently with regard to the orientation and/or distortion of the fibrils in our CARS images (Fig. 2). The same set of quantitative features was further used to train a classifier for separation of normal tissue from benign lesions. This set of 145 features consists of three feature categories widely used in image-based retrieval and pattern recognition: 85 wavelet features, 13 Haralick co-occurrence features,[21,22] and 47 Zernike moment features (the first and second out of 49 Zernike moment features are discarded due to they have the same values for all the images).[23] The wavelet features come from two important wavelet techniques: 70 Gabor wavelet features[24] and 15 Cohen–Daubechies–Feauveau wavelet (CDF9/7) features.[25]

2.2.2 *Separation of subtypes of cancers*

Segmentation of cancer cells. In contrast with noncancerous groups, cancerous samples show a high density of cancer cells whose nuclei can be identified by CARS because of their low CH_2 level (Fig. 2). Moreover, cellular details evident in CARS images enable us to measure additional morphological characteristics utilized by pathologists to identify different subtypes of cancer, including nuclear size, cell volume, and cell-cell distance, which correspond to such pathological criteria as pleomorphism and nucleus-to-cytoplasm (N/C) ratio. To perform measurements of these features, segmentation of cell nucleus is an essential step. Since CARS images bear a low level of contrast and a high level of noise, it is often not compatible with fully automatic detection approaches in identifying nuclear boundaries with a high degree of accuracy. As a result, a semi-automatic segmentation algorithm, which was fast enough to obtain segmentation results within minutes, was employed to precisely delineate the boundaries of cell nuclei. The algorithm consists of one manual step and four automatic steps to obtain an accurate nuclear boundary, as described in Ref. 19. In addition,

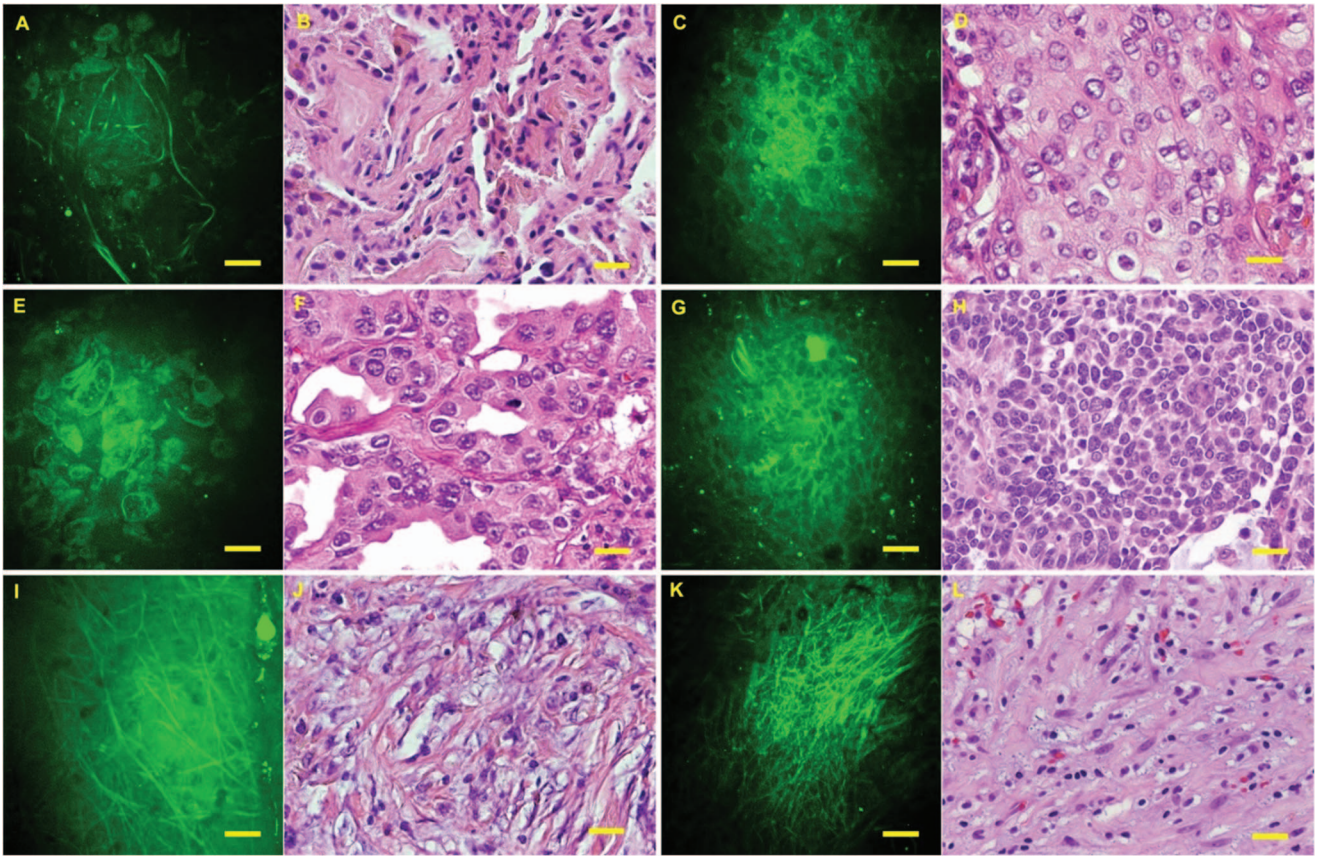

Fig. 2 *Ex vivo* images of human lung lesions. CARS and H&E images of (a) and (b) normal lung, (c) and (d) squamous cell carcinoma, (e) and (f) adenocarcinoma, (g) and (h) small cell carcinoma, (i) and (j) organizing pneumonia, and (k) and (l) organizing pneumonia derived from the same patient, respectively. Scale bars: 50 μm.

a manual ellipse fitting algorithm was developed to segment a rare fraction of cell nuclei that could not be well processed using the semi-automatic approach. In this algorithm, the user needs to select four points on the boundaries of the cell nucleus in order to generate accurate cell fitting.

Design of informative cellular features. Following nuclear segmentation, seven cellular features were designed and calculated to capture cellular signatures of cancer subtypes. These features include size of the nucleus, length of major and minor axes of the nucleus, area of Voronoi Tessellations[26–28] [Fig. 3(a)], as well as the maximum, minimum, and average neighbor distance of a cell based on the Delaunay Triangulation graph[29] [Fig. 3(b)]. These features describe both the attributes of individual cells and their relative spatial distribution. However, because of the diversity among different cells within each CARS image, the measurement of each feature resulted in producing a unique distribution.[19] Therefore, we made use of five additional parameters to describe each distribution type (i.e., mean value, standard deviation, skewness, kurtosis, and entropy), resulting in a grand total of 35 features.

2.2.3 *Differential diagnostic analysis*

Having extracted two sets of quantitative features (145/35), we could finally perform differential diagnostic analysis. There are two classification algorithms used in this paper. Partial least square regression (PLSR)[30,31] and support vector machine (SVM) with recursive feature elimination (RFE) (SVM-RFE).[32–34] Specifically, PLSR analysis[30] provides a global view of the distribution of different types of lesions by mapping the original feature space into a new space in which the predicted and investigated variables are maximally correlated.[30,31] This results in an optimal visual separation of samples in a three-dimensional (3D) space. Since the main advantage of PLSR lies in regression analysis, rather than classification analysis,[30] we further employed the SVM-RFE approach[32–34] for differential analysis and the investigation on changes of classification accuracy with different number of features through a feature selection process in order to overcome possible overfitting problems.

To validate the classification algorithm, a leave-one (patient)-out cross-validation analysis was used. In this step, experimental data from one of the patients were used for testing, while the remaining patient data were used to train the SVM classifier. The training and testing datasets were randomly selected and repeated 100 times to test the accuracies of classification. Since three images were acquired for each sampling point, the voting strategy, which determines the patient's lesion type according to the classification results of the majority of the three images, was used to adjust conflicting results among individual images.

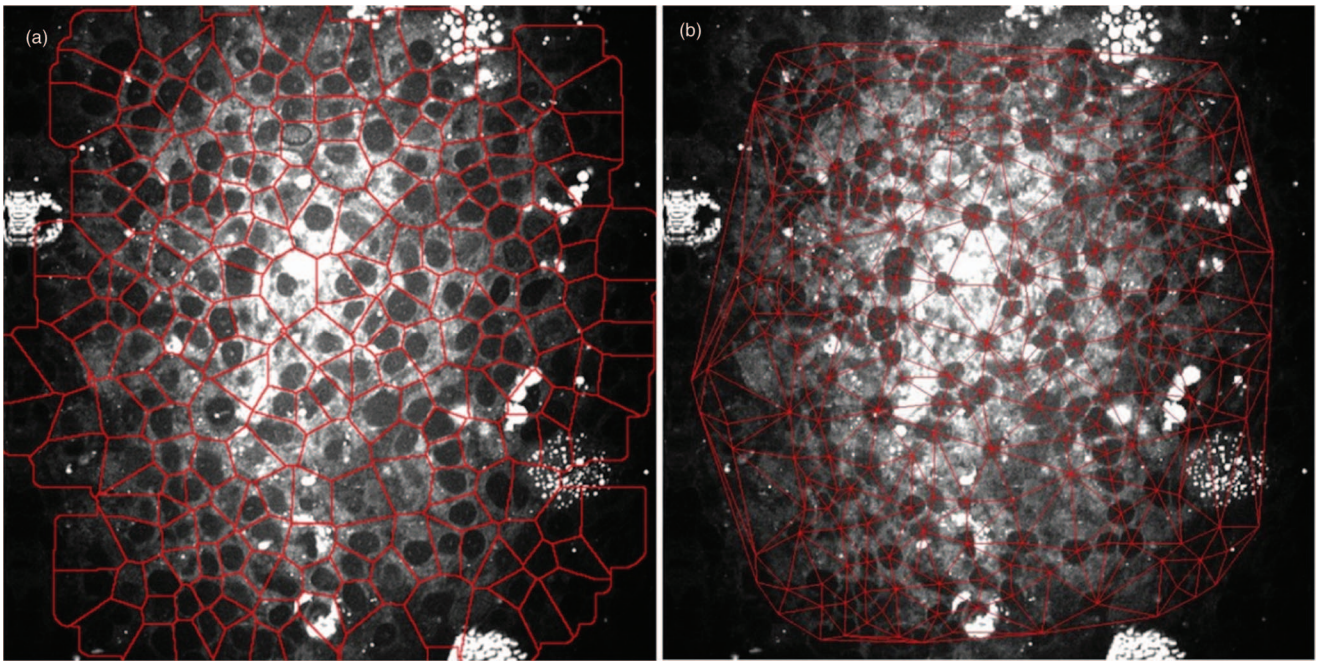

Fig. 3 Representative results of (a) Voronoi Tessellation and (b) Delaunay Triangulation on a small cell carcinoma image.

3 Results and Discussion

3.1 *Label-Free Molecular Vibrational Imaging of Different Types of Lung Lesions*

Figure 2 shows our representative CARS images and corresponding H&E results of normal, cancer and noncancer lesions. Tissue structures were clearly identified on the cellular level. The normal lung is predominantly composed of well-organized fibrous structures, consisting of the bronchi and supporting matrix for alveoli [Figs. 2(a) and 2(b)]. Cancer regions showed much denser cellularity compared with normal regions, and the size and configuration of the cells corresponded with these parameters, as shown by H&E stain [Figs. 2(c)–2(h)]. Commonly used pathological features were also identified for individual subtypes of cancers, including large polygonal malignant cells in sheets with abundant dense cytoplasm for squamous cell carcinoma, nested large round cells with abundant inhomogeneous cytoplasm for adenocarcinoma, and round or oval cells with minimal cytoplasm (high N/C ratio), and nuclear molding for small cell carcinoma. Meanwhile, organizing pneumonia and interstitial fibrosis [Fig. 2(i)–2(l)], two types of noncancerous lesions served as controls, showed dense fibrous structures with distortion of the normal lung architecture similar to those shown by the corresponding H&E stains. The lack of cellularity in the images of benign cases, as compared to both normal and cancerous cases, could be explained by the predominant signals from the fibrous tissue, as well as extracellular matrix.

3.2 *Differential Diagnostic Analysis*

3.2.1 *Separation of cancer from noncancer*

As indicated in Figure 2, normal and benign tissues possess clear fibrous structures, while cancer tissues possess high-density cellular features without obvious fibril formation. Using the 145 features, these fibrous and cellular signatures were numerically characterized, enabling the separation of normal, cancer and benign cases. Figure 4(a) illustrates the 3D spatial distribution of normal, benign, and cancer samples using PLSR analysis, in which all three groups are visually separated. The SVM-RFE approach was further employed to optimize classification accuracy and identify optimal feature combinations. Figure 4(b) shows the classification accuracy when different feature combinations (using 1 up to all 145 features) are used to separate cancer from noncancer cases. The results indicate that the accuracy reaches a stable peak level with 11 to 16 features. Figure 5(a) illustrates the classification accuracy with an optimal 11-feature set. Over 92% and 91% of samples from cancer and non-cancer tissues are correctly classified. Again, Haralick co-occurrence texture features show their importance in this separation step because all members (sum variance, sum of squares, contrast, difference variance, and sum average) from the top 5-feature subset after SVM-RFE belong to this category, which again demonstrates its superior ability to describe fibrils and cell structures for building a classifier.

Although all cancer samples can be visually well separated from non-cancer samples using their distinctive cellular features, the developed semi-automated strategy still shows difficulties in precisely extracting fibrous and cellular features to reach 100% accuracy. In our 145 features-based classification strategy, membranes around cell nuclei in cancer lesions were sometimes considered as fibers, while the dark holes between the fibers in the noncancer cases were confused with cell nuclei. Still, as an ancillary tool for clinical diagnosis, current accuracies are good enough to produce reliable results.

3.2.2 *Separation of normal from benign tissues*

As shown in Figure 2, normal tissues possess moderate fibrous structures with clear orientation while benign cases have predominant fibrous structures with distortions. Figure 4(c) shows

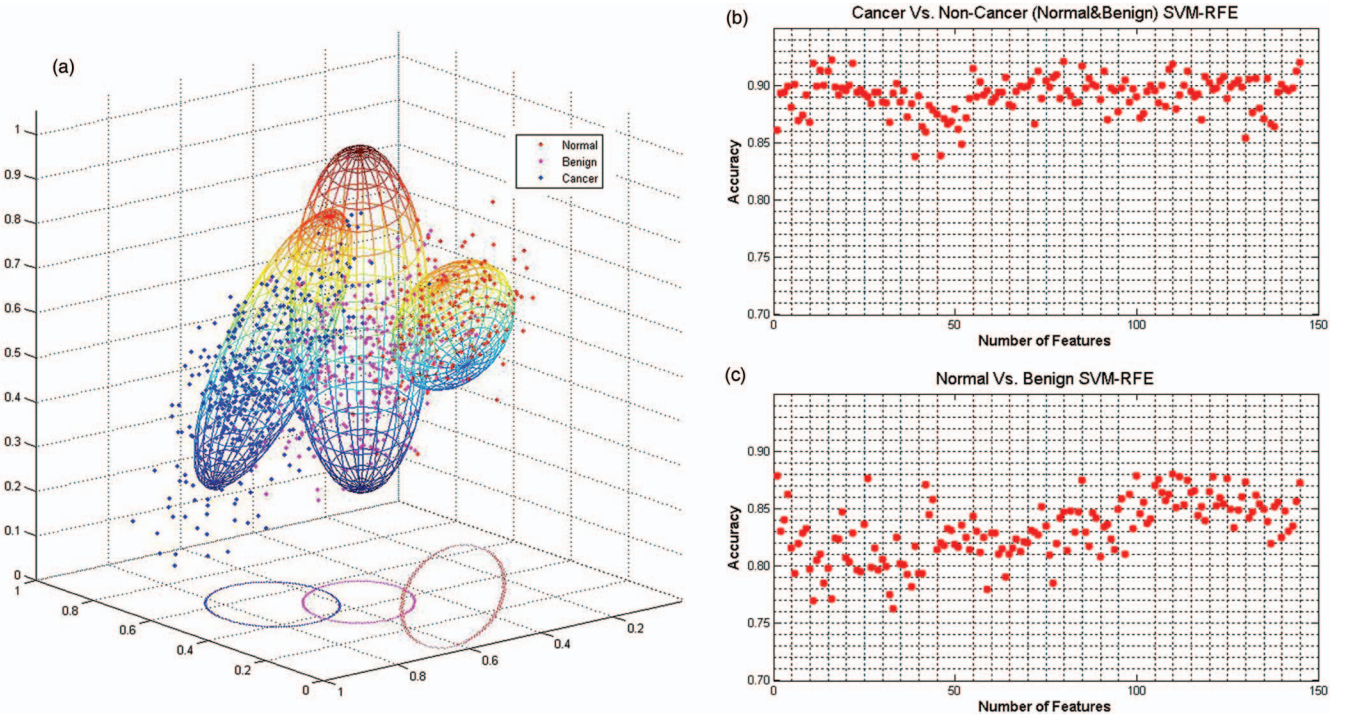

Fig. 4 (a) Spatial distributions of cancer, normal and benign groups using PLSR analysis. The top three-scored vectors were first calculated through PLSR analysis on all lung samples such that each sample could be represented by this three-dimensional vector with a scale normalized from 0 to 1. All lung samples were then plotted in a 3D space of these three vectors for separation. To enhance visualization, three ellipsoids were further fitted to the three subgroups with the ellipsoids projected onto the first and second components (in the PLSR analysis). (b) and (c) Classification analyses through SVM-RFE of cancer versus noncancer samples and normal versus benign samples on the 145 feature set, respectively.

A	Cancer	Non-Cancer
Cancer	91.24	8.76
Non-Cancer	7.47	92.53

B	Benign	Normal
Benign	92.5	7.5
Normal	11	89

C	Small Cell Carcinoma	Non-Small Cell Carcinoma
Small Cell Carcinoma	100	0
Non-Small Cell Carcinoma	0	100

D	Squamous Cell Carcinoma	Adenocarcinoma
Squamous Cell Carcinoma	83.57	16.43
Adenocarcinoma	23.22	76.78

Fig. 5 Classification accuracies of separating (a) cancers from noncancers using the top 11 features through SVM-RFE on the 145 feature set; (b) benign from normal cases using the top 110 features through SVM-RFE on the 145 feature set. (c) small cell from nonsmall cell carcinoma using the top 1 feature obtained through SVM-RFE on the 35 feature set. (d) Squamous cell carcinoma from adenocarcinoma using the top 25 features obtained through SVM-RFE on the 35 feature set.

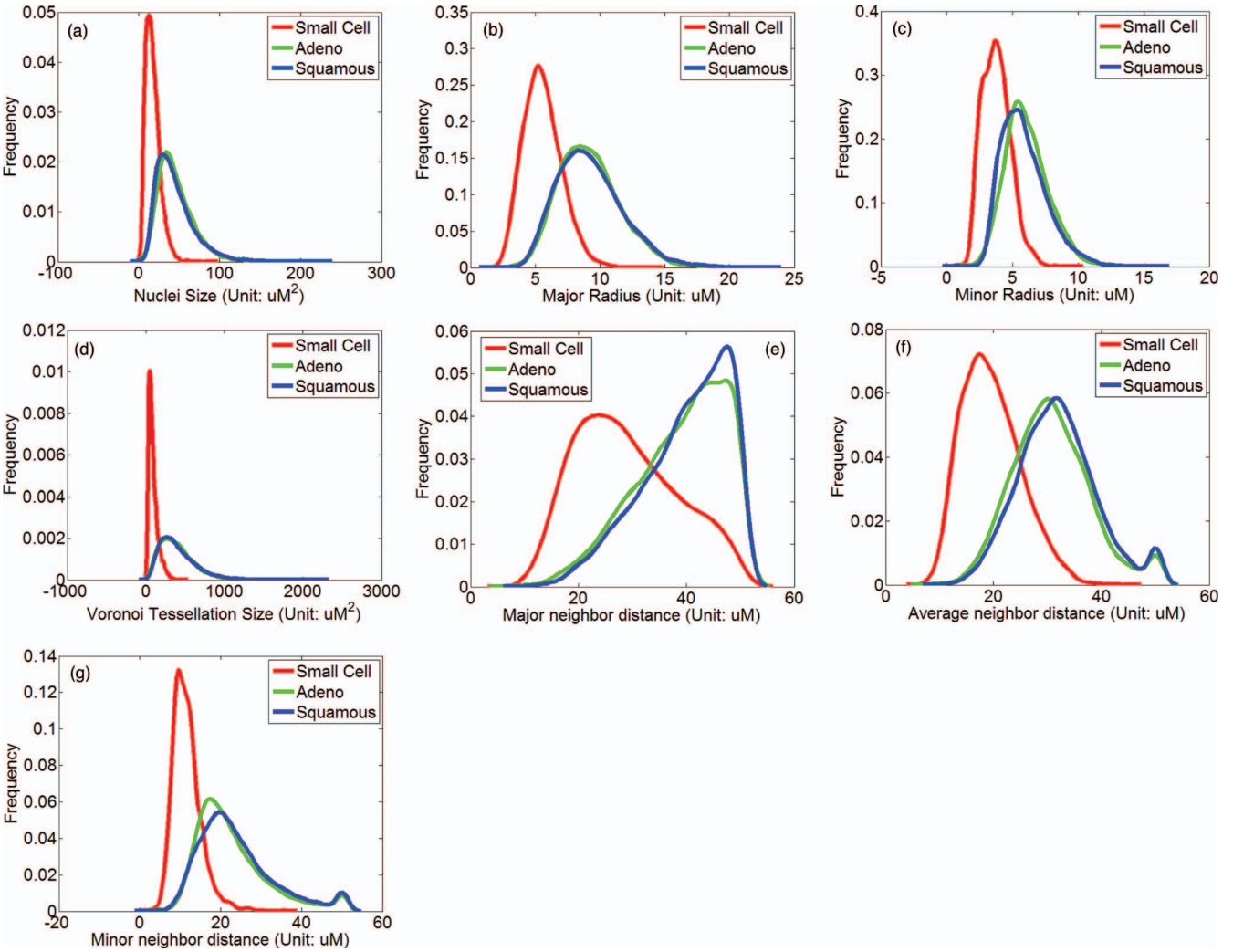

Fig. 6 Distributions of the seven features of three cancer subgroups: small cell carcinoma (red), adenocarcinoma (green), and squamous cell carcinoma (blue). We randomly chose 5000 cells from each subtype and investigated the distribution of their 7 features with respect to different subtypes. (Color online only.)

classification accuracies using different subsets of feature combinations through a SVM-RFE feature selection process, in which the accuracy reaches a stable level with the use of around 110 features. Using a 110-feature subset, over 92% and 89% accuracies were achieved for benign and normal samples, respectively [Fig. 5(b)]. To account for the percentage shortfall, we know that pathological changes in benign lesions are such that certain sampled locations possessed a lower level of abnormality, thus showing tissue structures similar to those of normal cases. However, in a real differential diagnostic process, current results are more than sufficient to achieve the main goal: delineating benign samples (over 92%) from the normal.

3.2.3 *Separation of subtypes of lung cancers*

Using the semi-automatic segmentation algorithm, we were able to delineate boundaries of an individual cell nucleus, enabling the measurements of 35 numerical features (Fig. 6). The spatial distribution of the three subtypes of cancers is illustrated in Fig. 7(a). All small cell cancer cases are well separated from nonsmall cell subtypes, while the latters overlap in a certain extent in distribution. SVM-RFE analysis showed that the classification accuracy between small cell and non-small cell cancers reaches 100%, even with only one feature [Fig. 5(c)], i.e., Voronoi Tessellation.

For separation of adenocarcinoma from squamous cell carcinoma, a subset of 25 features was chosen through feature selection. Classification based on these features showed mixed results with lower (75.5% and 71.5%) classification accuracies [Fig. 5(d)]. This overlap is not surprising and is in accordance with the clinical difficulty in differentiating these two subtypes using morphology alone.[35] Although separation of nonsmall cell carcinoma from small cell carcinoma has been traditionally adequate for clinical decision-making, this is no longer the case. Increasingly, definitive diagnosis of histologic subtype, often in conjunction with molecular tumor profiling, is needed.[36] In this regard, CARS has the potential to allow for real-time identification of nonsmall cell carcinomas, but tissue excision for additional work-up may still be necessary. Particularly, we have evaluated the time taken to classify each sample through the three-step process. We concluded that CARS imaging, together with computerized pattern recognition and classification,

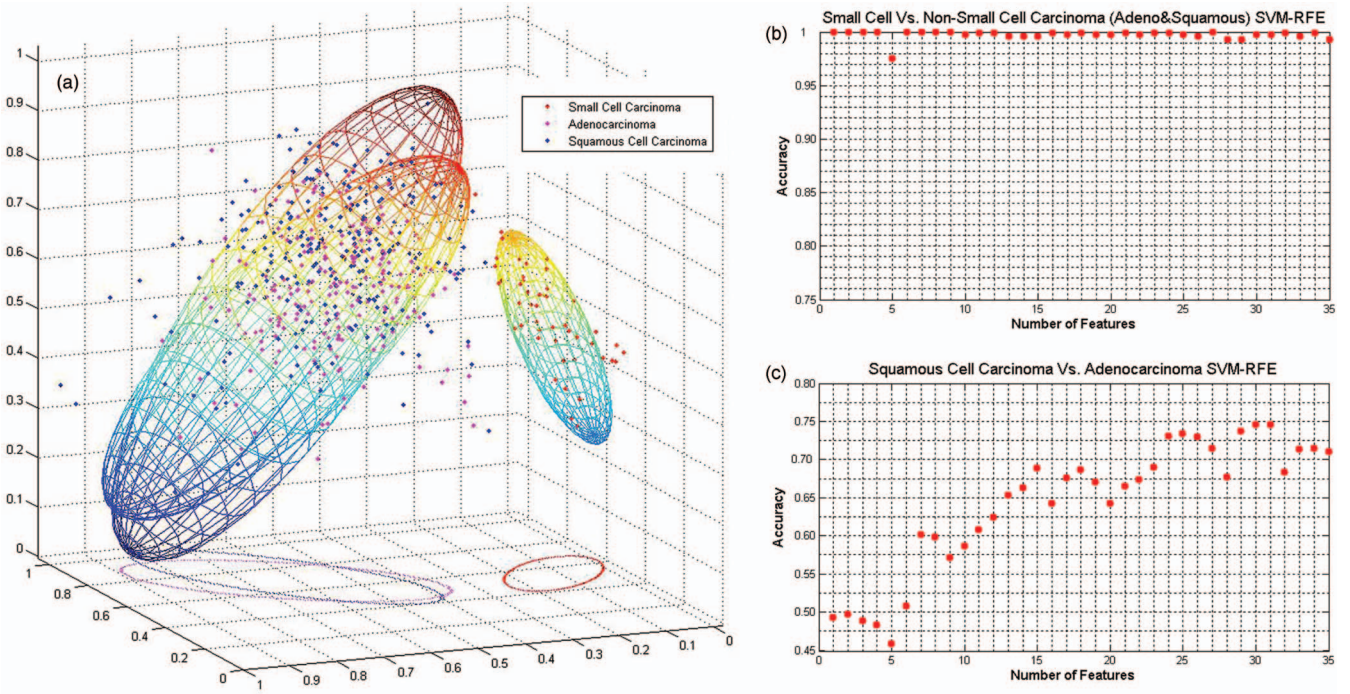

Fig. 7 (a) Spatial distribution of cancer subtypes using PLSR analysis. Similar to Fig. 4, the top three-scored vectors were calculated through PLSR analysis on all cancer samples to represent each cancer sample with a three-dimensional vector and a scale normalized from 0 to 1. Three ellipsoids were fitted in the same way for better visualization. (b) and (c) Classification analyses through SVM-RFE of small cell versus nonsmall cell carcinoma and squamous cell carcinoma versus adenocarcinoma on the 35 feature set, respectively.

only takes a mere few minutes to reach a final diagnosis in our study.

By looking at the five parameters used to describe the distribution of each feature, we found that mean and skewness play major roles in this 25-feature subset. In comparison, the kurtosis parameters, which measure "peakedness" of these distributions, are excluded, falling into the last 10 VIPs. Since all pictures were acquired on a two-dimensional scale through optical sectioning, a normally distributed background noise could be introduced to weaken any significant peak in a given distribution. For example, the same cell nucleus would be measured as different sizes (potentially from zero to the real size) from different imaging depths. As a result, the peak of cell nuclei will be less significant, even if the real size is quite uniform. Therefore, the difference in kurtosis will be reduced between different cancer subtypes, lowering the importance of this parameter. For the same reason, standard deviation and entropy may be weakened as well, while the mean and skewness are less likely to be affected in reflecting the difference between subtypes. One possible solution to avoid these artificial effects is to conduct measurements on 3D reconstructed data, which will better reflect the real size and distance between cells and lead to potential improvement of the accuracy in separating non-small cell subtypes.

3.3 *Future Work*

Screening for early cancer has attracted much attention since it could potentially increase survival rate. After initial screening, our novel CARS technique combines real-time and label-free imaging with cell feature classification to potentially facilitate biopsy yield, differential diagnosis, and subsequent treatment in a manner suggested by this report. Ongoing screening studies are using high-resolution CT to detect early stage lung cancers, but have not caused a decreased incidence of advanced lung cancers and the results on lung cancer mortality have not yet been finalized.[37–41] Though it is not yet clear whether screening can improve survival, we believe the increased biopsy yield can definitely reduce medical costs and patients suffering in addition to facilitating on-the-spot diagnosis.

An additional benefit of the CARS system involves the strong association between histologic cell type and subtypes and specific predictive biomarkers in terms of response to targeted molecular therapies for advanced stage lung cancer. Specifically, it has long been known that most lung cancers are histologically heterogeneous and that over 90% of adenocarcinomas have more than one histologic subtype. When treatment options were limited, this heterogeneity was not important, and it was only necessary to differentiate small cell from nonsmall cell carcinoma to select an appropriate treatment regimen. However, the advent of molecular targeted therapies makes identification of the various histologic types and subtypes within a given lung cancer more important. Therefore, another potential benefit of our strategy lies in its ability to differentiate cell types and subtypes within a heterogeneous lung cancer at the time of biopsy such that each of the different histologic types can be sampled sent for molecular analysis.

Although the current CARS-based diagnostic system has shown substantial efficacy, we acknowledge that the retrospective study pattern on patient selection, together with unequal sample sizes may have brought bias to the classification

accuracy, as reported in this study. To fully demonstrate this diagnostic strategy for clinical applications, a much larger patient group and a prospective study pattern is still needed for unbiased evaluation and further improvement of the platform.

4 Conclusions

In this paper, we introduced a new approach that integrates label-free, chemistry-sensitive CARS microscopy with advanced pattern recognition techniques to enable quantitative differentiation of human lung lesions and classification of lung cancer subtypes. We demonstrated the utility of the approach in differentiating among normal, benign, and malignant lung tissues, as well as different subtypes of cancerous tissue, in a manner that can be both visualized and quantified. Diagnostic features were chosen according to established pathological standards, enabling direct interpretation of the results. These excellent *ex vivo* results indicate the potential of the reported diagnostic system for the evaluation of fresh tumor specimens during intraoperative procedure or image-guided biopsy without waiting for pathological staining, this would result in accelerated diagnosis or improved clinical decision making. In addition, the demonstration of the strategy is a necessary step toward *in vivo* diagnosis of precancerous and cancerous lung lesions. The clinical potential is further strengthened by the efforts aimed at miniaturizing the CARS technique for fiber-based *in vivo* imaging.[42,43] In summary, the reported computerized and label-free imaging strategy could potentially improve and fundamentally change diagnostic approaches to early-stage lung cancer by offering an efficient way to characterize different types of lesions, enabling medical practitioners to obtain essential information in real time and, when coupled with fiber-based imaging when available, would reduce the need for excisional tissue biopsies while facilitating definitive treatment.

Acknowledgments

The funding of this research was initiated and supported by the Department of Systems Medicine and Engineering (SMAB) of The Methodist Hospital Research Institute (TMHRI), Weill Cornell Medical College and John S. Dunn Research Foundation to STCW. The authors would like to thank Dr. David P. Bernard and Pam McShane of the Department of Pathology for providing human tissue samples from the tissue bank of TMH, Dr. Zhong Xue, Dr. Kemi Cui, Dr. Xiaofeng Xia, Dr. Yong Mao, Dr. Hai Li, and Dr. Pengfei Luo from SMAB-TMHRI, Dr. Brian Saar, Dr. Gary Holtom, Dr. Wei Min, and Dr. Sunney Xie of Harvard University, as well as Wei Wang of Thorlabs, Inc. for helpful suggestions.

References

1. H. Hashizume, P. Baluk, S. Morikawa, J. W. McLean, G. Thurston, S. Roberge, R. K. Jain, and D. M. McDonald, "Openings between defective endothelial cells explain tumor vessel leakiness," *Am. J. Pathol.* **156**(4), 1363–1380 (2000).
2. D. R. Youlden, S. M. Cramb, and P. D. Baade, "The international epidemiology of lung cancer: geographical distribution and secular trends," *J. Thorac. Oncol.* **3**(8), 819–831 (2008).
3. D. M. Parkin, F. Bray, J. Ferlay, and P. Pisani, "Global cancer statistics, 2002," *Ca-Cancer J. Clin.* **55**(2), 74–108 (2005).
4. S. Diederich, "Lung cancer screening: status in 2007," *Radiologe* **48**(1), 39–44 (2008).
5. C. I. Henschke, D. F. Yankelevitz, D. M. Libby, M. W. Pasmantier, J. P. Smith, and O. S. Miettinen, "Survival of patients with stage I lung cancer detected on CT screening," *N. Engl. J. Med.* **355**(17), 1763–1771 (2006).
6. A. McWilliams, C. MacAulay, A. F. Gazdar, and S. Lam, "Innovative molecular and imaging approaches for the detection of lung cancer and its precursor lesions," *Oncogene* **21**(45), 6949–6959 (2002).
7. J. Hung, S. Lam, J. C. LeRiche, and B. Palcic, "Autofluorescence of normal and malignant bronchial tissue," *Lasers Surg. Med.* **11**(2), 99–105 (1991).
8. T. Gabrecht, T. Glanzmann, L. Freitag, B. C. Weber, H. van den Bergh, and G. Wagnieres, "Optimized autofluorescence bronchoscopy using additional backscattered red light," *J. Biomed. Opt.* **12**(6), 064016 (2007).
9. F. Stanzel, "Fluorescent bronchoscopy: contribution for lung cancer screening?," *Lung Cancer* **45** Suppl 2, S29–S37 (2004).
10. T. Xie, G. Liu, K. Kreuter, S. Mahon, H. Colt, D. Mukai, G. M. Peavy, Z. Chen, and M. Brenner, "In vivo three-dimensional imaging of normal tissue and tumors in the rabbit pleural cavity using endoscopic swept source optical coherence tomography with thoracoscopic guidance," *J. Biomed. Opt.* **14**(6), 064045 (2009).
11. J. Lademann, J. Shevtsova, A. Patzelt, H. Richter, N. D. Gladkowa, V. M. Gelikonov, S. A. Gonchukov, W. Sterry, A. M. Sergeev, and U. Blume-Peytavi, "Optical coherent tomography for in vivo determination of changes in hair cross section and diameter during treatment with glucocorticosteroids–a simple method to screen for doping substances?," *Skin Pharmacol. Appl. Skin Physiol.* **21**(6), 312–317 (2008).
12. M. D. Duncan, J. Reintjes, and T. J. Manuccia, "Scanning coherent anti-Stokes Raman microscope," *Opt. Lett.* **7**(8), 350–352 (1982).
13. C. L. Evans and X. S. Xie, "Coherent Anti-Stokes Raman Scattering Microscopy: Chemical Imaging for Biology and Medicine," *Annu. Rev. Anal. Chem.* **1**, 883–909 (2008).
14. J.-X. Cheng and X. S. Xie, "Coherent anti-Stokes Raman scattering microscopy: instrumentation, theory, and applications," *J. Phys. Chem. B* **108**(3), 827–840 (2004).
15. C. L. Evans, E. O. Potma, and X. S. Xie, "Coherent anti-stokes raman scattering spectral interferometry: determination of the real and imaginary components of nonlinear susceptibility chi(3) for vibrational microscopy," *Opt. Lett.* **29**(24), 2923–2925 (2004).
16. C. L. Evans, E. O. Potma, M. Puoris'haag, D. Cote, C. P. Lin, and X. S. Xie, "Chemical imaging of tissue in vivo with video-rate coherent anti-Stokes Raman scattering microscopy," *Proc. Natl. Acad. Sci. U.S.A.* **102**(46), 16807–16812 (2005).
17. P. D. Chowdary, Z. Jiang, E. J. Chaney, W. A. Benalcazar, D. L. Marks, M. Gruebele, and S. A. Boppart, "Molecular histopathology by spectrally reconstructed nonlinear interferometric vibrational imaging," *Cancer Res.* **70**(23), 9562–9569 (2010).
18. V. Kumar, A. K. Abbas, N. Fausto, and J. Aster, Eds., *Pathologic Basis of Disease*, 8th Ed., Saunders, Philadelphia (2009).
19. L. Gao, H. Zhou, M. J. Thrall, F. Li, Y. Yang, Z. Wang, P. Luo, K. K. Wong, G. S. Palapattu, and S. T. Wong, "Label-free high-resolution imaging of prostate glands and cavernous nerves using coherent anti-Stokes Raman scattering microscopy," *Biomed. Opt. Express* **2**(4), 915–926 (2011).
20. L. Gao, Y. Yang, J. Xing, M. Thrall, Z. Wang, P. Luo, K. Wong, H. Zhao, and S. Wong, "Diagnosing lung cancer using coherent anti-Stokes Raman scattering microscopy," *Proc. SPIE* 7890 (2011).
21. R. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE* **67**, 786–804 (1979).
22. Haralick texture features: http://murphylab.web.cmu.edu/services/SLF/features.html.
23. F. Zernike, "Beugungstheorie des schneidencerfarhens undseiner verbesserten form, der phasenkontrastmethode," *Physica* **1**, 689–704 (1934).
24. B. S. Manjunatha and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 837–842 (1996).
25. A. Cohen, I. Daubechies, and J. C. Feauveau, "Bi-orthogonal bases of compactly supported wavelets," *Commun. Pure Appl. Math.* **45**, 485–560 (1992).

26. T. Jones, A. Carpener, and P. Golland, "Voronoi-based segmentation of cells on image manifolds," *Lect. Notes Comput. Sci.* **3765**, 535–543 (2005).
27. C. C. Bilgin, P. Bullough, G. E. Plopper, and B. Yener, "ECM-aware cell-graph mining for bone tissue modeling and classification," *Data Min. Knowl. Discov.* **20**(3), 416–438 (2010).
28. F. Li, X. Zhou, and T. C. S. Wong, "Optimal Live Cell Tracking for Cell Cycle Study Using Time-lapse Fluorescent Microscopy Images" in *International Workshop on Machine Learning in Medical Imaging (MLMI 2010)*, pp. 124–131, Springer Lecture Notes in Computer Science, Beijing, China (2010).
29. F. Li, X. Zhou, J. Ma, and S. T. Wong, "Multiple nuclei tracking using integer programming for quantitative cancer cell cycle analysis," *IEEE Trans. Med. Imaging* **29**(1), 96–105 (2010).
30. P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Anal. Chim. Acta* **185**, 1–17 (1986).
31. H. Abdi, "Partial least squares (PLS) regression," *Encyclopedia of Social Sciences, Research Methods*, Sage, Thousand Oaks, California (2003).
32. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.* **2**(2), 1–47 (1998).
33. V. Vapnik, *Statitical Learning Theory*, Wiley Interscience (1998).
34. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning* **46**(1), 389–422 (2002)
35. J. Terry, S. Leung, J. Laskin, K. O. Leslie, A. M. Gown, and D. N. Ionescu, "Optimal immunohistochemical markers for distinguishing lung adenocarcinomas from squamous cell carcinomas in small tumor samples," *Am. J. Surg. Pathol.* **34**(12), 1805–1811 (2010).
36. C. J. Langer, B. Besse, A. Gualberto, E. Brambilla, and J. C. Soria, "The evolving role of histology in the management of advanced non-small-cell lung cancer," *J. Clin. Oncol.* **28**(36), 5311–5320 (2010).
37. D. R. Aberle, C. D. Berg, W. C. Black, T. R. Church, R. M. Fagerstrom, B. Galen, I. F. Gareen, C. Gatsonis, J. Goldin, J. K. Gohagan, B. Hillman, C. Jaffe, B. S. Kramer, D. Lynch, P. M. Marcus, M. Schnall, D. C. Sullivan, D. Sullivan, and C. J. Zylak, "The National Lung Screening Trial: overview and study design," *Radiology* **258**(1), 243–253 (2011).
38. E. Warner, A. Jotkowitz, and N. Maimon, "Lung cancer screening–are we there yet?," *Eur. J. Intern. Med.* **21**(1), 6–11 (2010).
39. L. R. Chirieac and D. B. Flieder, "High-resolution computed tomography screening for lung cancer: unexpected findings and new controversies regarding adenocarcinogenesis," *Arch. Pathol. Lab Med.* **134**(1), 41–48 (2010).
40. P. B. Bach, "Is our natural-history model of lung cancer wrong?," *Lancet Oncol.* **9**(7), 693–697 (2008).
41. P. B. Bach, J. R. Jett, U. Pastorino, M. S. Tockman, S. J. Swensen, and C. B. Begg, "Computed tomography screening and lung cancer outcomes," *JAMA, J. Am. Med. Assoc.* **297**(9), 953–961 (2007).
42. Z. Wang, Y. Yang, P. Luo, L. Gao, K. K. Wong, and S. T. Wong, "Delivery of picosecond lasers in multimode fibers for coherent anti-Stokes Raman scattering imaging," *Opt. Express* **18**(12), 13017–13028 (2010).
43. F. Legare, C. L. Evans, F. Ganikhanov, and X. S. Xie, "Towards CARS endoscopy," *Opt. Express* **14**(10), 4427–4432 (2006).