COMPRESSION-BASED UNSUPERVISED CLUSTERING OF SPECTRAL SIGNATURES

D. Cerra, J. Bieniarz, J. Avbelj, P. Reinartz, and R. Mueller

German Aerospace Center (DLR)
Earth Observation Center (EOC)
Muenchner str. 20, 82234 Wessling, Germany

ABSTRACT

This paper proposes to use compression-based similarity measures to cluster spectral signatures on the basis of their similarities. Such universal distances estimate the shared information between two objects by comparing their compression factors, which can be obtained by any standard compressor. Experiments on rocks categorization show that these methods may outperform traditional choices for spectral distances based on vector processing.

Index Terms— Spectral distance, similarity measure, data compression

1. INTRODUCTION

The processing of hyperspectral images for detection and classification purposes often relies on estimating the similarities between spectra, represented by vectors composed of the values in each image element (or pixel) across all the spectral bands. Spectral matching has at its core the use of a distance measure as a mean to quantify the distance between any pair of such spectra. Among the adopted measures, often having their origins in vector processing, popular choices are the Euclidean distance (ED), the Spectral Angle (SA) [1], the Spectral Correlation (SC) [2], and the Spectral Information Divergence (SID) [3]. The performances of these spectral distances have been compared in [4] and [5], with both works agreeing on considering SID as a slightly more discriminative distance among the mentioned ones.

This paper proposes to use compression-based similarity measures as a valid alternative to quantify the similarity between spectral signatures. These measures employ general off-the-shelf compressors in an unusual way, by exploiting them to estimate the amount of information shared by two objects. They can be employed for clustering and classification on diverse data types, outperforming general distance measures [6]. Experiments on satellite images using these techniques have been presented in [7].

To assess the quality of the distances obtained with the proposed method we perform an unsupervised hierarchical clustering with all the distances mentioned above on a set of spectral signatures, collected on the field and related to different kinds of rocks. Results suggest that compression-based methods could outperform traditional similarity measures employed in spectral matching at capturing similarities between the spectra which could be not obvious at a first inspection.

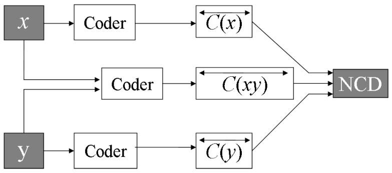

Fig. 1. Computation of a distance between two general objects x and y by means of a standard compressor C. The sizes of the objects compressed separately and jointly are compared, yielding a distance ranging from 0 to 1.

The work is structured as follows. Section 2 introduces the proposed Normalized Compression Distance (NCD), while Section 3 presents a brief reminder on well-known spectral distances. Section 4 reports experiments on rocks categorization. We conclude in Section 5.

2. NORMALIZED COMPRESSION DISTANCE

The most widely known and used compression based similarity measure for general data is the Normalized Compression Distance (NCD), defined for any two objects x and y as:

$$NCD(x,y) = \frac{C(x,y) - \min\{C(x),C(y)\}}{\max\{C(x),C(y)\}}, \qquad (1)$$

where $C(x)$ represents the size of x after being compressed by a general off-the-shelf compressor (such as Gzip), and $C(x,y)$ is the size of the compressed version of x appended to y (Fig. 1). The NCD ranges approximately from 0 to 1, representing maximum and minimum similarity, respectively. The idea is that if x and y share common information they will compress better together than separately, as the compressor will be able to reuse recurring patterns found in one of them to more efficiently compress the other. One of the main advantages of such distance is its parameter-free approach, which makes it applicable to diverse data types [8], as the NCD only depends on the compressor adopted and its internal parameters, with performance comparisons for general compression algorithms showing this dependance to be loose [9]. To compute the NCD between two spectra we apply a compressor belonging to the lz-family [10] to the spectra converted into ASCII text files.

3. SPECTRAL DISTANCES

In the following definitions, unless otherwise stated, x and y are assumed to be n-dimensional vectors representing spectra, with n being the number of bands in each spectrum.

3.1. Euclidean Distance

The Euclidean Distance (ED) quantifies the distance between two vectors x and y as:

$$ED(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

As simple as it may be, this distance often gives the best results in several data mining problems [6].

3.2. Spectral Angle

The Spectral Angle (SA) measures the angle between two spectra [1]:

$$SA(x,y) = \cos^{-1} \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2}} \quad (3)$$

3.3. Spectral Correlation

We compute the Spectral Correlation (SC) between two spectra x and y as:

$$SC(x,y) = \sqrt{\frac{1 - r(x,y)}{2}}, \quad (4)$$

where $r(x,y)$ is the correlation between x and y:

$$r(x,y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad (5)$$

with σ_{xy} being the covariance between x and y, and σ_x and σ_y the standard deviations of x and y [2].

3.4. Spectral Information Divergence

The spectral information divergence (SID) [3] derives from information theory notions. If we consider two spectra x and y as two probability distributions $p_x(i)$ and $p_y(i)$, the SID is given by:

$$SID(x,y) = d(p_x(i)||p_y(i)) + d(p_y(i)||p_x(i)), \quad (6)$$

where

$$d(p_x(i)||p_y(i)) = \sum_{i=1}^{n} p_x(i) \log \frac{p_x(i)}{p_y(i)}. \quad (7)$$

4. EXPERIMENTAL RESULTS

We tested the discriminative power of the previously introduced distances on a set of spectral signatures, chosen from different materials divided into some categories.

For this purpose, from the ASTER 2.0 spectral library [11] a set of 41 spectra has been randomly selected, categorized as in Fig. 2. Being the spectral range not constant across all the spectra, each spectrum has been resampled to the 244 bands of the future EnMAP mission's sensor [12], spanning the interval 0.42-2.45 μm, as described in [13]. The dataset looks a difficult one at first sight, as in some occasions the spectra exhibit similar behaviour or overlap (Fig. 3).

Class	Igneous				Sedimentary
Subclass	Mafic	Felsic			Shale
Name	Basalt	Tuff	Rhyolite	Dacite	Phosphorite
Origin	8 locations	Nevada	Spain	Spain	Idaho
Samples	20	4	10	2	5

Fig. 2. Categories of the rocks related to the analyzed spectra.

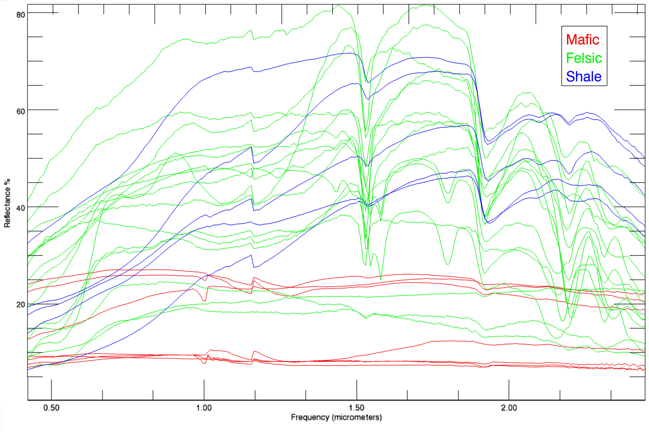

Fig. 3. The 41 spectra analyzed, belonging to three classes of rocks.

We computed a distance matrix related to the 41 spectra according to all the introduced distances. Then we performed on each distance matrix an unsupervised hierarchical clustering, by deriving a dendrogram (binary tree) which represents the matrix in 2 dimensions, as described in [8]. Results are reported in Fig. 4. Each leaf represents a spectrum, with the spectra which behave more similarly appearing as siblings. The evaluation is done by visually inspecting if spectra belonging to the same class are correctly clustered in some branch of the tree, i.e. by checking how much each class can be isolated by ”cutting” the tree at convenient points. The NCD is the only method yielding a good separation between the clusters, with the exception of the acceptable results obtained by the SA. It is surprising how the SID, which outperforms other distances in [4] and [5], results in a quite confused dendrogram. For the NCD the values have been first quantized in bytes to provide a meaningful data representation to the compressor used [6]. We also tested the other distances on the quantized data, with no improvement in the results.

5. CONCLUSIONS

In this paper we propose to use a general similarity measure based on data compression, the Normalized Compression Distance (NCD), to categorize spectra belonging to different kinds of rocks. Being the spectra extracted from different materials, the task looks demanding (Figs. 2 and 3). An unsupervised hierarchical clustering, carried out on the basis of the NCD distances between the spectra, results in a better performance with respect to traditional distances used in spectral matching. We argue that small differences between the spectra can be treated as noise, as we are analyzing different subclasses of the same minerals, so the analysis benefits from considering only the general behaviour of the data. This is well captured by the compressor in NCD, which implicitly focuses on the relevant information and is resistant to noise [14]. This suggests that the NCD could be able to capture information inside the spectra which does not result obvious.

The spectra have been analyzed in a spectral range characteristic of many hyperspectral sensors such as AVIRIS, HyMAP, the future EnMAP, and Hyperion. The proposed technique could be then successfully employed to characterize the contents of a scene acquired by such sensors.

6. REFERENCES

[1] F.A. Kruse et al., “The Spectral Image Processing System (SIPS) - Interactive Visualization and Analysis of Imaging Spectrometer Data,” *Remote Sensing of Environment*, vol. 44, pp. 145–163, 1993.

[2] O.A. de Carvalho and P.R. Meneses, “Spectral Correlation Mapper (SCM): An Improvement on the Spectral Angle Mapper (SAM),” in *NASA JPL AVIRIS Workshop*, 2000.

[3] H. Du, C.I. Chang, H. Ren, C.C. Chang, J.O. Jensen, and F.M. D’Amico, “New hyperspectral discrimination measure for spectral characterization,” *Optical Engineering*, vol. 43, no. 8, pp. 1777–1786, 2004.

[4] S.A. Robila and A. Gershman, “Spectral matching accuracy in processing hyperspectral data,” in *Signals, Circuits and Systems, 2005. ISSCS 2005. International Symposium on*, 2005, vol. 1, pp. 163–166.

[5] F.D. van der Meer, “The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery,” *International journal of applied earth observation and geoinformation*, vol. 8, no. 1, pp. 3–17, 2006.

[6] E. Keogh, S. Lonardi, and C.A. Ratanamahatana, “Towards parameter-free data mining,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, p. 215.

[7] D. Cerra, A. Mallet, L. Gueguen, and M. Datcu, “Algorithmic Information Theory-Based Analysis of Earth Observation Images: An Assessment,” *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 8–12, 2010.

[8] R. Cilibrasi and P. M. B. Vitányi, “Clustering by compression,” *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1523–1545, 2005.

[9] M. Cebrian, M. Alfonseca, and A. Ortega, “Common pitfalls using the normalized compression distance: What to watch out for in a compressor,” *Communications in Information and Systems*, vol. 5, no. 4, pp. 367–384, 2005.

[10] J. Ziv and A. Lempel, “Compression of individual sequences via variable-rate coding,” *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[11] A. M. Baldridge, S. J. Hook, C. I. Grove, and G. Rivera, “The aster spectral library version 2.0,” *Remote Sensing of Environment*, vol. 113, no. 4, pp. 711–715, Apr. 2009.

[12] R. Mueller et al., “The processing chain and cal/val operations of the future hyperspectral satellite mission enmap,” in *IEEE Aerospace Conference*, 2010.

[13] F.D. van der Meer, S.M. De Jong, and W. Bakker, *Imaging spectrometry: basic principles and prospective applications, Chapter 2*, Kluwer Academic Publishers, 2001.

[14] M. Cebrian, M. Alfonseca, and A. Ortega, “The normalized compression distance is resistant to noise,” *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1895–1900, May 2007.

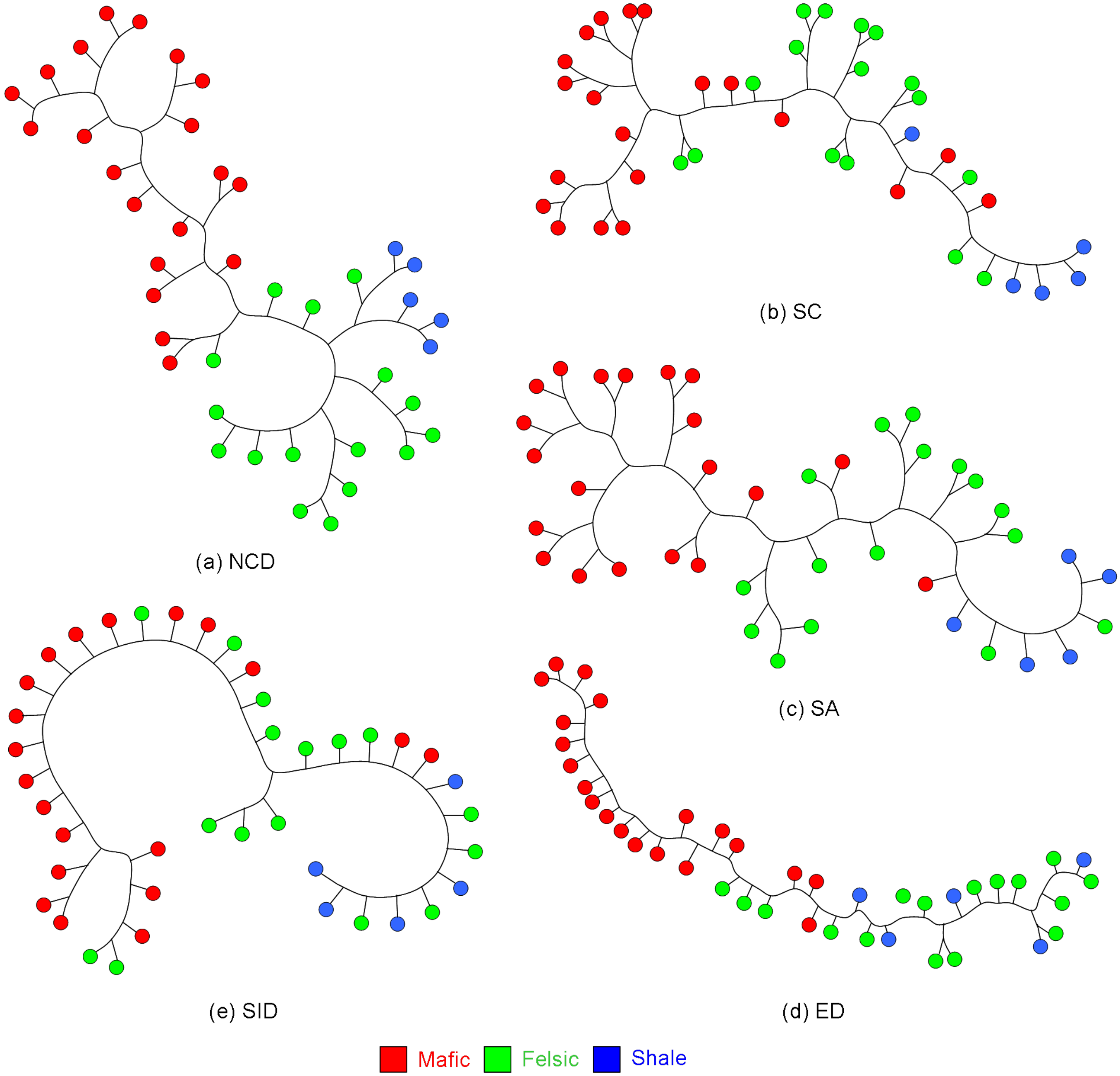

Fig. 4. Hierarchical clusterings for the dataset in Fig. 3, with each node in the tree representing an object, color-coded as in the reported legend. From top-left in clockwise order: results for NCD, SC, SA, ED, and SID distances.