DOI 10.1007/s10812-015-0128-6
Journal of Applied Spectroscopy, Vol. 82, No. 3, July, 2015 (Russian Original Vol. 82, No. 3, May–June, 2015)

DISCRIMINATION OF BREAST CANCER FROM NORMAL TISSUE WITH RAMAN SPECTROSCOPY AND CHEMOMETRICS

Q.-B. Li,[a] W. Wang,[a] Ch.-H. Liu,[b] and G.-J. Zhang[a,*]

UDC 535.375.5:616-006.6

Conventional Raman spectra of normal and cancerous breast tissues were acquired at an excitation wavelength of 785 nm and subjected to a discrimination analysis. First the spectra were pretreated with wavelet transform and polynomial fitting; next, cancerous tissue was identified by applying an adaptive local hyperplane K-nearest neighbor (ALHK) method to the pretreated spectra. The best discrimination accuracy of the ALHK method was 93.2%. In summary, normal and cancerous breast tissue were accurately distinguished by a miniature laser Raman spectrometer and the chemometrics method (ALHK), which might prove to be a portable and accessible diagnostic system.

Keywords: *breast cancer, miniature Raman spectrometer, adaptive local hyperplane K-nearest neighbor (ALHK).*

Introduction. Breast cancer is among the major causes of female mortality. In 2012, about 63,300 cases of breast carcinoma *in situ* were newly diagnosed in the United States. Breast cancer presumably accounts for 14% of all female cancer deaths in that country, second only to lung cancer [1]. In China, the incidence of breast cancer has also increased significantly in recent years. In some large cities, such as Beijing, Shanghai, and Tianjin, breast cancer is the top-ranking malignant tumor in women (in terms of incidence) [2].

Early cancer diagnosis is crucial for implementing timely, effective, and ultimately successful treatments. As a form of molecular spectroscopy, Raman spectroscopy can detect cancer-induced changes in the molecular structure and composition of breast tissue. Before the appearance of clinical symptoms, cancer alters the structure and concentration of the main biomolecules constituting the cells and tissues. Therefore, molecular spectroscopy is a potential tool for early tumor diagnosis [3–7].

Raman spectroscopy has only recently emerged as a diagnostic technology for breast cancers. Thus far, diagnoses have been made by Fourier transform Raman spectroscopy (FTRS), confocal Raman microspectroscopy (CRS), resonance Raman spectroscopy (RRS), surface-enhanced Raman spectroscopy (SERS), and conventional Raman spectroscopy (RS). The SERS technology cannot easily detect human cancer *in vivo*, as the samples must be attached to a SERS-active substrate, which is cumbersome to operate [8, 9]. FTRS, CRS, and RRS allow reduced fluorescence and higher resolution of the Raman spectra and have been extensively investigated as breast cancer diagnosis tools [10–14]. However, these technologies generally require a large Raman spectrometer or a large desktop microscope, increasing the expense and reducing the portability of clinical diagnosis. In contrast, conventional Raman spectrometers are small, portable and low-cost. Combined with an optical fiber probe, these spectrometers hold much promise for *in vivo* and *in situ* cancer detection.

The disadvantages of miniature Raman spectrometers are strong fluorescence background interference and low spectral signal-to-noise ratio, both of which lower the discrimination accuracy of common data analysis methods. Few studies have considered the miniature Raman spectrometer as an RS tool for detecting breast cancer [15–17]. Therefore, if one could identify a discrimination analysis method with high prediction accuracy, the usefulness of the miniature Raman spectrometers for this purpose would be greatly enhanced. In this paper, the conventional Raman spectra of breast tissues were acquired by

*To whom correspondence should be addressed.

[a]School of Instrumentation Science and Opto-Electronics Engineering, Precision Opto-Mechatronics Technology Key Laboratory of Education Ministry, Beihang University, Xueyuan Road No. 37, Haidian District, Beijing, 100191, China; e-mail: qblee@126.com; [b]Institute for Ultrafast Spectroscopy and Lasers, The Department of Physics of the City College of the City University of New York, New York, USA. Published in Zhurnal Prikladnoi Spektroskopii, Vol. 82, No. 3, pp. 441–446, May–June, 2015. Original article submitted July 10, 2014.

 0021-9037/15/8203-0450 ©2015 Springer Science+Business Media New York

a miniature laser Raman spectrometer with the excitation wavelength of 785 nm. Prior to analysis, the noise and fluorescence background were eliminated by wavelet transform and polynomial fitting, respectively. Finally, cancerous and normal breast tissues were separated by a new classification algorithm called adaptive local hyperplane *K*-nearest neighbor (ALHK), a variant of the adaptive local hyperplane (ALH) algorithm, applied to the preprocessed spectra. The present study successfully detected cancer by a miniature Raman spectrometer, promoting the development of a portable clinical diagnostic technology.

Materials and Methods. *Tissue specimens.* Normal and malignant samples of human breast tissue were obtained from the National Disease Research Interchange (NDRI) and the Cooperation Human Tissue Network (CHTN). The cancerous tissues exhibited various stages of disease. All cancer tissue specimens were invasive ductal carcinoma (IDC), but two of them were ductal carcinoma *in situ* (DCIS). Most of the tissues were sourced from female patients aged 32 to 71 years (median age 55 years). One normal tissue specimen was taken from a female aged 16 years.

The tissue specimens were not chemically treated prior to spectroscopic analysis. They were maintained in liquid nitrogen before being packed in dry ice and shipped. All tissues arrived still frozen on dry ice, uncut and irregularly shaped. When required for spectroscopic study, they were removed from storage at –80^{o}C and thawed to ambient room temperature.

Acquisition of NIR conventional Raman spectra of breast tissues. In total, 368 Raman spectra were acquired from 11 tissue samples (four normal, seven cancerous) by an R-2000 NIR-Raman spectrometer (America Ocean Optics Inc.), a miniature spectrometer that excites at 785 nm. This Raman system comprises a multimode solid-state diode laser with an output power of 500 mW at room temperature and an "all-in-one" fiber optic probe with a spectral resolution of 15 cm^{-1}. At the focal point in the tissue specimen, the laser power is 175 mW, and the excitation spot size is 0.5 mm. The Raman spectra were acquired in the backscattered direction with an integration time of 30 s. Three spectra were collected at each location and averaged to reduce the noise level.

Spectra preprocessing method. The spectra collected by the Ocean Optics R-2000 Raman spectrometer were noisy and contained a strong fluorescence background. First the noise was removed by wavelet transform; next, the fluorescence background was removed by fitting the smoothed spectra to a third-order polynomial function using Matlab R2011b software. The wavelet transform [18] is briefly described below.

The discrete wavelet transform is given by

$$f(t) = \sum_{k\in z} c_{J,k}\psi_{J,k}(t) + \sum_{j=1}^{J}\sum_{k\in z} d_{j,k}\psi_{j,k}(t) \ ,$$

where $\psi_{J,k}(t)$ is the wavelet basis function, $c_{J,k}$ is the Jth approximation coefficient of the spectral signal (denoting the low-frequency coefficient), and $d_{j,k}$ is the jth detail coefficient of the spectral signal (denoting the higher-frequency coefficients).

The wavelet transform basically projects the spectral signal in the wavelet basis function, decomposes the spectrum signal into its time and frequency components, and obtains the wavelet approximation and high-resolution coefficients. The high-resolution signals reflect the local nuances, but most of the very high-frequency components constitute noise. Thus, the wavelet transform is used to remove the noise from the Raman spectra and to optimize the spectral quality.

The spectral preprocessing is implemented as follows.

Step 1. Choose a wavelet function and a decomposition scale.

Step 2. Remove the high-frequency coefficients of the wavelet decomposition by threshold filtering. Here we adopt a soft threshold function [19]:

$$\hat{w}_{j,k} = \begin{cases} \text{sgn}\ (w_{j,k})(|w_{j,k}| - \lambda) & |w_{j,k}| \geq \lambda; \\ 0 & |w_{j,k}| < \lambda\ ; \end{cases} \qquad \lambda = \sigma * \sqrt{2\ \log\ (N)}.$$

Step 3. Reconstruct the spectrum signal from the lowest-frequency coefficient (the Jth coefficient) and the higher-frequency coefficients (1–j) that have passed the threshold processing.

Discrimination analysis method. The cancerous and normal tissue samples were classified by the ALHK classifier, a variant of the adaptive local hyperplane (ALH) that was recently proposed by Yang et al. [20, 21]. The ALHK classifier operates similarly to ALH but adopts a different neighborhood selection procedure. More specifically, it constructs the *K*-neighborhood as the set of *K* training samples in each class with the smallest spatial separation from the given query q.

The ALHK algorithm proceeds as follows.

Suppose that the training set contains L samples in J classes. Each training sample consists of d input features $x_i = (x_{i1}, ..., x_{id})^T$ with known class labels $y_i = c (i = 1, \ldots, L; c = 1, \ldots, J)$. The goal is to predict the class label of a query with input vector $q = (q_1, \ldots, q_d)^T$.

Step 1. Calculate the feature weight w of the training sample as follows:

$$r_j = \frac{\sum_i \sum_c I(y_i = c)(\bar{x}_{cj} - \bar{x}_j)^2}{\sum_i \sum_c I(y_i = c)(x_{ij} - \bar{x}_{cj})^2},$$

$$R_j = r_j / \max(r_j),$$

$$w_j = \exp(TR_j) \Big/ \sum_{j=1}^{d} \exp(TR_j), \qquad \forall j = 1, ..., d,$$

where $\bar{x}_j$ denotes the jth component of the grand class centroid and $\bar{x}_{cj}$ denotes the jth component of the centroid of class c. The indicator function $I(\bullet)$ equals 1 when $y_i = c$, and 0 otherwise. T is a positive parameter that controls the influence of R_j on w_j.

Step 2. Calculate the weighted Euclidean distance metric D between x_i and q:

$$D(x_i, q) = \sqrt{\sum_{j=1}^{d} w_j (x_{ij} - q_j)^2}.$$

Step 3. Based on the Euclidean distance D, select the K nearest neighbors of class c, $p_c = (p_{c1}, ..., p_{cK})$, for the given query q, then construct the local hyperplane of class c containing the p_c:

$$LH_c(q) = \{s \mid s = \sum_{i=1}^{K} \alpha_i V_{\cdot i} + m_c\},$$

$$m_c = \frac{1}{K} \sum_{i=1}^{K} p_{ci}, \qquad V_{\cdot i} = p_{ci} - m_c, \qquad \alpha = (\alpha_1, ..., \alpha_K)^T.$$

Step 4. Calculate the minimum distance between q and $LH_c(q)$:

$$J_c(q) = \min_{\alpha} \sum_{j=1}^{d} w_j (V_{j\cdot}\alpha + m_{cj} - q_j)^2 + \lambda \alpha^T \alpha = \min_{\alpha} (s-q)^T W (s-q) + \lambda \alpha^T \alpha,$$

$$W = \mathrm{diag}(w_1, ..., w_d),$$

where the regularization parameter λ prevents α from becoming too large. Solving the equation $\partial J_c(q)/\partial \alpha = 0$, we obtain $\alpha = (U^T V + \lambda I_{nc})/(U^T (q - m_c))J$, where $U^T = V^T W$.

Step 5. Assign a class label to q: label(q) = argmin$_c J_c(q)$.

Results and Discussion. *Conventional NIR laser Raman spectra.* Each Raman spectrum was labeled according to the pathological diagnosis of the tissue. The scan region of each spectrum was 700–1800 cm^{-1}. In this study, noise was reduced by a Symmlet-5 wavelet filter and a four-decomposition scale, and the fluorescent background was removed by a third-order polynomial. Typical Raman spectra of normal and cancerous tissues before and after preprocessing are shown in Fig. 1.

Although clear peaks appear in the raw spectrum of the normal tissue, the peaks in the spectrum of the cancerous tissues are obscured by noise and the fluorescent background (Fig. 1a). The quality of the Raman spectra was greatly improved by preprocessing. In Fig. 1b, the Raman spectra are smoother, and the Raman peaks of both tissue types are clearer, than in Fig. 1. Most importantly, the preprocessing highlights the differences between the Raman spectra of normal and cancerous tissues. The Raman peaks occur at 827, 1078, 1305, 1447, 1653, and 1747 cm^{-1} in normal tissues, and at 815, 1078, 1243, 1308, 1453, 1663, and 1750 cm^{-1} in cancerous tissues. The assignments of the individual peaks are listed in Table 1 [10].

According to Fig. 1 and Table 1, normal tissue displays four peaks attributable to lipid molecules (1078, 1447, 1653, 1747 cm^{-1}), whereas cancerous tissue displays only two lipid peaks (1078 and 1750 cm^{-1}). Moreover, the peaks at 1078 and 1747 cm^{-1} are attenuated in the cancerous tissues. Meanwhile, peaks representing protein molecules appear at 1243, 1308,

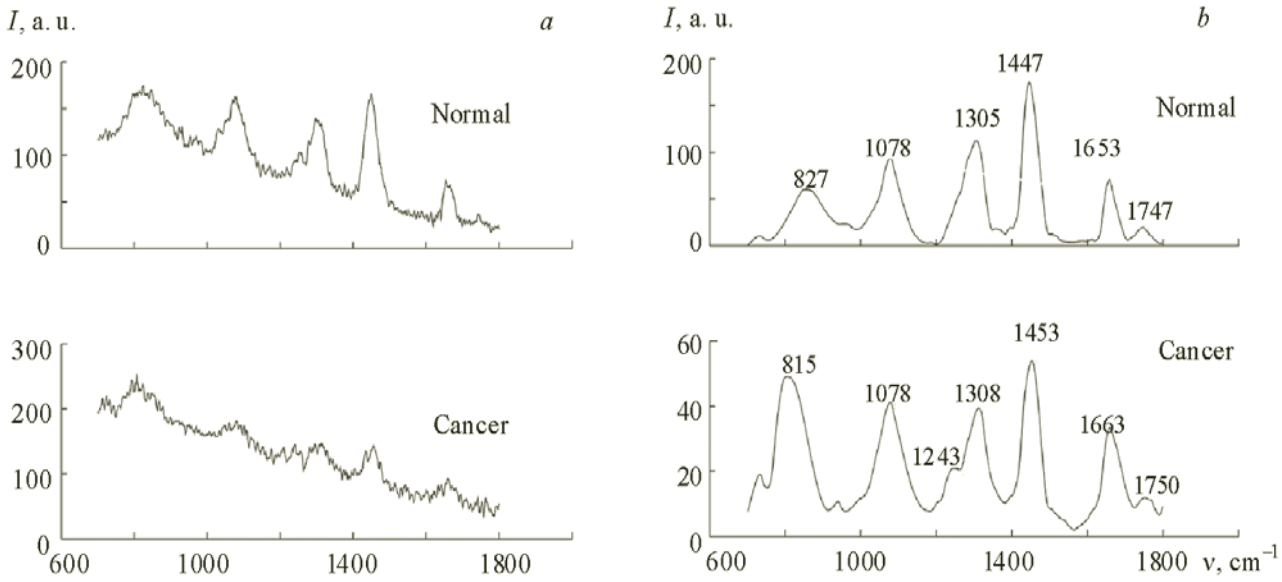

Fig. 1. Typical Raman spectra before (a) and after pre-processing (b).

TABLE 1. Peak Raman Positions and Assignments in Breast Tissue

Peak position, cm^{-1}	Major assignment
827/815	O–P–O stretch (nucleic acid)
1078	C–C or C–O stretch (lipid)
1243	Amide III(C–N stretch) (protein)
1305/1308	Amide III, α-helix, C–C str&C–H (protein)
1447	Scissoring mode of methylene (CH2) (lipid)
1453	CH2 deformation (protein)
1653	lipid
1663	Amide I(C=O stretch) (protein)
1747/1750	C=O stretch (lipid)

1453, and 1663 cm^{-1} in cancerous tissues. In contrast, normal tissue displays a sole prominent protein peak at 1305 cm^{-1}; the others are submerged by the scattering spectra of the lipid molecules. The peak representing nucleic acid molecules appears at 815 and 827 cm^{-1} in cancerous and normal tissues, respectively, and is attenuated in the cancerous tissue. These changes reflect the changing configurations and components and quantities of proteins, lipids, and nucleic acids during tumor formation. The proportions of proteins and lipids were significantly increased and decreased respectively in the cancerous tissues, as reported in previous studies [10, 22].

Statistical analysis. After detecting the outliers, the conventional NIR-Raman spectra of 145 normal tissues and 205 cancerous tissues were classified by ALHK. Two-thirds of the spectra from the normal and cancerous tissues were randomly selected as the training set; the remaining one-third was reserved as the test set. The classification procedure first normalizes the training set to zero mean and unit variance, then normalizes the test set to the corresponding training mean and variance. Finally, the test set is classified by ALHK. To demonstrate the effect of spectral preprocessing, the classification was performed on both raw and preprocessed spectra.

In this study, the three ALHK parameters K, T, and λ were varied as 1–30, 1–10 (in 0.1 increments), and 1–10, respectively. The result yielding the highest testing accuracy was assumed as the optimized classification result. As shown in Table 2, the highest testing accuracy attained by ALHK was 90.6% on the raw spectra, improving to 93.2% on the

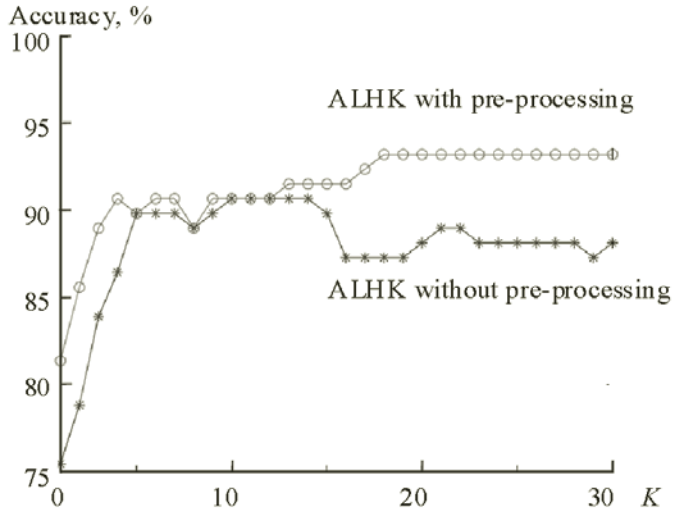

Fig. 2. Classification accuracy as a function of K (the best classification accuracy obtained by optimal choice of parameters T and λ is shown for each K).

TABLE 2. Classification Results of the Test Set

Method	Sensitivity, %	Specificity, %	Predictive value of a positive test, %	Predictive value of a negative test, %	Accuracy, %
ALHK	97.1	81.6	88.1	95.2	90.6
Pre-processing +ALHK	95.6	89.7	92.9	93.6	93.2

preprocessed spectra. The optimal parameters were determined as $K = 10$, $T = 6.6$, and $\lambda = 1$ before preprocessing, and as $K = 18$, $T = 2.4$, and $\lambda = 10$ after preprocessing. The classification accuracies of ALHK for different K are presented in Fig. 2. Note that the classification accuracy is never worsened by the preprocessing but improves it by up to 6.8%, depending on the value of K. This result confirms that preprocessing effectively improves the classification accuracy.

Conclusions. Conventional laser Raman spectra of normal and cancerous breast tissues were acquired by a miniature spectrometer excited at 785 nm. After preprocessing by wavelet transform and polynomial fitting, the major Raman peaks of the normal tissue spectra appeared at 1078, 1447, 1653, and 1747 cm^{-1}, providing information on lipid molecules, whereas the spectra of cancerous tissues peaked at 1243, 1308, 1453, and 1663 cm^{-1}, indicative of protein molecules. Therefore, tumor development is characterized by a significant increase in protein content and a large reduction in lipid content. Finally, we classified the preprocessed Raman spectra by an ALHK classifier. The highest prediction accuracy was 93.2%. For a given K, preprocessing improved the classification accuracy by up to 6.8%, indicating that ALHK with preprocessing can effectively recognize cancerous tissue. We conclude that the miniature spectrometer is a viable diagnostic tool for breast cancer and could prove to be a portable clinical diagnosis technology.

REFERENCES

1. R. Siegel, D. Naishadham, and A. Jemal, *CA-Cancer J. Clin.*, **62**, 10–29 (2012).
2. G. Z. Yu, *Natl. Med. J. China*, **90**, 505–507 (2010).
3. R. R. Alfano, G. Tang, A. Pradhan, W. Lam, D. S. J. Choy, and E. Opher, *IEEE J. Quant. Electron*, **23**, 1806–1811 (1987).
4. Q. B. Li, X. J. Sun, Y. Z. Xu, L. M. Yang, Y. F. Zhang, S. F. Weng, J. S. Shi, and J. G. Wu, *Clin. Chem.*, **51**, 346–350 (2005).
5. C. Murali Krishna, G. D. Sockalingum, Rani A. Bhat, L. Venteo, Pralhad Kushtagi, M. Pluot, and M. Manfait, *Anal. Bioanal. Chem.*, **387**, 1649–1656 (2007).

6. Z. F. Zhuang, N. Li, Z. Y. Guo, M. F. Zhu, K. Xiong, and S. J. Chen, *J. Biomed. Opt.*, **18**, 031103-1-3 (2013).
7. S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, and Z. Huang, *Br. J. Surg.*, **97**, 550–557 (2010).
8. J. C. Zhu, J. Zhou, J. Y. Guo, W. Y. Cai, B. Liu, and Z. G. Wang, *Chem. Cent. J.*, **7**, 1–5 (2013).
9. J. Yang, Z.Y. Wang, S. F. Zong, C. Y. Song, R. H. Zhang, and Y. P. Cui, *Anal. Bioanal. Chem.*, **402**, 1093–1100 (2012).
10. C. H. Liu, Y. Zhou, Y. Sun, J. Y. Li, L. X. Zhou, S. Boydston-White, V. Masilamani, K. Zhu, Y. Pu, and R. R. Alfano, *TCRT*, **12**, 371–382 (2013).
11. R. A. Bitar, H. S. Martinho, C. J. Tierra-Criollo, R. L. N. Zambelli, M. M. Netto, and A. A. Martin, *J. Biomed. Opt.*, **11**, 054001-1–5 (2006).
12. A. F. García, L. Raniero, R. A. Canevari, K. J. Jalkanen, and R. A. Bitar, *Theor. Chem. Acc.*, **130**, 1231–1238 (2011).
13. C. Yu, E. Gestl, K. Eckert, D. Allara, and J. Irudayaraj, *Cancer Detect. Prev.*, **30**, 515–522 (2006).
14. A. Zoladek, F. C. Pascut, P. Patel, and L. Notingher, *J. Raman Spectrosc.*, **42**, 251–258 (2011).
15. A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, J. Lyons, D. Hicks, M. Fitzmaurice, R. R. Dasari, J. P. Crowe, and M. S. Feld, *Cancer Res.*, **66**, 3317–3322 (2006).
16. A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, R. Shenk, N. Wang, R. R. Dasari, M. Fitzmaurice, and M. S. Feld, *J. Biomed. Opt.*, **14**, 054023-1-8 (2009).
17. M. V. P. Chowdary, K. K. Kumar, S. Mathew, L. Rao, C. M. Krishna, and J. Kurien, *Biopolymers*, **91**, 539–546 (2009).
18. S. G. Mallat, *IEEE T. Pattern Anal.*, **11** (1989).
19. H. B. Qi, X. F. Liu, and C. Pan, *Int. Con. Intel. Comput. Tec. Aut.*, **2**, 126–129 (2010).
20. T. Yang and V. Kecman, *Neurocomputing*, **71**, 3001–3004 (2008).
21. T. Yang, V. Kecman, L. B. Cao, C. Q. Zhang, and J. Z. Huang, *Exp. Syst. Appl.*, **38**, 12348–12355 (2011).
22. M. V. P. Chowdary, K. K. Kuntar, J. Kurien, S. Mathew, and C. M. Krishna, *Biopolymers*, **83**, 556–569 (2006).