34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Segmentation of Diffuse Reflectance Hyperspectral Datasets with Noise for Detection of Melanoma*

Ricky Hennessy, Sheldon Bish, James W. Tunnell, *Member, IEEE*, Mia K. Markey, *Senior Member, IEEE*

***Abstract*— We present a segmentation algorithm that allows optical properties to be extracted from diffuse reflectance hyperspectral datasets with a speedup of three orders of magnitude when compared to current methods. Such data could be used for the detection of melanoma. The algorithm first performs dimensionality reduction using principal component analysis, and then the image is segmented using k-means clustering. The mean spectrum from each cluster is then calculated and can be used to extract chemical information. By reducing the number of spectra to be analyzed, extraction of physiological information can be achieved three orders of magnitude faster than methods requiring the analysis of every spectrum in the hyperspectral dataset. The effect of noise on the ability of the algorithm to accurately segment images was tested using digital phantoms, for which the noise level was under the control of the investigators. The analysis showed a linear relationship between the level of noise and the smallest difference in scattering that the algorithm was able to accurately detect and segment. This finding can be used to determine the maximum amount of noise in the imaging system that will still allow detection of the difference in optical properties between non-melanoma and melanoma.**

I. INTRODUCTION

Skin cancer is the most common form of malignancy. For melanoma skin cancer, early detection is critical to patient survival: The five-year survival rate for early stage melanoma is 98%, whereas the five-year survival rate for late stage melanoma is 16% [1]. Optical techniques offer a noninvasive alternative to tissue biopsy for determining disease status [2]. The interaction of light with the lesion provides information about tissue morphology, function, and biochemical composition. As these physiological parameters change with disease progression, optical methods offer a means to measure that progression noninvasively [3].

Current methods for diffuse reflectance spectroscopy acquire only a single spectrum at a single point, and are unable to provide spatial information. To overcome the inaccuracies due to sampling error that arise from single point measurements, Wang et al. have developed a spectroscopic imaging system capable of collecting hyperspectral images [4]. The resulting images contain individual spectra vectors collected from thousands of discrete pixel points in the sample. In a hyperspectral image, each pixel is a single N-dimensional data point, where the number of dimensions (N) is equal to the number of spectral bands (the length of the spectral vector). Multivariate analysis methods can be used to extract chemical information from the individual spectra to reconstruct images from the hyperspectral dataset [5]. Miljkovic et al. have shown that unsupervised machine learning methods are capable of detecting spectral features; however, it is unclear which method is most sensitive and generally applicable [6]. Unsupervised methods have the advantage of not requiring any knowledge of sample composition, but unsupervised methods are unable to provide quantitative information in terms of biochemical differences between the spectral classes.

To extract quantitative information from diffuse reflectance spectra, Rajaram et al. developed a lookup table-based inverse model [7]. This method relies on a lookup table (LUT) generated from experimental measurements on tissue-simulating phantoms with known optical properties. To fit the measured reflectance spectra and extract optical properties, they implemented a nonlinear optimization fitting routine. The average time to fit a single spectrum is approximately two seconds. While this amount of time is trivial if only a few spectra are being analyzed, using the same method to analyze a hyperspectral image containing over 10,000 spectra would take too long to work as an effective on-site clinical diagnostic tool.

Numerous other methods have been proposed for the analysis of hyperspectral images of biological tissue [6], [8], [9], [10], [11] Techniques involving cluster analysis have successfully shown that it is possible to segment hyperspectral images into tissue types [12]. More scalable techniques, such as k-means, provide a rapid method for segmentation of hyperspectral images [13]. Clustering approaches offer the advantage of providing mean spectra with low noise for each cluster, which can then be used to extract quantitative information.

Dimensionality reduction is another commonly used technique in hyperspectral image analysis [6], [14]. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. This facilitates segmentation, classification, and visualization of high-dimensional data, such as hyperspectral images. Dimensionality reduction techniques can be described as either linear or nonlinear. Nonlinear techniques do not rely on the linearity assumption and can identify more complex

*Funding for this project was provided in part by two grants from the NIH (5T32EB007507, R21RR026259)

Ricky Hennessy is with the Biomedical Engineering Department, The University of Texas, Austin, TX 78751 hennessy@utexas.edu

Sheldon Bish is with the Biomedical Engineering Department, The University of Texas, Austin, TX 78751 sfb@utexas.edu

James W. Tunnell is with the Biomedical Engineering Department, The University of Texas, Austin, TX 78751 jtunnell@mail.utexas.edu

Mia K. Markey is with the Biomedical Engineering Department, The University of Texas, Austin, TX 78751. She is also with the Imaging Physics Department at MD Anderson Cancer Center, Houston, Texas 77030 mia.markey@mail.utexas.edu

embeddings of the data in the high-dimensional space. However, when the number of dimensions (e.g., the number of spectral bands) is much less than the number of data points (e.g., the number of pixels), non-linear techniques have a large computational disadvantage when compared to linear techniques. In addition, this increase in computational costs is not accompanied by an improvement in performance [15].

In this work we present an algorithm for the analysis of diffuse optical reflectance hyperspectral images to detect skin cancer. The algorithm consists of principal component analysis (PCA) for dimensionality reduction, followed by k-means cluster analysis; the mean spectrum from each cluster is calculated and saved for further analysis. The limitations and properties of the proposed algorithm are tested using digital phantoms composed of spectra with known properties created using lookup table model [7].

II. BACKGROUND

A. Diffuse Reflectance Spectroscopy (DRS)

Diffuse reflectance is a function of the scattering and absorption properties of tissue, meaning DRS can be used to acquire information about the tissue morphology and function. A model-based analysis of diffuse reflectance can provide quantitative measures of the wavelength-dependent reduced scattering (μ_s) and absorption (μ_a) coefficients [7]. Scattering properties of the tissue can be used to determine size and density of the primary tissue scatterers, and the absorption properties of the tissue can be used to determine physiological parameters such as blood volume fraction oxygen saturation, blood vessel diameter, and melanin concentration. Changes in cellular structure and organization in the epidermis, changes in the extra-cellular matrix, and angiogenesis are important hallmarks of progression from normal skin to cancer. Researchers have used diffuse reflectance to optically visualize these changes and diagnose melanoma [16], [17], [18], [19] and non-melanoma skin cancers [20].

B. Hyperspectral Imaging

A hyperspectral image (HSI) is an image where each pixel represents a spectral vector. For example, an image could be collected using spectral bands with a width of $20nm$ from $400nm$ to $700nm$. Then each pixel would be a spectrum of 16 wavelength bands. The 16 wavelength bands are called the image variables [21]. Figure 1 shows a hyperspectral image composed of 16 different bands. Notice that the hyperspectral image is actually a stack of grayscale images, with each grayscale image representing a different spectral band, or dimension. Each pixel in the image can be thought of as a 16 dimensional data point.

C. Principal Component Analysis (PCA)

The goal of PCA is to find a new set of dimensions that better captures the variability of the data. The first dimension is chosen to capture as much variability as possible. The second dimension is orthogonal to the first and captures as much of the remaining variability as possible. This can continue for any number of dimensions up to the original dimensionality of the data. PCA has several characteristics that make it ideal for hyperspectral image analysis. First, it has a computational complexity of $O(N^3)$, where N is the number of dimensions in the original data. In a hyperspectral image, N is the number of spectral bands. This is advantageous for hyperspectral image data, where N is typically much less than the number of data points, or pixels. Second, most of the variability in hyperspectral images can be captured in a small number of dimensions, meaning PCA can result in low dimensional data and it may be possible to apply techniques that dont work well with high-dimensional data [15].

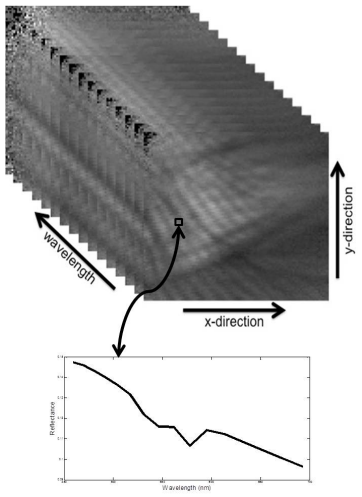

Fig. 1. Hyperspectral image collected using diffuse reflectance spectroscopic imaging. Each pixel in a hyperspectral image is a spectrum.

Principal components are computed in the following way. Given a P by N data matrix D, whose P rows are data objects and N columns are attributes. For a hyperspectral image, N is the number of spectral bands and P is the number of pixels. If the data matrix D is preprocessed so that the mean of each attribute is 0, then we can calculate the covariance matrix S, where $S = DD^T$. Let U be the matrix of eigenvectors of S. These eigenvectors are ordered such that the i^{th} eigenvector corresponds to the i^{th} largest eigenvalue. The data matrix $D' = DU$ is the set of transformed data, where the new attributes are the principal components, which are linear combinations of the original attributes. The dimensionality of the dataset can then be reduced by retaining only the first few dimensions of the transformed dataset.

D. K-Means Cluster Analysis

K-means is a clustering technique that attempts to find a user-specified number of clusters (K), which are represented by their centroids. The k-means technique works by first choosing K initial centroids, where K is a user specified parameter. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then recalculated.

Reassignment of points and centroid updating is repeated until no point changes clusters [22].

III. Methods and Materials

A. Algorithm

An algorithm involving dimensionality reduction and clustering is used to analyze the hyperspectral datasets. First the dataset is loaded into MATLAB as a hyperspectral data cube. Next, dimensionality reduction is performed by using PCA. Then, the dataset is segmented using k-means clustering. Lastly, the mean spectra from each cluster are calculated and saved for further analysis.

Dimensionality reduction was performed using PCA. The first two principal components are then retained for further analysis. More than 95% variance can be captured in the first two principal components. Skala et al. have shown that the first principal component of diffuse reflectance spectroscopy data of epithelial neoplasias is highly correlated with the reduced scattering coefficient of the tissue, and the second principal component is correlated with absorption properties of the tissue [23].

Before clustering, the two principal components that are retained are normalized to have a mean of one. This is done to ensure each principal component has equal influence on the clustering results. After normalization, k-means clustering is performed, with each pixel representing a two dimensional data point composed of the first and second principal components. The number of clusters should be based on knowledge of the tissue being imaged. This algorithm was analyzed using digital phantoms containing two different tissue types, so for our purposes, two clusters were formed. The two initial cluster centers are chosen randomly from data points within the image. Clustering is performed 3 separate times, and the clustering that gives the smallest sum of within cluster variances is used for further analysis.

After clustering, the mean cluster spectra can be calculated. This is done by taking the mean of all spectra from the original data set within each cluster. Because many spectra are averaged, much of the noise within each individual spectrum is cancelled out and we are left with low-noise mean cluster spectra that can be related to biochemical properties of tissue areas. These properties can be calculated using the method described by Rajaram et al. [7] and the computational time is no longer an issue since the number of spectra to analyze is reduced by over three orders of magnitude.

B. Algorithm Analysis

Figure 2 describes the process used to determine the relationship between the amount of noise in the hyperspectral images and the minimum difference in scattering that the algorithm is able to detect without misclassifying more than 5% of the pixels in the segmented image. The pixel misclassification percentage is calculated by comparing the segmentation results after the addition of noise to the ideal segmentation results. The number of pixels that differ between the two segmented images is then divided by the total number of pixels. When more than 5% of the pixels are misclassified, it is no longer possible to accurately calculate the optical properties based on the mean spectra from each cluster.

A lookup table (LUT) was used to generate spectra. The LUT was created by measuring the functional form of the reflectance using tissue phantoms with known optical properties. These phantoms were fabricated using polystyrene microspheres and India ink dissolved in water to simulate scattering and absorption respectively [7]. Mie theory was used to calculate μ_s of the tissue phantoms and μ_a was measured using a spectrophotometer. A matrix $(4x6)$ of 24 tissue phantoms with varying scattering and absorption parameters was created. The probe was placed in contact with the surface of the tissue phantoms, and white-light spectra from the phantoms were recorded. Reflectance was calculated by dividing white-light intensity measured from the phantom by the intensity from a reflectance standard. The sparse matrix was then interpolated to create a grid of uniformly spaced data points of s and a to obtain a LUT for diffuse reflectance spectra.

A hyperspectral image ($100x100$ pixels) was simulated from the spectra generated using the forward model created. Two different spectra were used, with one having a higher scattering coefficient than the other. A circular area with a radius of 25 pixels contained the spectrum with the higher scattering coefficient, and the rest of the image contained the other spectrum. After creation of the hyperspectral image, noise from a normal distribution was added to the image. The main advantage of using digital phantoms is that the level of noise could be selected by the investigator.

Figure 2 outlines the process used to determine the relationship between the amount of noise in the data and the minimum difference between spectra that the algorithm is able to detect without misclassifying more than 5% of the pixels in the segmented image. This was done by creating 40 different phantoms where the difference in scattering between the two regions ranged from $0.05mm^{-1}$ to $2.00mm^{-1}$. Noise from a normal distribution of mean zero was then added to the images starting with a standard deviation of 0.001 and the noise was increased in increments of 0.001 until more than 5% of the pixels in the segmented image were misclassified. For each level of scattering difference and noise, the process was repeated 10 times and the percentage of misclassified pixels was averaged. This was necessary to account for the addition of random noise.

IV. Results and Discussion

This paper describes an algorithm for the segmentation of diffuse reflectance hyperspectral datasets with noise. The algorithm is able to rapidly segment the data into similar clusters and is also robust to noise. The algorithm segments the hyperspectral images by first reducing the number of dimensions with PCA and then clustering using K-Means clustering.

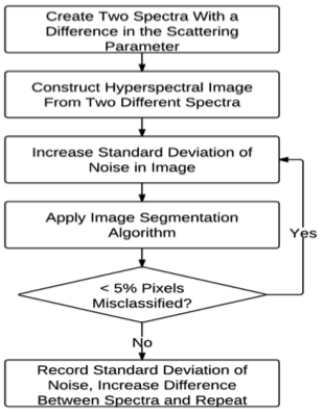

Fig. 2. Flowchart describing the process used to determine the maximum amount of noise that still allows for an accurate segmentation.

Digital phantoms were created and noise was added to determine the effect of noise on the ability of the algorithm to distinguish between tissue types. The analysis showed a linear relationship between the amount of noise in the images and the minimum detectable difference in scattering between the two tissue segments. It has been shown that the difference in μ_s between normal and cancerous human skin in the visible range is approximately 0.15 [24]. We can use this information and our findings to determine the maximum level of noise in the data that will still allow the detection of skin cancer.

The proposed algorithm is able to decrease the time to extract chemical information from the spectra by three orders of magnitude compared to methods that analyze every spectrum in the hyperspectral dataset. This is done by reducing the number of spectra to be fitted, and allows results to be obtained in a few seconds. Another advantage is that the effect of noise in the individual spectra from the original dataset is greatly reduced by analyzing the average spectrum from each cluster. By averaging the thousands of spectra within each cluster, most of the noise is eliminated.

In future work, we will investigate this algorithm on real datasets containing melanocytic regions. Segmented images will be created and the mean spectra from each cluster in the segmented image will be analyzed using an inverse model lookup table approach. This will allow the extraction of optical properties which can be used to quantify relevant physiological parameters such as blood volume fractions, scattering, and hemoglobin concentration. With the collection of a large dataset, these parameters can be used to develop a classifier that will assign a probability of disease to each cluster in the segmented image.

REFERENCES

[1] A. Jemal, R. Siegel, J. Q. Xu, and E. Ward, "Cancer Statistics, 2010," Ca-a Cancer Journal for Clinicians, vol. 60, pp. 277-300, 2010.

[2] D. Gareau, R. Hennessy, E. Wan, G. Pellacani, and S. L. Jacques, "Automated detection of malignant features in confocal microscopy on superficial spreading melanoma versus nevi," Journal of Biomedical Optics, vol. 15, pp. 06173, 2010.

[3] N. Rajaram, J. S. Reichenberg, M. R. Migden, T. H. Nguyen, and J. W. Tunnell, "Pilot Clinical Study for Quantitative Spectral Diagnosis of Non-Melanoma Skin Cancer," Lasers in Surgery and Medicine, vol. 42, pp. 716-727, 2010.

[4] Y. M. Wang, S. Bish, J. W. Tunnell, and X. J. Zhang, "MEMS scanner enabled real-time depth sensitive hyperspectral imaging of biological tissue," Optics Express, vol. 18, pp. 24101-24108, 2010.

[5] L. Gorlitz, B. H. Menze, B. M. Kelm, and F. A. Hamprecht, "Processing spectral data," Surface and Interface Analysis, vol. 41, pp. 636-644, 2009.

[6] M. Miljkovic, T. Chernenko, M. J. Romeo, B. Bird, C. Matthaus, and M. Diem, "Label-free imaging of human cells: algorithms for image reconstruction of Raman hyperspectral datasets," Analyst, vol. 135, pp. 2002-2013, 2010.

[7] N. Rajaram, T. H. Nguyen, and J. W. Tunnell, "Lookup table-based inverse model for determining optical properties of turbid media," Journal of Biomedical Optics, vol. 13, 2008.

[8] B. Bird, M. Miljkovic, M. J. Romeo, J. Smith, N. Stone, M. W. George, and M. Diem, "Infrared micro-spectral imaging: distinction of tissue types in axillary lymph node histology.," BMC Clin Pathol, vol. 8, p. 8, 2008.

[9] M. Hanselmann, U. Kothe, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. A. Heeren, and F. A. Hamprecht, "Toward Digital Staining using Imaging Mass Spectrometry and Random Forests," Journal of Proteome Research, vol. 8, pp. 3558-3567, 2009.

[10] M. Isabelle, K. Rogers, and N. Stone, "Correlation mapping: rapid method for identification of histological features and pathological classification in mid infrared spectroscopic images of lymph nodes," Journal of Biomedical Optics, vol. 15, 2010.

[11] R. Jolivot, P. Vabres, and F. Marzani, "Reconstruction of hyperspectral cutaneous data from an artificial neural network-based multispectral imaging system," Computerized Medical Imaging and Graphics, vol. 35, pp. 85-88, 2011.

[12] B. Wood, L. Chiriboga, H. Yee, M. Quinn, D. McNaughton, and M. Diem, "Fourier transform infrared (FTIR) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium," Gynecologic Oncology, vol. 93, pp. 59-68, 2004.

[13] B. Andreopoulos, A. J. An, X. G. Wang, and M. Schroeder, "A roadmap of clustering algorithms: finding a match for a biomedical application," Briefings in Bioinformatics, vol. 10, pp. 297-314, 2009.

[14] J. Wang and C. I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," Ieee Transactions on Geoscience and Remote Sensing, vol. 44, pp. 1586-1600, 2006.

[15] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality Reduction: A Comparative Review," vol. 2009, ed. Tilburg Univeristy Technical Report, 2009.

[16] G. W. Juette and L. E. Zeffanella, Radio noise currents n short sections on bundle conductors (Presented Conference Paper style), presented at the IEEE Summer power Meeting, Dallas, TX, June 2227, 1990, Paper 90 SM 690-0 PWRS.

[17] J. G. Kreifeldt, An analysis of surface-detected EMG as an amplitude-modulated noise, presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.

[18] J. Williams, Narrow-band analyzer (Thesis or Dissertation style), Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

[19] N. Kawasaki, Parametric study of thermal and chemical nonequilibrium nozzle flow, M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

[20] J. P. Wilkinson, Nonlinear resonant circuit devices (Patent style), U.S. Patent 3 624 12, July 16, 1990.

[21] H. F. Grahn and P. Geladi, "Techniques and Applications of Hyperspectral Image Analysis," West Sussex, England: John Wiley and Sons Ltd, 2007.

[22] A. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, pp. 651-666, 2010.

[23] M. C. Skala, G. M. Palmer, K. M. Vrotsos, A. Gendron-Fitzpatrick, and N. Ramanujam, "Comparison of a physical model and principal component analysis for the diagnosis of epithelial neoplasias in vivo using diffuse reflectance spectroscopy.," Opt Express, vol. 15, pp. 7863-75, 2007.

[24] E. Salomatina, B. Jiang, J. Novak, and A. N. Yaroslavsky, "Optical properties of normal and cancerous human skin in the visible and near-infrared spectral range.," J Biomed Opt, vol. 11, p. 064026, 2006.