Terahertz imaging diagnostics of cancer tissues with a chemometrics technique

Sachiko Nakajima, Hiromichi Hoshina,[a] Masatsugu Yamashita, and Chiko Otani
RIKEN, 519-1399 Aramaki-Aoba, Aoba-ku, Sendai, Miyagi 980-0845, Japan

Norio Miyoshi
University of Fukui, 23-3 Shimoaizuki, Matsuoka, Yoshida-gunn, Fukui 910-1193, Japan

(Received 4 October 2006; accepted 18 December 2006; published online 22 January 2007)

Terahertz spectroscopic images of paraffin-embedded cancer tissues have been measured by a terahertz time domain spectrometer. For the systematic identification of cancer tumors, the principal component analysis and the clustering analysis were applied. In three of the four samples, the cancer tissue was recognized as an aggregate of the data points in the principal component plots. By the agglomerative hierarchical clustering, the data points were well categorized into cancer and the other tissues. This method can be also applied to various kinds of automatic discrimination of plural components by terahertz spectroscopic imaging. © *2007 American Institute of Physics*.
[DOI: 10.1063/1.2433035]

By virtue of the penetrative property and the ability to identify chemicals by their spectra, terahertz radiation has been spotlighted in various application fields,[1] especially by applying spectroscopic imaging.[2] In many cases, however, it is difficult to identify the unknown materials straightforwardly, because the spectra can be easily affected by the sample's conditions, such as crystal structure, hydration, optical interference, and scattering.[3,4] A more sophisticated analysis to reveal the underlying characteristics of the observed information is required for the practical use of the terahertz waves.

In the medical field, a pathologic diagnosis using the difference in terahertz absorbance between cancer and normal tissues was reported.[5–7] There are two major merits in the cancer diagnosis: First, it covers a wider area compared to the optical microscopy measurements, and second, once the systematic analysis is established, it can provide an objective judgment which is independent of the skill of medical doctors. However, there still remains an uncertainty in the judgment because of the plain spectral features of the cancer tissues with no remarkable absorption peak. Moreover, due to the variety of the tissues, their characteristic information is not clearly apparent by a simple comparison between the obtained spectra and images.

In this letter, we have introduced the chemometrics[8,9] technique [the principal component analysis (PCA) and the hierarchical clustering analysis] for the systematic and automatic analysis of the terahertz spectroscopic images and applied it to the pathologic diagnosis of the four tumor samples. In three of them, the cancer tissues can be recognized as aggregates of the data points in the principal component plots. By the agglomerative hierarchical (AH) clustering (Ward's linkage),[9,10] the data points were well categorized into cancer and the other tissues. Such a procedure can be applied not only for various kinds of diagnosis of pathological samples from various body parts but also for other various kinds of diagnosis which require an automatic component discrimination.

Four slices of paraffin-embedded liver cancer tissues (samples A, B, C, and D) were prepared without staining. The samples have been treated by standard procedures for *histopathological* examination.[11] Transmission multispectral terahertz images of the samples were taken with an imaging system based on terahertz time domain spectroscopy (made in collaboration with Tochigi Nikon Co.). During the measurement, the spectrometer was purged with nitrogen gas to avoid absorption of the water vapor in the air. The absorbance and refractive index spectra were measured from 0.5 to 2.25 THz with 25 GHz frequency resolution. At each pixel with the size of 250 μm^2, 40 spectra were averaged. The obtained spectral images were analyzed by the PCA and the hierarchical clustering by using commercial software, MINITAB 14 (Minitab Inc.).

Figure 1 shows the photographs and the terahertz images of the four samples. The cancer tissue, identified by optical microscope, is indicated on the photographs as hatched areas. The second and the third columns show the images of the absorbance and the refractive index at 1.7 THz, respectively, at the same scale as the photographs. In these results, no remarkable spectral features were found either in absorbance or in the refractive index except for a small oscillation due to the optical interference. Therefore, the images at the other frequencies are similar to those in Fig. 1. The distinction between the cancer and the other tissues can be partly recognized in the terahertz images. In sample A, the green area with the weaker absorption is in good agreement with the cancer tissue. However, the cancer tissues of sample C show greater absorbance than the other tissues. The cancer information is clearly seen on the refractive index plot of sample B. No clear correlation was found in sample D. Thus, the correlation between the cancer areas and the areas on the terahertz images is different from sample to sample.

In this experiment, the spectra are influenced by a series of factors such as the variety of diseases, the sample condition, interference, and scattering. To extract the cancer information systematically out of the huge amount of multidimensional spectra, PCA has been applied to the observed data.[8] By applying the singular value decomposition, the original data matrix can be projected into a low dimensional principal

[a]Electronic mail: hoshina@riken.jp

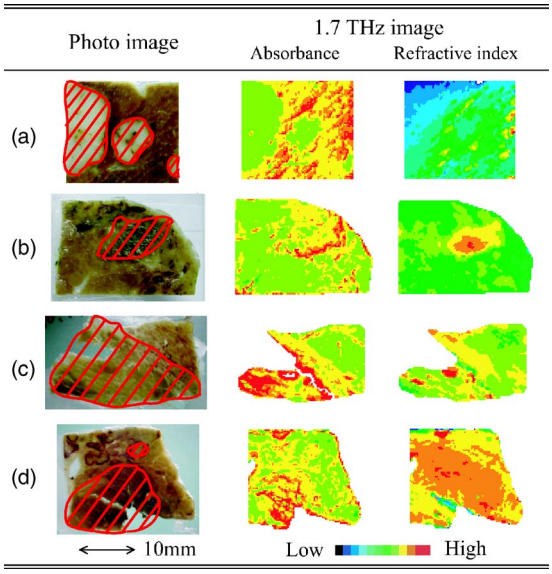

FIG. 1. Photographs and terahertz images of samples A–D. The first column shows photographs of the samples, with the cancer areas marked by red lines. The second and the third columns show maps of the absorbance and the refractive index at 1.7 THz, respectively, in the same scale as the photographs. The absorbance is shown in color, from 0 to 2.0 for samples A, B, and D, and from 0 to 0.7 for sample C. The value of the refractive index is shown in color from 1.0 to 1.7 for all samples.

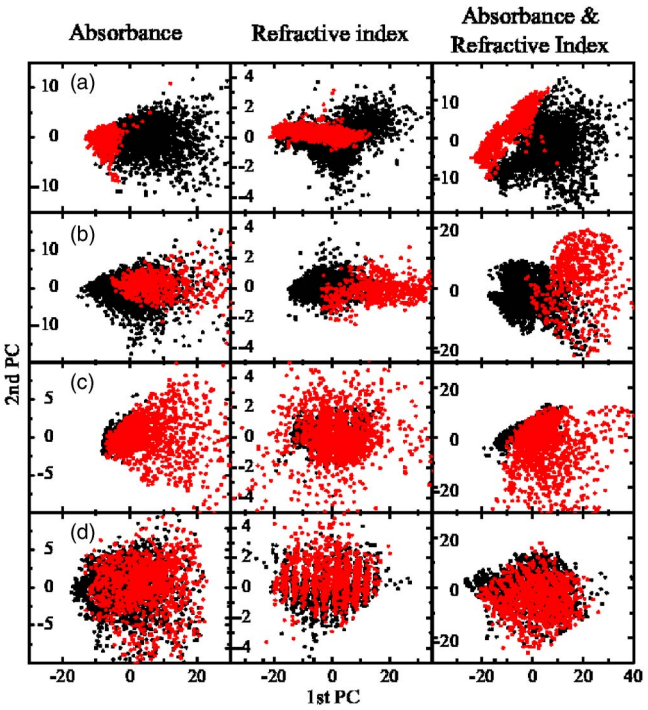

FIG. 2. Score plots with the first (PC1) and the second (PC2) PCs. Each column shows the result of PCA relative to the absorbance spectra, to the refractive index spectra, and to both. The pixels of the cancer tumor are shown in red.

component (PC) space. Since PCs are set to make the largest variance between data points, the information of the original data can be accounted for a smaller number of variables.

Figure 2 shows the score plot of the first PC versus the second PC. In this figure, each point corresponds to a single pixel of the image, with the red points representing the pixels in the area marked in red in Fig. 1. Before applying PCA, the data matrices of the spectra were normalized by their standard deviation for the entire images. In the first and the second columns, PCA was applied to the observed absorbance and refractive index images, respectively. On these plots, more than 90% of the variance was explained within the first and the second PCs. Since the similarity of the spectra appears as the nearness on the PC plots, the cancer tumor can be recognized as an aggregate of points. A separate aggregate of red points is seen in the absorbance plot of sample A and

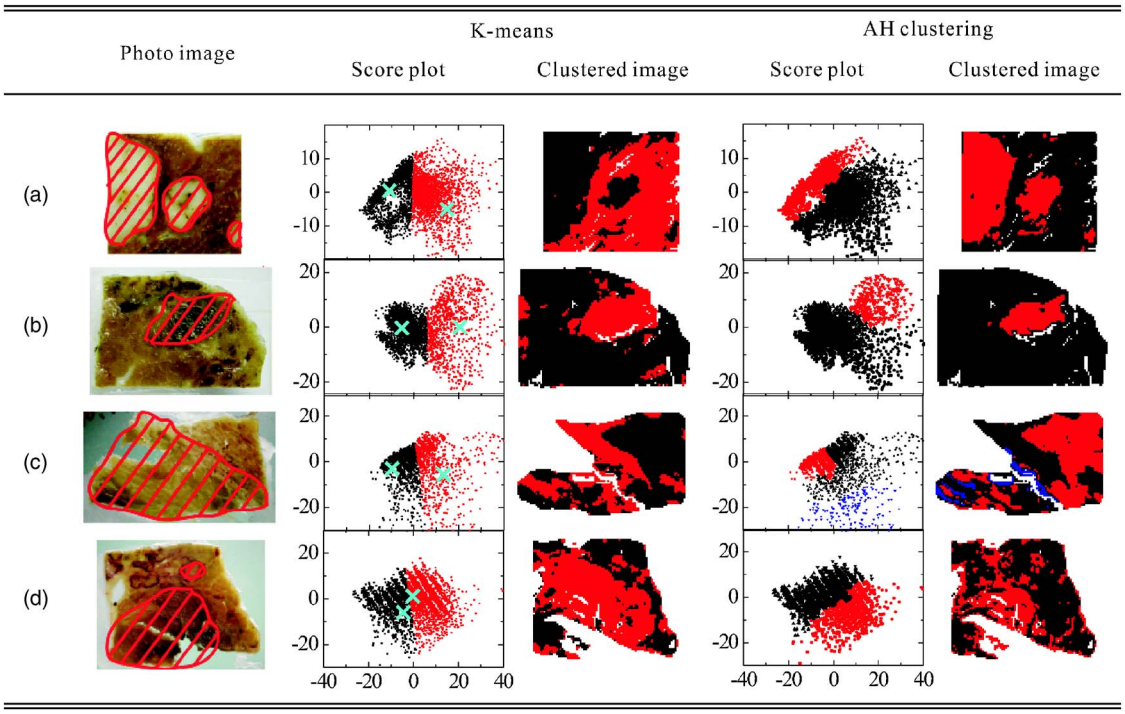

FIG. 3. Results of the analysis by *k*-means clustering (with crosses marking the initial points) and by hierarchical clustering (Ward's algorithm with the square of Pearson's distance).

in the refractive index plot of sample B. However, the red and the black points are not clearly separated in the case of sample C. This is probably because the difference of the spectrum is too small. No separation has been found at sample D.

On the third column of Fig. 2, PCA was simultaneously applied to both the absorbance and the refractive index spectra. In all four samples, more than 90% of the variance was explained by the first and the second PCs. The first PC tends to reflect the magnitude of the refractive index, and the second PC reflects the magnitude of the absorbance. In plots A–C, the aggregates of the red points can be seen. Note that the cancer points of sample C appear separately only by this procedure. Thus, the cancer tissue identification is much easier on the score plots using the merged data of the absorbance and refractive index spectra.

In order to categorize the aggregates systematically into cancer and the other tissues, the clustering was applied to the score plot.[9] Based on the assumption that the distances of points reflect the similarity of their properties, the clustering analysis categorizes the data points into several groups by their distance.[12] This analysis has been used for various imaging spectra in other frequencies.[13,14] As discussed in these studies, the result of the clustering analysis highly depends on the algorithm.

Two kinds of general approach, AH clustering and nonhierarchical cluster analysis (k-means[12]), have been tested. The former can be made without *a priori* assumptions, but the latter needs to give initial points by hand. Figure 3 shows the results of k-means clustering and the AH clustering. Initial points of the k-means clustering are marked by crosses. The left-side plots of each column show the result on the score plot, and the right-side figures depict them on the spatial images. It was found that the k-means clustering simply separates the area with a straight line so that the result does not reflect the cluster shape well (see the third column of Fig. 2); on the other hand, the AH clustering shows a better correspondence with the actual cancer tumor area in samples A–C.

In this study, various methods of the AH clustering have been tried, such as average linkage, centroid linkage, complete linkage, median linkage, and Ward's linkage.[10] Among them, Ward's linkage with the square of Pearson's distance shows the best agreement. Note that this method tends to be affected by the points corresponding to the defects of the samples. Therefore, in sample C, we classified the points into three clusters, i.e., cancer tissues, the other tissues, and the sample defects. Thus the preparation of defectless sample is important for this method.

By applying the clustering analysis on PC plots, the systematic diagnosis of the disease area was realized. We emphasize that this method works complementally to the diagnosis done by medical doctors using the optical microscope because the clustering does not specify which the cancer area is. In addition, as seen in sample D, the cancer tumor cannot always be distinguished by the terahertz spectra. By the optical microscopic observation, it was found that viable cancer cells were dominant in samples A–C, whereas considerable amount of necrotic cells were found in sample D. Such difference of the organ condition may possibly be the reason for the difficulty in sample D. More research for the particular sample type is necessary.

In conclusion, terahertz spectroscopic images of cancer tissues have been recorded by a terahertz time domain spectrometer and processed using the PCA and the clustering technique for a potential systematic identification of cancer tumors. In three of the four samples, the cancer tissue was classified by the agglomerative hierarchical clustering on the principal component plots. We showed that the chemometrics could work effectively in terahertz applications for the purpose of simplifying complicated spectral datasets.

The authors are grateful to K. Kawase for the valuable discussion and encouragement.

[1] D. L. Woolard, W. R. Loerop, and M. S. Shur, *Terahertz Sensing Technology: Emerging Scientific Applications & Novel Device Concepts* (World Scientific, Singapore, 2004).
[2] K. Kawase, Y. Ogawa, and Y. Watanabe, Opt. Express **11**, 2549 (2003).
[3] P. Tady, I. Bradley, D. Arnone, and M. Pepper, J. Pharm. Sci. **92**, 831 (2003).
[4] C. F. Zhang, E. Tarhan, A. K. Ramdas, A. M. Weiner, and S. M. Durbin, J. Phys. Chem. B **108**, 10077 (2004).
[5] D. Arnone, C. Clesla, and A. Corchia, Proc. SPIE **3823**, 209 (1999).
[6] R. Woodward, B. Cole, V. Wallace, R. Pye, D. Arnone, and E. Linfield, Phys. Med. Biol. **47**, 3853 (2002).
[7] P. Knobloch, C. Schildknecht, T. Kleine-Ostmann, M. Koch, S. Hoffmann, M. Hoffmann, E. Rehberg, M. Sperling, K. Donhuijsen, G. Hein, and K. Pierz, Phys. Med. Biol. **47**, 3875 (2002).
[8] J. N. Miller and J. C. Miller, *Statistics and Chemometrics for Analytical Chemistry* (Prentice Hall, New York, 2005).
[9] H. H. Harman, *Modern Factor Analysis* (University of Chicago Press, Chicago, 1976).
[10] J. Ward, J. Am. Stat. Assoc. **58**, 236 (1991).
[11] L. Carlos and J. Carneiro, *Basic Histology* (McGraw-Hill, New York, 2003).
[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification* (Wiley, New York, 2000).
[13] P. Lasch, W. Haensch, D. Naumann, and M. Diem, Biochim. Biophys. Acta **1688**, 176 (2004).
[14] M. Jackson, B. Ramjiawan, M. Hewko, and H. H. Mantsch, Cell Mol. Biol. (Paris) **44**, 89 (1998).