doi:10.1017/S1431927611003692

Automated Classification of SIMS Images Using K-means Clustering

A.R. Konicek,* J. Lefman,*† and C. Szakal*

* Surface and Microanalysis Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899

† Now at the Army Corps of Engineers

Computational analysis of hyperspectral secondary ion mass spectrometry (SIMS) image data provides far better insight and interpretation than can be gained from a manual inspection. Multivariate analysis routines such as principal component analysis (PCA) [1-3] characterize the data by looking for the maximum variance in the spectral dimension of a 3D data set. However, the resulting multivariate component spectra are mathematical constructs of the original data that have a nontrivial relationship to the spatial differentiation within mass spectral images.

We present a new method for SIMS data analysis that focuses on classifying SIMS data an approach designed specifically for image analysis. Raw data images are processed using a *k*-means clustering analysis. The outputs are clusters of unique pixels, called centroids, which are average groups of images that have a minimized Euclidean distance to their centroid. The analysis then correlates all of the images to every centroid, and classifies the images based on the highest correlation value. This approach groups images with like spatial distribution, and therefore similar chemical constituents. Since each image is unique to a particular mass value, composite images and spectra can be created to represent the different constituents that create each spatially distinct component.

An example data set was created by laserjet printing of ink from CMYK cartridges in standardized patterns (FIG. 1). Each pattern includes a single region of pure cyan, magenta, and yellow, and then each includes a mixture of the three to produce red, green, and/or blue. There are also regions in each image that do not have ink and are from the paper. SIMS data was taken in negative ion mode, and subsets of the extracted mass-specific images reflect the different chemical (and therefore spatial) composition of the inks (FIG. 2).

We will provide detailed information on the algorithm performance for the aforementioned dataset as a proof-of-concept. Preliminary data for more challenging datasets will be presented to illustrate the broad utility of this image-based data processing tool. The methodology can be theoretically be extrapolated to any hyperspectral image data set.

References

[1] M.S. Wagner, D.G. Castner, *Appl. Surf. Sci.*, 203 (2003) 698.
[2] M.R. Keenan, P.G. Kotula, *Surf. Interface Anal.* 36 (2004) 203.
[3] V.S. Smentkowski *et al.*, *Anal. Chem.* 77 (2005) 1530.

https://doi.org/10.1017/S1431927611003692 Published online by Cambridge University Press

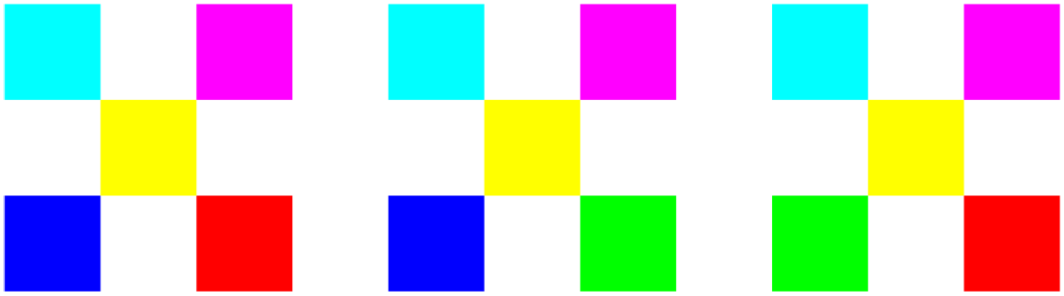

FIG. 1. Schematic of the ink printing shapes. Colors are cyan and magenta on top, yellow in middle, and then two colors of blue, red, and/or green. Overall printed shape is roughly 500 μm x 500 μm.

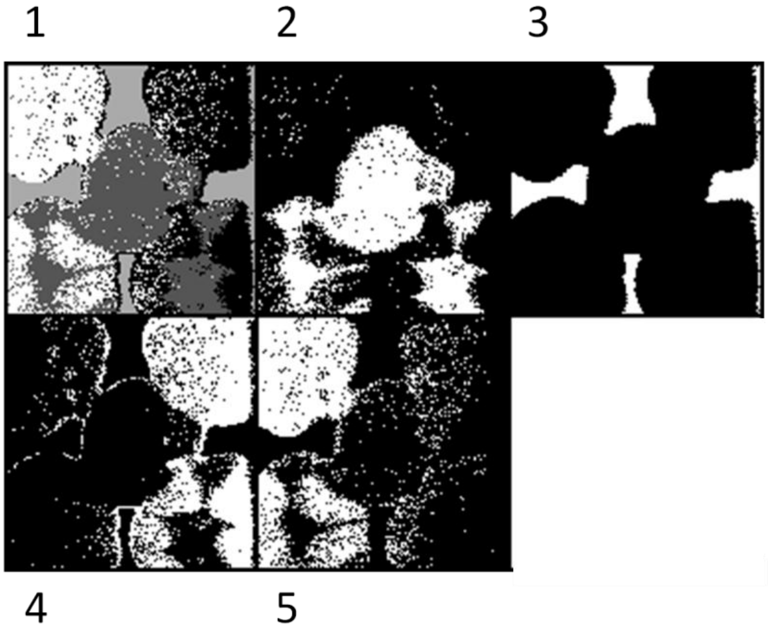

FIG. 2. Output centroids from k-means analysis of the left pattern in FIG. 1. Panel 1 is the combination of the four centroids with different gray level values for each centroid. The remaining panels are centroids that represent the yellow (2), paper (3), magenta (4), and cyan (5). It can be seen that the yellow is in both of the lower colored regions as it is a CMYK component of both blue (cyan and yellow) and red (yellow and magenta).

https://doi.org/10.1017/S1431927611003692 Published online by Cambridge University Press