SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM

JBEMi JOURNAL OF BIOMEDICAL ENGINEERING AND MEDICAL IMAGING

ISSN: 2055-1266
VOLUME 3 ISSUE 1

In vivo Tumor Spatial Classification using PCA and K-means with NIR-Hyperspectral Data

[1]Mai Kasai, [2]Yuya Yasuda, [1]Hiroshi Takemura, [1]Hiroshi Mizoguchi, [2]Kohei Soga, and [3]Kazuhiro Kaneko
[1]Faculty of Science and Technology, Tokyo University of Science, Japan;
[2]Faculty of Industrial Science and Technology, Tokyo University of Science, Japan;
[3]National Cancer Center Hospital East, Japan;
7515613@ed.tus.ac.jp

ABSTRACT

This paper presents new method of spatial classification and wavelength bands reduction of near-inferred (NIR) hyperspectral imaging data for medical application. Hyperspectral imaging data have more than several hundred wavelength bands. However hyperspectral data have sometimes redundant information to detect region of interest. The aim of this research is to archive became possible that a region of interest is distinguished by observing the particular wavelength bands without any markers in order to develop a special application such as a surgery supporting system. NIR light with wavelengths of 800-2000 nm, called as the 'biological window,' has received particular attention given that water and biological tissues have minimal optical loss caused by scattering and absorption at these wavelengths. NIR light can penetrate/see through deep tissues. NIR endoscope have a great potential as the surgery supporting system, however wavelength bands needs to reduce according to the limitation of NIR endoscope hardware performance. To consider only several wavelength bands are sometimes much effective case than to consider all wavelength bands. In this paper, we proposed the method of spatial classification and reduction a number of wavelength bands simultaneously by combined PCA and k-means, and assessed the cancer-caring nude mouse. The experimental results demonstrate that the proposed method can select valuable wavelength bands to distinguish the region of interest with comparable accuracy of the conventional method.

Keywords: Spatial Classification, PCA, k-means, Hyperspectral data, Near-Infrared

1 Introduction

Hyperspectral imaging is a combination of imaging and spectroscopic technology [1]. Hyperspectral imaging data have more wavelength bands than RGB images taken by the ordinary RGB camera. Each pixel of hyperspectral imaging has a spectrum data. Differences that appear subtle to the human eye could be significant when looking at the detailed spectrum. Hyperspectral imaging data is utilized in several fields: remote sensing, biological engineering, agriculture, food engineering, and so on. In the medical field, a region of interest (ROI) means an image area required at a diagnosis [2-3]. If it became possible that ROI of a nerve or a lesion part is distinguished by observing the particular wavelength bands without any markers, a special application as a surgery supporting system is realized. Near-inferred light (NIR) imaging has such a potential. NIR light with wavelengths of 800-2000 nm, called as

DOI: 10.14738/jbemi.31.1892
Publication Date: 27th February 2016
URL: http://dx.doi.org/10.14738/jbemi. 31.1892

the 'biological window,' has received particular attention given that water and biological tissues have minimal optical loss caused by scattering and absorption at these wavelengths. NIR light can penetrate/see through deep tissues. NIR endoscope have a great potential as the surgery supporting system [4-5], however, hyperspectral data have sometimes redundant information to detect region of interest. To consider only several wavelength bands are sometimes much effective case than to consider all wavelength bands. And wavelength bands need to reduce according to the limitation of NIR endoscope hardware performance.

There are many analysis techniques to detect the ROI. These include Gabor filter, grey level co-occurrence matrix (GLCM), principle component analysis (PCA), minimum noise fraction (MNF), wavelet transform, etc. [6-10]. Multitude studies apply the analysis of a machine leaning method using all of wavelength bands, spectral peak and extracted features. To distinguish the ROI from enormous multidimensional data by using a conventional machine learning method, it is necessary to acquire a large amount of the dataset. However the collecting vast quantities of information on medical field such as rare intractable diseases is very hard. This is because that the conventional machine learning method is difficult to apply directory to the medical application. Additionally, hyperspectral images have redundant information about an object, and often require application of dimensionality reduction methods such as PCA to remove the redundant information. One example, Naganathan et al. [11] develop and test hyperspectral imaging system to predict tenderness of beef from hyperspectral images. This method combined PCA, GLCM and canonical discriminant analysis (CDA). Combined with these methods, this analysis technique is implemented to extract features from the hyperspectral images. However this method cannot be applied to unknown sample of composition and spectrum. On the other hand, k-means clustering [12-13] is the analysis to distinguish the ROI without the machine learning. Hyperspectral data, however, have sometimes redundant information not to enhance more accuracy using only k-means. To consider only several wavelength bands are sometimes much effective case than to consider all wavelength bands. The wavelength bands need to reduce according to an application of NIR endoscope, combined with the NIR-fluorescent imaging system [14].

We propose the method of spatial classification for a tumor and reduction a number of wavelength bands simultaneously. By combining PCA [15] and k-means clustering, we visualize the analyzed results and reduce wavelength bands based on a factor loading value of PCA. The proposed method applies the tumor region of the cancer-caring nude mouse distinguish experiments. In the experiments, the selected wavelength bands calculated by the proposed method is performed favorable behavior compared to original ROI calculated by all wavelength bands.

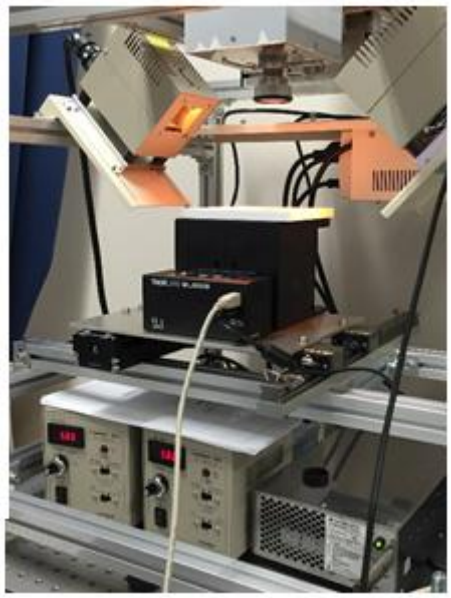


Figure 1. Compovision®(Sumitomo Electric Industries, Ltd., CV-N800HS).

2 Proposed Methods

Figure 1 shows the NIR-hyperspectral imaging camera, Compovision (Sumitomo Electric Industries, Ltd., CV-N800HS) using in this study. The wavelength resolution of the Compovision is approximately 6nm from 913.78 to 2522.44 nm, and this is a line scan camera, 320 pixel par line. Each pixel has 256 wavelength bands. The flowchart of the proposed method is shown in Figure 2. The proposed method is mainly composed of the following six process.

2.1 Preprocessing

The purpose of preprocessing is to calibrate and reduce noise of the captured hyperspectral imaging data. Each pixel of hyperspectral imaging data is converted to 234 wavelength bands from 256. The first 11 bands and the last 11 bands do not use for the analysis due to low signal to noise ratio. As a reflectance calibration, original data *V* is corrected from the dark and white data of the camera. The dark data *D*, obtained by turning off the light and covering the lens with a lens cap, is the background response of camera. The white data *W* is obtained with the standard reflector, which is not absorption in near-infrared wavelength bands to correct a wavelength sensitivity characteristic by the camera and light. Corrected reflectance value *f* is calculated as follows:

$$f = \frac{V - D}{W - D} * 65535 \quad (1).$$

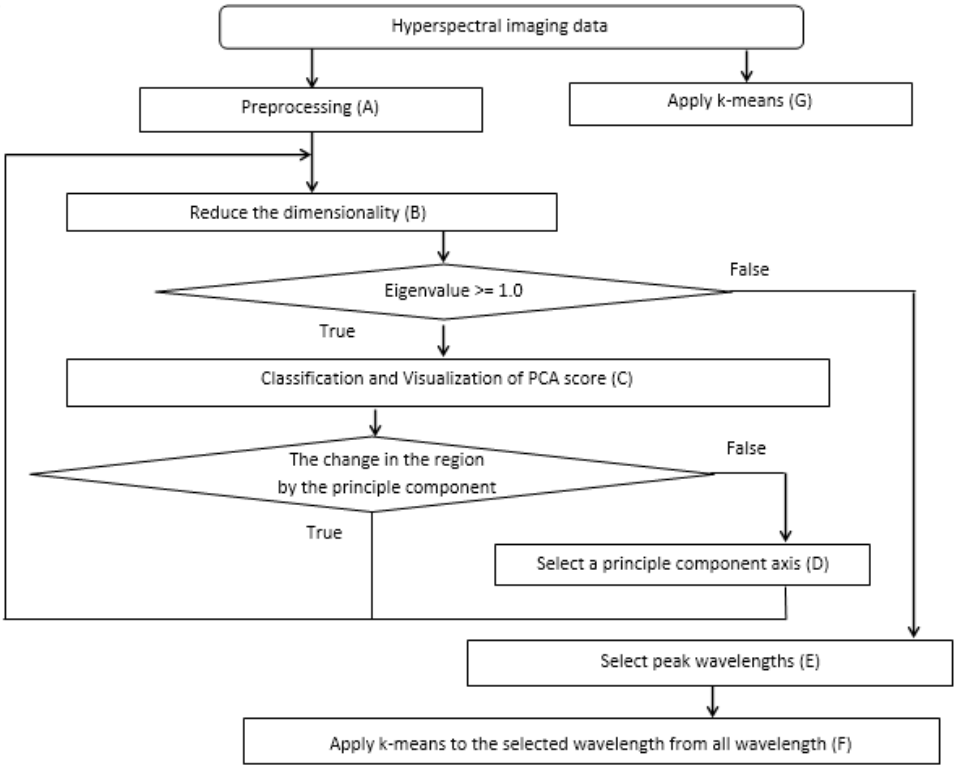


Figure 2 Flow chart of the proposed method

2.2 Reduce the dimensionality (by PCA)

The multi-dimensional wavelength bands are reduced by PCA as information content is maintained. PCA is a technique of multivariate analysis that reduces the dimensionality. PCA has been utilized to reduce dimension, extract feature, compress data, and identify key factor. If an eigenvalue of PCA is more than 1.0, the PCA score are classified and visualized as the result image calculated by using k-means clustering.

2.3 Classify and Visualize of PCA score (by k-means)

K-means clustering is a representative technique as a partitioned optimization clustering technique that search optimal partition by an evaluation function. PCA score is divided into some cluster by similarity on principle component axis. The evaluation function is defined by

$$E = \sum_{i=1}^{k} \sum_{x \in c_i} (dist(\boldsymbol{x}, \boldsymbol{c}_i))^2 \quad (2)$$

Where $\boldsymbol{x}$ is PCA score calculated by (B), k is number of clusters, $\boldsymbol{c}_i$ indicates centroid points of each cluster, and $dist(\boldsymbol{x}, \boldsymbol{c}_i)$ measures the Euclidean distance between a pattern $\boldsymbol{x}$ and its cluster centroid point $\boldsymbol{c}_i$. PCA score is partitioned into k clusters to minimize value of the function E. Each cluster is distinguished and described on the 2D image, which captured with the hyperspectral image data at the same time, by using different colors to visualize clustering results.

2.4 D. Select the principle component axis

The proposed method need to select a principle component axis to visualize k-means clustering results. PCA continues to apply to the captured hyperspectral image data until an eigenvalue of PCA is less than 1.0 or a cumulative contribution ratio of PCA are more than 99.98 %. An integer number of n is defined as maximum number of principal components at this time. The k-means clustering result images of PCA are calculated based on each principal component from 1 to n, respectively. If a different between the clustering result image of i th and $i+1$ st principal components is more than a threshold, $i+1$ st principle component that contained feature information to divide image region is selected. If the region of the result image does not change as a whole, all n th principle components are selected.

2.5 Select peak wavelength bands

The peak wavelength bands are selected as significantly peak values of respective axis according to a loading factor value as the following equation.

$$\begin{cases} \text{if} & (\sigma_m^2 > 0.8 * \sigma^2 \text{ or } \mu < 0.2 * max) \\ & \text{Select the wavelength bands.} \\ & \left(\text{satisfied the conditions that loading factor value takes more than max*0.95} \right) \\ \text{else} & \text{Don't select the wavelength band.} \end{cases} \quad (3)$$

Where σ^2 is the dispersion of loading factor, σ_m^2 is the dispersion of loading factor from the third-quartile to maximum value (*max*), and μ is average value of loading factor. In PCA, the loading factor indicates the correlation between the principle component and the identified the factor that affects each principle component. Loading factors value of the selected principle component is compared and extracted the peak wavelength bands as feature values. If the loading factor value does not show the significantly peak, the loading factor is regarded as no factor that concerned principle component strongly and extracted as no wavelength from this axis. Regarding the peak wavelength bands of the loading factor value in selected principle component axis by changed region, the required feature values are extracted to discriminate particular region from multidimensional and enormous informative data. The results are evaluated by an error between the correct answer data.

Figure 3. Pseudo-color image of the target. Orange area show the tumor area.

2.6 Apply k-means clustering to the selected wavelength bands from all wavelength bands.

This section describes visualizing evaluating of the proposed method. The reduced data only selected wavelength bands are applied k-means clustering. The analyzed results are visualized by the 2D image. The original data (234 wavelength bands) is applied k-means clustering similarly for comparison. The result using all wavelength bands (G) is compared with the result of the proposed method using the selected wavelength bands (F) for evaluation. The total number of pixels of a target ROI where is deciding comparison objective region manually calculated as the correct answer data. The visualized target ROI is calculated by the proposed method and compared in the result images.

3 Evaluation Experiment

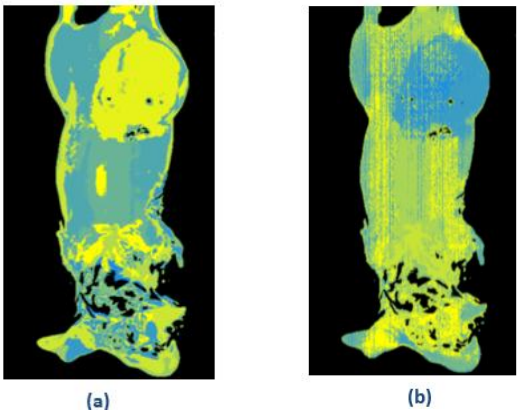
3.1 Experimental Condition

The proposed method is applied to distinguish by the tumor of mains. The target is the cancer-caring nude mice. This study received ethics committee approval of Tokyo University of Science. We acquired the hyperspectral image data that is 320 (pixels per line) x 234 (wavelength bands per pixel) x 720 (number of lines) and the resolution is about 0.10[mm^2/pixel]. One image data is less than 2-3 [s] to scan. The mouse mask binary imaging data is generated by using the brightness value of 1394.64 nm. Pseudo-color image of the cancer-caring nude mouse shows in Figure 3. Orange area indicate the tumor area.

3.2 Cross-Validation Method

The proposed method selects the wavelength bands as feature values using changing area by visualize PCA score. The usability of selected wavelength bands from a mouse is assessed to apply the selected wavelength bands to other mouse data. The distinguished area that is acquired applying the data of selected wavelength bands to k-means clustering is compared with the ROI area in visible image and evaluated by Equation (2).

$$\text{Sensitivity} = \frac{distinguished\ area\ [pixel]}{target\ area\ [pixel]}$$
$$\text{Specificity} = \frac{not\ distinguished\ area\ [pixel]}{not\ target\ area\ [pixel]} \quad (2)$$



(a) (b)

Figure 4. Results of distinguished area by k-means. Result using all wavelength bands data (a), Result using only selected wavelength bands data (b)

4 Results

The results of distinguished area by k-means show Figure 4. The result using all wavelength data is Figure 4-(a) and using only selected wavelength bands data is Figure 4-(b). The sensitivity and specificity of the all results area are calculated by comparing with the visible image of tumor area (Figure 3). In Table 1, "Mouse 1" is target data used to select the wavelength bands and the sensitivity and specificity are calculated by using other mics. "ALL" is the results using all wavelength bands data, and "SELECT" is the results using only the selected wavelength bands data. The 2, 3, 5, 13, 18th principle components and seven wavelength bands are selected by the proposed method. The sensitivity and specificity of the selected wavelength bands data differs little from that of the all wavelength bands data. The original data are reduced to only seven wavelength bands from 234 wavelength bands. The selected wavelength bands as the feature values of each mouse by the proposed method are the same wavelength bands by coincidence.

Table 1. Error Identification Rate

		"Mouse 1"	"Mouse 2"	"Mouse 3"	"Mouse 4"	Average
Sensitivity	ALL	0.87	0.24	0.75	0.57	0.61
	SELECT	0.86	0.27	0.77	0.55	0.61
Specificity	ALL	0.77	0.86	0.93	0.94	0.88
	SELECT	0.85	0.81	0.83	0.87	0.84

5 Discussion

5.1 Reduce the wavelength bands

The approach of extracting feature value by PCA is the major method in data analysis research field. However the selected feature by PCA tend to extract the more salient and major difference feature.

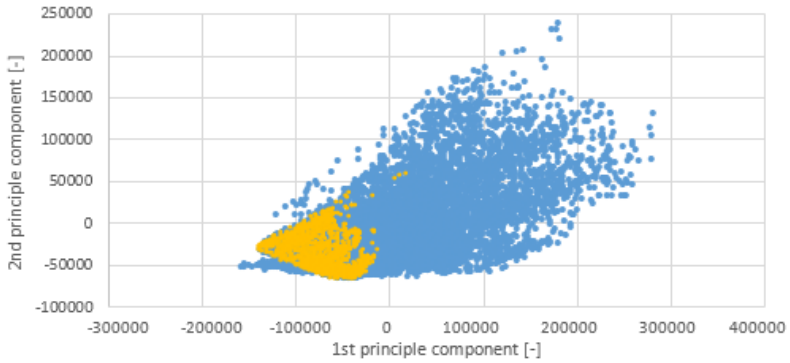


Figure 5. Distribution of principle component scores. The horizontal axis is 1st principle component and the vertical axis is 2nd principle component. Orange color plots show the tumor area and blue color plots show other area (Fig. 3.).

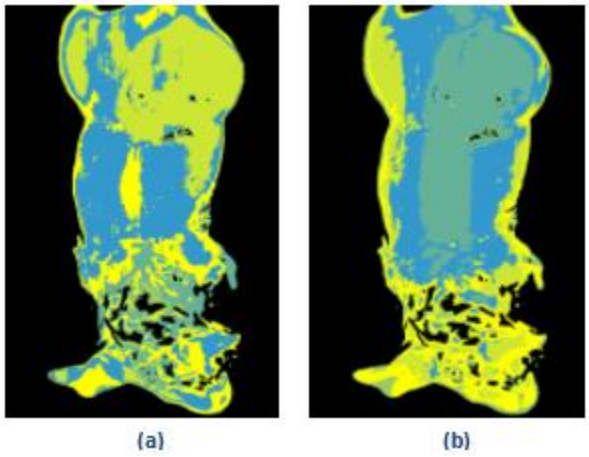


Figure 6. Result image only using the 1st (a) and 2nd (b) principle component data respectively.

Extracting the small/ minor difference in similar properties area, it is difficult to select the principle component axis only. Figure 5 shows the distribution of PCA score between 1st and 2nd principal component axis. The vertical axis is 1st and the horizontal axis is 2nd principle component axis. Orange dots indicate the tumor area and blue dots indicate other area every 10 [pixel]. Figure 5 shows that the tumor area is difficult to distinguish from other area of mouse using only the 1st and 2nd PCA scores. In particular, the tumor area is divided due to the 1st principle component. The distinguished results show at Figure 6 using only the 1st and 2nd principle component, respectively. The results indicate that the 1st and 2nd principle component has keep information (the cumulative contribution ratio of 1st and 2nd principal component is more than 99.98%), however, has redundant information to distinguish the tumor area.

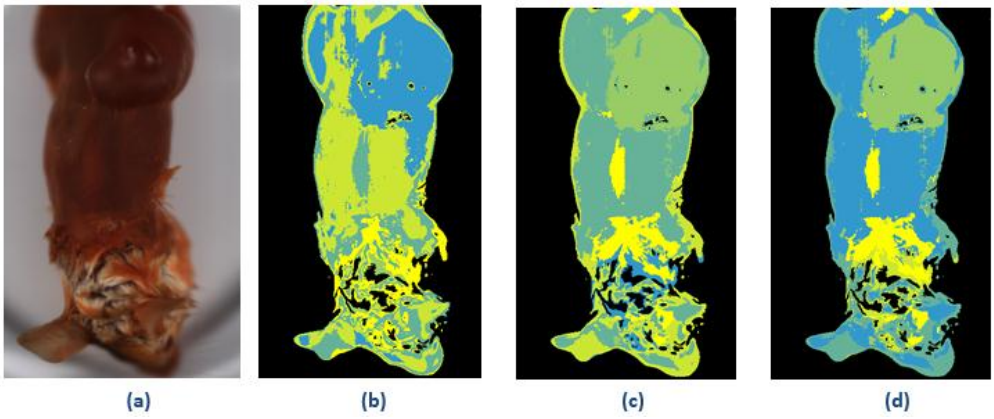


Figure 7. Images show changing the distinguished result by the number of reduced dimensions. (a) shows the Pseudo-color image, (b)~(d) show the result area 2th ~ 4th principle components.

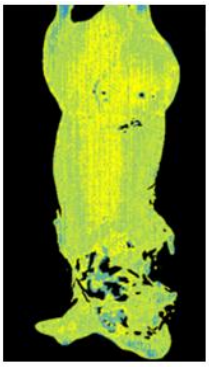


Figure 8. Result image using only the 3rd principle component axis

The proposed method can select the principle component axis focusing on the identification result area depend on small/minor difference feature, not salient/major difference feature. Changing the distinguished result by the number of reduced dimensions show in Figure 7-(b) ~ (d). When compared the result area between 2th ~ 4th principle components, 3th principle component axis is selected because the difference of area of Figure 7-(b) ~ (d) is depended on 3th axis. However change at area having no connection with ROI is ignored. The wavelength bands is selected to distinguish the ROI by the factor loading value of every principle component axis. However, when the dispersion of the factor loading value is more than 0.8, a wavelength bands is not selected. Figure 8 shows the factor loading of 3th principle component axis. The 2, 3, 5, 13, 18th principle component axis, and seven wavelength bands are selected. This result suggested that the 3th axis has useful information to distinguish of the tumor area. This flow is had done until satisfied the conditions that a cumulative contribution ratio of PCA takes more than 99.98% and the principal component axis is selected.

5.2 Investigate the selected wavelength bands

The selected wavelength bands is used to investigate for other mouse data. The imaging conditions is the same and target is other mouse's tumor. Table 1 shows the distinguished result using the selected wavelength bands data by the proposed method. The data of 2~4 mouse except the mouse used to select the wavelength bands is acquired, and same as "Mouse 2", "Mouse 3" and "Mouse 4," respectively. In the case of "Mouse 1" used to select wavelength bands, the result area is distinguished with high precision than using all of wavelength bands data. In the case of "Mouse 2", discrimination precision is low. However in the "Mouse 2~4", the difference of the sensitivity and specificity is small compared with the case of using all of wavelength bands. Necessary information can be extracted to distinguish the ROI.

6 Conclusion

This paper proposed the method of spatial classification and reduction a number of wavelength bands to distinguish the region of tumor by combined PCA and k-means. The proposed method can select the valuable principle component axis focusing on the identification result area depend on changing the number of dimensions reduced by PCA to extract the feature in resemblance area, not large /major difference feature. The proposed method also select the wavelength bands as the feature value using the factor loading value of the each selected principal component axis. The selected wavelength bands are assessed by data of four tumor mouse by compared with the all wavelength bands. Applying the

selected wavelength bands to the mouse average sensitivity is 0.61, and applying the all wavelength bands average sensitivity is 0.61. The specificity of seven wavelength bands distinguished is 0.84. Although there is not large difference comparing with using all of wavelength bands (0.88), the proposed method can reduce dimension and select only seven wavelength bands from 234 wavelength bands. The proposed method is effective for the situation that the composition and spectrum of the target is unknown. To select the wavelength bands has valuable information using the proposed method, learning analysis applies the hyperspectral data of unknown subject efficiently.

REFERENCES

[1]. Schultz R. A., Nielsen T., Zavaleta J. R., Ruch R., Wyatt R., Ganner H. R., Hyperspectral Imaging: A Novel Approach For Microscopic Analyasis, Cytometry, Vol. 43, Issue 4, 2001, pp. 239-247.

[2]. Guolan Lu, Baowei Fei, Medical hyperspectral imaging: a review, Journal of Biomedical Optics, Vol. 19(1), 2014, 010901.

[3]. Guolan Lu, Luma Halig, Dongsheng Wang, Zhuo Georgia Chen, and Baowei Fei, Hyperspectral Imaging for Cancer Surgical Margin Delineation: Registration of Hyperspectral and Histological Images, Proc SPIE. NIH Public Access Author Manuscript, 2014, No. 92036, pp.1-11.

[4]. Zako T., Ito M., Hyodo H., Yoshimoto M., Watanabe M., Takemura H., Kishimoto H., Kaneko K., Soga K., Maeda M., Extra-iluminal detection of assumed colonic tumor site by near-infrared laparoscopy, Surgical Endoscopy, 2015, pp. 1-7.

[5]. Zako T., Hyodo H., Tsuji K., Tokuzen K., Kishimoto H., Ito M., Kaneko K., Maeda M. and Soga K., Development of near infrared-fluorescent nanophosphors and applications for cancer diagnosis and therapy, Journal of Nanomaterials, 2010 , Vol. 2010, pp.1-7.

[6]. Amgren M., Hansen PW., Eriksen B., Larsen J., Larsen R., Analysis of Pregerminated Barley using Hyperspectral Image Analysis, Journal of Agricultural and Food Chemistrt, Vol. 59, 2011, pp. 11385-11394.

[7]. Yaguchi A., Kobayashi T., Watanabe K., Iwata K., Hosaka T., Out N., Cancer Detection From Biopsy Images using Probabilistic and Discriminative Features, 2011 18th IEEE International Conference on Image Processing (ICIP), 2011, pp. 1609-1612.

[8]. Serranti S., Cesare D., Marini F., Bonifazi G., Classification of oat and groat kernels using nir hyperspectral imaging, Talanta, Vol. 103, 2013, pp. 276-284.

[9]. Okamoto, H., Murata, T., Kataoka, T. and Hata, S., Plant classification for weed detection using hyperspectral imaging with wavelet analysis, Weed Biology and Management, 2007, Vol. 7, pp.31-37.

[10]. Jeng-Ren Duann, Chia-Ing Jan, Mnang Ou-Yang, Chia-Yi Lin, Jen-Feng Mo, Yung-Jiun Lin, Ming-Hsui Tsai, Jin-Chern Chiou, Separation spectral mixtures in hyperspectral image data using independent component analysis: validation with oral cancer tissue sections, Journal of Biomedical Optics, 2013, Vol. 18, No. 12, 126005.

[11]. Naganathan G. K., Grimes L. M., Subbiah J., Calkins C. R., Samal A., Meyer G. E., Visible/Near-infrared Hyperspectral Imaging for Beef Tenderness Prediction, Computers and Electronics in Agriculture, Vol. 64, 2008, pp. 225-233.

[12]. MacQueen J., Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1: Statistics, 1967, pp. 281-297.

[13]. Wagstaff K., Cardie C., Rogers S., Schroedl S., Constrained K-means Clustering with Background Knowledge, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 577-584.

[14]. Komoriya K., Takemura H., Mizoguchi H., Soga K., Hyodo H., Kishimoto H., Kaneko K., NIR-fluorescent Imaging by Head-Scanning Mechanism for Near-Infrared Endoscope, Transaction of Japanese Society for Medical and Biomedical Engineering, Vol. 51, No. 2, 2013, pp.135-141.

[15]. Peason K., On lines and planes of closest fit to systems of point in space, Philosophical Magazine, Vol. 2, 1901, pp.559-572.