REDUCING THE DIMENSIONALITY OF HYPERSPECTRAL DATA USING DIFFUSION MAPS

Louis du Plessis[*]*, Steven Damelin*[†‡] *and Michael Sears*[*]

[*]School of Computer Science
University of the Witwatersrand
Johannesburg, South Africa

[†]Department of Mathematical Science,
Georgia Southern University, Statesboro
GA 30460, USA

[‡]School of Computational and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa

1. INTRODUCTION

Core samples are long cylindrical pieces of rock drilled in prospective mining sites. The decision to mine or not is dependent on the analysis of core samples. AngloGold Ashanti has developed the Hyperspectral Core Imager (HCI) to produce hyperspectral images of core samples. By analyzing the spectra of different points in the resulting image it is possible to determine the mineral composition of a core sample. However, an accurate automated method remains elusive, due to the complexity and the size of the data.

The core supposition of this paper is that the data produced by the HCI is highly redundant. The HCI produces images with 640 bands, using three different spectrometers with overlapping ranges. Because the spectrometers are not exactly spatially aligned, it is not possible to use all the bands at the same time. Instead, only the bands from the spectrometer located in the short-wave infrared area (the area most sensitive to minerals) are used. The resulting data is very noisy, suggesting that it can be considerably simplified. The redundancy in the data is exploited by reducing the dimensionality further using a diffusion map. An unsupervised clustering method is used to partition the resulting lower-dimensional image into its constituent mineral clusters. By embedding template spectra in the original higher-dimensional data, the dominant mineral in each cluster may be identified. This paper builds on research published in [1] and [2].

2. DIFFUSION MAPS

Diffusion maps provide a nonlinear method for reducing the dimensionality of a dataset, while preserving local connectivity. A Markov random walk is defined on the data, and the diffusion distance is a metric on the resulting Markov matrix. The diffusion distance compares all possible paths between the two points being compared, making it very robust to noise [3].

The diffusion distance can be written as a function of the eigenvalues and eigenvectors of the resulting Markov matrix. The distance can be accurately approximated by only using the largest eigenvalues and corresponding eigenvectors. A diffusion map maps the data to a space with dimension equal to the number of eigenvalues and eigenvectors used to approximate the diffusion distance. This space also has the property that the ordinary Euclidean distance is equivalent to the diffusion distance in the original high-dimensional representation.

More information on diffusion maps can be found in [3] and [4].

The authors would like to thank AngloGold Ashanti as well as the Anglo Technical Division, Geoscience Research Group. Louis du Plessis and Michael Sears would like to acknowledge support from the School of Computer Science at the University of the Witwatersrand. Steven Dameilin's research was funded in part by the NSF, EPRSC and the School of Computational and Applied Mathematics at the University of the Witwatersrand.

(a) One band of the original image.

(b) Clustering produced.

Fig. 1. An example clustering produced by separately reducing the dimension of three sub-images and clustering each sub-image using k-means. The results were merged to produce this image.

3. EXPERIMENTAL RESULTS

Because the diffusion map essentially squares the amount of data, it is at present infeasible to perform the process on complete datasets. To remedy this problem the dataset is divided into overlapping sub-images, which are individually transformed into their lower-dimensional representations. The composition of clusters are found by embedding template spectra of all expected minerals into the untransformed higher-dimensional data. By correlating the spatial location of clusters and the composition of clusters, the sub-images can be merged to produce a map of the complete dataset.

At present, k-means [1] and fuzzy ART [2] have been used to cluster the lower dimensional representation of the data. The results are promising and consistent with expectations and results obtained by AngloGold's proprietary method. An example is given in figure 1. This image was produced by performing a diffusion map on three overlapping sub-images. The dimension of the data was reduced from 100 to 20 dimensions. The three resulting 20-dimensional representations were then clustered using k-means, and the results merged to produce the image given in figure 1.

4. REFERENCES

[1] K. Cawse, S. Damelin, L. du Plessis, R. McIntyre, M. Mitchley, and M. Sears, "An investigation of data compression techniques for hyperspectral core imager data," in *Proceedings of the Mathematics in Industry Study Group, South Africa – MISG2008*, To appear.

[2] Rui Xu, Louis du Plessis, Steven Damelin, Michael Sears, and Donald C. Wunsch II, "Analysis of Hyperspectral Data with Diffusion Maps and Fuzzy ART," in *Proceedings of the 2009 International Joint Conference on Neural Networks*, To appear.

[3] Ronal R. Coifman and Stephane Lafon, "Diffusion maps," *Journal of Applied and Computational Harmonic Analysis*, pp. 5–30, April 2006.

[4] Stephane Lafon and Ann B. Lee, "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning and data set parameterization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1393–1403, September 2006.