# K-means and Hierarchical Cluster Analysis as segmentation algorithms of FTIR hyperspectral images collected from cutaneous tissue

Cassio Lima
*Center for Lasers and Applications*,
*IPEN/CNEN-SP*
*Universidade de Sao Paulo*
Sao Paulo, Brazil
cassiolima@usp.br

Luciana Correa
*School of Dentistry*
*Universidade de Sao Paulo*
Sao Paulo, Brazil
lcorrea@usp.br

Hugh Byrne
*Facility for Optical Characterisation and Spectroscopy - FOCAS*
*Dublin Institute of Technology*
Dublin, Ireland
hugh.byrne@dit.ie

Denise Zezell
*Center for Lasers and Applications,*
*IPEN/CNEN-SP*
*Universidade de Sao Paulo*
Sao Paulo, Brazil
zezell@usp.br

***Abstract*— Fourier Transform Infrared (FTIR) spectroscopy is a rapid and label-free analytical technique whose potential as a diagnostic tool has been well demonstrated. The combination of spectroscopy and microscopy technologies enable wide-field scanning of a sample, providing a hyperspectral image with tens of thousands of spectra in a few minutes. In order to increase the information content of FTIR images, different clustering algorithms have been proposed as segmentation methods. However, systematic comparative tests of these techniques are still missing. Thus, the present paper aims to compare the ability of K-means Cluster Analysis (KMCA) and Hierarchical Cluster Analysis (HCA) as clustering algorithms to reconstruct FTIR hyperspectral images. Spectra for cluster analysis were acquired from healthy cutaneous tissue and the pseudo-color reconstructed images were compared to standard histopathology in order to assess the number of clusters required by both methods to correctly identify the morphological skin components (stratum corneum, epithelium, dermis and hypodermis).**

***Keywords*— FTIR microspectroscopy; Image segmentation; KMCA; HCA***

## I. INTRODUCTION

Fourier Transform Infrared (FTIR) spectroscopy is a rapid and label-free analytical technique based on the active infrared vibrations to interrogate the chemical profile of a sample. Biological molecular bonds with an electric dipole moment that can change by atomic displacement due to natural vibrations are infrared active and therefore are quantitatively measured by infrared spectroscopy [1, 2]. Biochemical information retained by spectral data have been fully succeeded discriminating cancer from benign tissue in a non-subject manner, indicating the potential of the technique as a diagnostic tool [2-6].

In the last decades, the combination of spectroscopy and microscopy technologies enabled wide-field scanning of a sample, providing a hyperspectral image with tens of thousands of spectra in a few minutes [1]. Spectral images provide spatial and compositional information similar to the results obtained by immunohistochemistry [3]. However, FTIR does not require staining and provide signatures containing molecular interactions occurring between biomolecules and not only on a single one as obtained using immunohistochemical imaging technology [7].

Different clustering algorithms have been proposed as segmentation methods aiming to increase the information content of FTIR images [8]. In general, spectral data are used as input for some clustering method in order to partitionate the dataset into sub-groups (clusters) of spectra with similar spectral characteristics. The differences between the data within each cluster are minimized, whereas the differences between clusters are maximized. At the end of the segmentation process, each cluster represents regions of the image with similar biochemical profiles.

Several multivariate pattern recognition methods have been proposed as segmentation methods to hyperspectral images in order to differentiate distinct tissue structures, as well as for identifying tissue pathologies [1, 8, 9]. However, systematic comparative tests of these techniques are still missing. Thus, the present paper aims to compare the ability of KMCA and HCA as clustering algorithms to reconstruct FTIR hyperspectral images. Spectra for cluster analysis were acquired from healthy cutaneous tissue and the results of the distinct cluster imaging techniques are compared to standard histopathology.

## II. MATERIAL AND METHODS

### A. Sample preparation

Cutaneous tissue were extracted from healthy Swiss mice aged up to 10 weeks and body mass about to 25 g after approval by the institutional Committee on Animal Research and Ethics of Instituto de Pesquisas Energeticas e Nucleares (IPEN/CNEN-SP, n° 164/15). Specimens were kept in formalin fixation for 24 h and paraffin-embedded

978-1-5386-6702-6/18/$31.00©2018 IEEE

(FFPE). Longitudinal sections of hematoxylin-eosin stained tissue with 5 µm of thickness were placed on calcium fluoride crystal slides for spectroscopic measurements (Crystran, Poole, Dorset, UK).

### B. *FTIR microspectroscopy*

FTIR hyperspectral images were acquired in transmission mode using a Spotlight 400N FTIR imaging system (Perkin Elmer, Waltham, MA, USA) equipped with an AutoImage microscope system operating with a ×40 Cassegrain objective. Pijanka *et al* have studied the effects of H&E staining on the infrared spectra of cells and demonstrated the appearance of a new band at 1378 $cm^{-1}$ and the disappearance of the bands peaking at 2850 $cm^{-1}$ and 2920 $cm^{-1}$. The new emerging band was attributed to the staining, while the removal of bands was associated to the ethanol washings during the staining procedure [10]. In light of this, spectral measurements were acquired over the spectral range 3000-4000 $cm^{-1}$ in order to avoid spectral changes associated to the staining procedure. Spectral data were recorded with a pixel size of 6.25 µm × 6.25 µm at a spectral resolution of 4 $cm^{-1}$. Background measurements were acquired on a region with no tissue with 120 scans per pixel whereas 8 scans per pixel were recorded from the sample.

### C. *Computational data analysis*

The methods used to preprocess the spectral data may result different reconstructed images. Thus, in order to avoid alterations not related to the clustering algorithm, the data used as input were equally preprocessed. Spectra were cut over the range of 3100-3800 $cm^{-1}$, vector normalized and submitted to noise reduction and atmospheric correction. KMCA and HCA were applied to the hyperspectral images using CytoSpec software package (CytoSpec, Berlin, Germany). Pseudo-color maps were reconstructed using both algorithms varying the number of clusters and compared to photomicrographs of hematoxylin-eosin stained tissue. Each clustering algorithm requires specific internal parameters to reconstruct the images, so that different results may be expected varying these parameters. Both methods are based on the statistical distance calculated from data as statistical measure. Thus, in order to avoid changes associated to the internal parameters of each method, Euclidean distance was used as statistical measure in both algorithms. Centroid random initialization was used to reconstruct the images using KMCA with 100 iterations of training cycles. Ward's algorithm was used as linkage method for hierarchical clustering method.

### D. *Histopathological architecture of skin*

The skin is the largest organ of the body and it is composed by three layers: epidermis, dermis and hypodermis, which are depicted in Figure 1A by the numbers 3, 2 and 1, respectively. Hypodermis is the deeper subcutaneous tissue and it is made of fat and connective tissue. Dermis, the middle layer, is composed by connective tissue, hair follicles and sweet glands. Epidermis, the outermost layer of skin, is made of two layers: Epithelium (5 in Figure 1B) and stratum corneum (6 in Figure 1B). In general, the epithelium is composed by 3 or 4 keratinocyte layers.

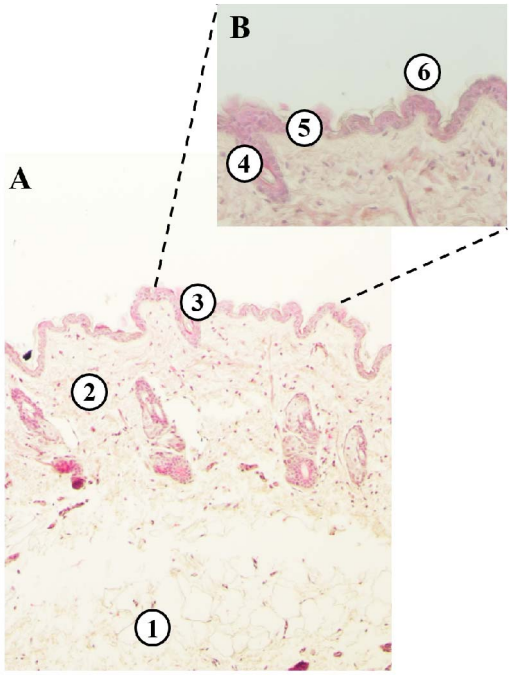


Fig. 1. Representative histopathology aspect of healthy skin. **A:** Healthy skin (10× original magnification): Adipose tissue (1), dermis (2), epithelium (3). **B:** Epithelium (30× original magnification): Keratinocyte layer (4), hair follicle (5) and stratum corneum (6).

### E. *FTIR image of skin (chemical map)*

Figure 2A depicts a functional group mapping (chemical map) of a hyperspectral image collected on transmission mode. Each pixel represent a spectrum collected (Figure 2B) over 3100-3800 $cm^{-1}$ spectral range. The different colors observed in false color map are associated to the different values of total absorbance under the curve.

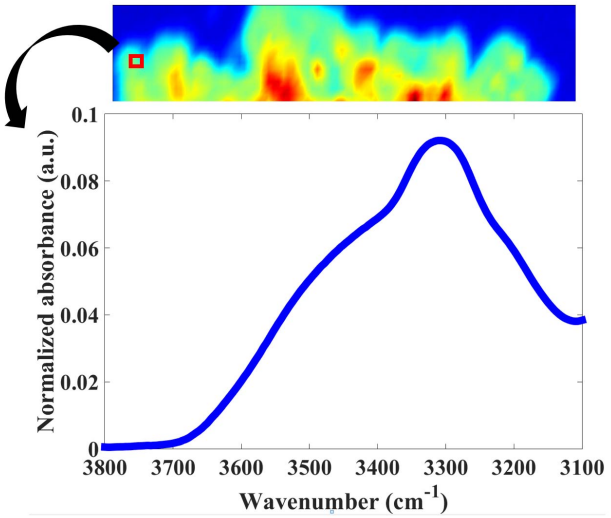


Fig. 2 Normalized FTIR spectra obtained from a pixel of hyperspectral image

Although it is possible to distinguish regions of tissue and substrate in the hyperspectral image, the skin components are not evidenced.

### F. *Cluster imaging using KMCA*

KMCA is one of the simplest unsupervised learning algorithms often used for spectral image analysis [11].

KMCA is a nonhierarchical clustering technique which uses an iterative algorithm to update randomly initial cluster centers. Objects are initially assigned to primary clusters according to the minimal distance values. In the following, cluster centroids are calculated and the distance between the centroids and each of the objects are recalculated. The object is relocated to the cluster with the closest centroid. The centroid's positions are recalculated every time that a component change its cluster membership. This process continues until none of the objects has been reassigned [8].

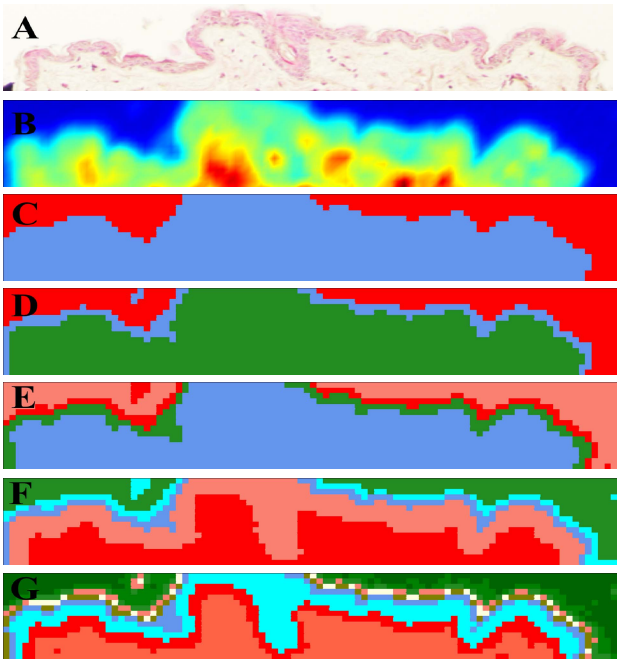


Fig. 3. **A:** H&E-stained specimen. **B:** Hyperspectral image. **C-G:** IR images reconstructed by KMCA clustering algorithm varying the number of clusters (2, 3, 4, 5 and 10, respectively).

Figure 3.C-G depicts the images assembled by KMCA clustering varying the number of clusters (2, 3, 4, 5, and 10, respectively). A photomicrograph of the H&E-stained tissue is also shown in Figure 3.A for comparison with histopathology, as well as a chemical map (Figure 3.B).

### *G. Cluster imaging using HCA*

HCA is an unsupervised and hierarchical clustering method which attributes color codes to spectral clusters based on the similarity of all spectra in the dataset. First, a distance matrix between all spectra is calculated. In our case, Euclidian distance was used to compare the spectra. Two similar objects are merged into a cluster and, subsequently, the distances of the new clusters are recalculated and then merged into a new cluster according to their similarity. At the end of the algorithm, there will be only one cluster.

Pseudo-color images reconstructed using HCA varying the number of clusters (2, 3, 4, 5, and 10) are depicted in the Figure 4.C-G. Again, a photomicrograph of the H&E-stained specimen and chemical map are shown.

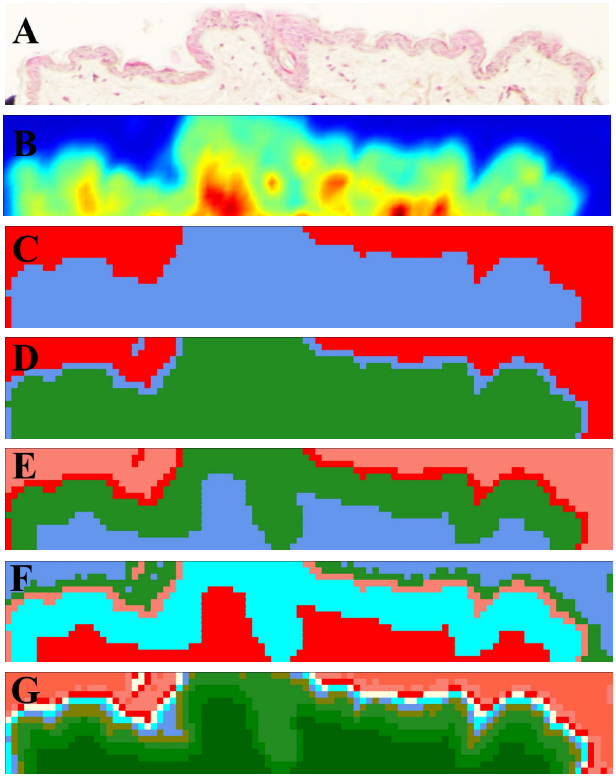


Fig. 4. **A:** H&E-stained specimen. **B:** Hyperspectral image. **C-G:** IR images reconstructed by HCA clustering algorithm varying the number of clusters (2, 3, 4, 5 and 10, respectively).

### *H. Correlation of spectral and histopathological images*

At a first glance, the results of FTIR microspectroscopy and light microscopy are highly correlated. Reconstructed images using the lowest number of clusters already enable the identification of regions with tissue and substrate (Figures 3C and 4C). However, with no significant difference between both images.

Increasing the number of clusters in the segmented images provide the identification of histological cutaneous structures. In Figures 3D and 4D, 3 clusters are used to built the images, in which green pixels encode dermis, blue pixels are epithelium and red represents glass with no tissue. Figures 3E and 4E (4 clusters) enable the identification of the keratin layer (dark red for both images), dermis (blue regions), epithelium (green) and regions of glass with no tissue (soft red). Although both methods are able to identify the skin structures, it is possible to observe a discrepancy in the clusters assigned to the epithelium and dermis layers. The crypt in the epithelium presented by the image HCA reconstructed represents the hair follicle identified as number 4 in the Figure 1A, which is only observed in the image segmented by KMCA using 5 clusters. In addition, the green cluster from image reconstructed by HCA is thicker than green cluster obtained by KMCA image. The assembled images using 5 and 10 clusters presented high degree of correlation.

## III. Conclusions

In this study, KMCA and HCA were used as clustering algorithms to reconstruct FTIR hyperspectral images collected from healthy cutaneous tissue. Pseudo color images were compared to standard histopathology in order

to assess the number of clusters required by both methods to correctly identify the morphological skin components. Both methods presented high correlation with specimen photomicrograph and increased the information content of IR images. Regarding tissue structure differentiation, HCA clustering method showed better results using a less number of clusters.

## ACKNOWLEDGMENT

This study was supported by CEPID-FAPESP 05/51689-2, CNPq (INCT-465763/2014-6, PQ-309902/2017-7, PhD grant-141629/2015-0), CAPES (PROCAD-88881.068505/2014-01, PDSE-88881.132771/2016-01).

## REFERENCES

[1] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sule-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, and F. L. Martin, "Using Fourier transform IR spectroscopy to analyze biological materials," *Nat Protoc,* vol. 9, no. 8, pp. 1771-91, Aug, 2014.

[2] C. A. Lima, V. P. Goulart, L. Correa, T. M. Pereira, and D. M. Zezell, "ATR-FTIR spectroscopy for the assessment of biochemical changes in skin due to cutaneous squamous cell carcinoma," *Int J Mol Sci,* vol. 16, no. 4, pp. 6621-30, 2015.

[3] M. Diem, A. Mazur, K. Lenau, J. Schubert, B. Bird, M. Miljkovic, C. Krafft, and J. Popp, "Molecular pathology via IR and Raman spectral imaging," *J Biophotonics,* vol. 6, no. 11-12, pp. 855-86, Dec, 2013.

[4] C. A. Lima, V. P. Goulart, L. Correa, and D. M. Zezell, "Using Fourier transform infrared spectroscopy to evaluate biological effects induced by photodynamic therapy," *Lasers Surg Med,* vol. 48, no. 5, pp. 538-45, Jul, 2016.

[5] P. Bassan, J. Mellor, J. Shapiro, K. J. Williams, M. P. Lisanti, and P. Gardner, "Transmission FT-IR chemical imaging on glass substrates: applications in infrared spectral histopathology," *Anal Chem,* vol. 86, no. 3, pp. 1648-53, Feb 04, 2014.

[6] H. J. Byrne, M. Baranska, G. J. Puppels, N. Stone, B. Wood, K. M. Gough, P. Lasch, P. Heraud, J. Sule-Suso, and G. D. Sockalingum, "Spectropathology for the next generation: quo vadis?," *Analyst,* vol. 140, no. 7, pp. 2066-73, Apr 07, 2015.

[7] S. Kalmodia, S. Parameswaran, W. Yang, C. J. Barrow, and S. Krishnakumar, "Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy: An analytical technique to understand therapeutic responses at the molecular level," *Sci Rep,* vol. 5, pp. 16649, Nov 16, 2015.

[8] P. Lasch, W. Haensch, D. Naumann, and M. Diem, "Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis," *Biochim Biophys Acta,* vol. 1688, no. 2, pp. 176-86, Mar 2, 2004.

[9] H. J. Byrne, P. Knief, M. E. Keating, and F. Bonnier, "Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells," *Chem Soc Rev,* vol. 45, no. 7, pp. 1865-78, Apr 07, 2016.

[10] J. K. Pijanka, N. Stone, G. Cinque, Y. Yang, A. Kohler, K. Wehbe, M. Frogley, G. Parkes, J. Parkes, P. Dumas, C. Sandt, D. G. van Pittius, G. Douce, G. D. Sockalingum, and J. Sulé-Suso, "FTIR microspectroscopy of stained cells and tissues. Application in cancer diagnosis.," *Spectroscopy,* vol. 24, pp. 73–78, 2010.

[11] S. M. Ali, F. Bonnier, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T. C. Lee, F. M. Lyng, and H. J. Byrne, "A comparison of Raman, FTIR and ATR-FTIR micro spectroscopy for imaging human skin tissue sections," *Analytical Methods,* vol. 5, no. 9, 2013.