Band Clustering for the Lossless Compression of AVIRIS Hyperspectral Images

Raffaele Pizzolante and Bruno Carpentieri
Dipartimento di Informatica, Università degli Studi di Salerno, I-84084, Fisciano (SA), Italy
rpizzolante@unisa.it, bc@dia.unisa.it

Abstract—Hyperspectral images can be efficiently compressed through a linear predictive model, as for example the one used in the SLSQ algorithm. In this paper we exploit this predictive model on the AVIRIS images by individuating, through an off-line approach, a common subset of bands, which are not spectrally related with any other bands. These bands are not useful as prediction reference for the SLSQ 3-D predictive model and we need to encode them via other prediction strategies which consider only spatial correlation. We have obtained this subset by clustering the AVIRIS bands via the clustering by compression approach. The main result of this paper is the list of the bands, not related with the others, for AVIRIS images. The clustering trees obtained for AVIRIS and the relationship among bands they depict is also an interesting starting point for future research.

Index Terms—lossless compression; hyperspectral images; band ordering; band clustering.

I. INTRODUCTION

Hyperspectral remote sensors produce daily a huge amount of hyperspectral images, that are obtained from the reflected light of the visible and the near-infrared spectrum. NASA Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) [10] sensors measure the spectrum from 400 up to 2500 nanometers. A hyperspectral image produced by this kind of sensor has generally 224 spectral bands.

Lossless data compression is generally needed to store or transmit hyperspectral images. The choice of a lossless compression algorithm is due to the high acquisition costs, and to the sophisticated types of analysis that this kind of material has a different spectral signature. Hence, a band can be predicted by another reference band, generally, the previous in the natural ordering.

Moreover it is important to consider that hyperspectral images generally need to be compressed “on board”, on airplanes or satellites and that the hardware capabilities available for such compression might be limited.

In earlier work [14], we proposed a pre-processing schema which performs a band ordering based on Pearson’s Correlation, and improves the compression performances of the state-of-art low-complexity lossless compression algorithm for hyperspectral images: SLSQ (Spectral-oriented Least SQuares, see [9, 13-17]).

In this paper we use the CompLearn clustering approach (see [3]), to off-line cluster the bands of AVIRIS hyperspectral images. The result of this clustering can be generalized to identify in the AVIRIS images a sub-set of the bands, that are not spectrally related with other bands, and that need to be compressed only by exploiting the spatial redundancy.

Generally these sub-sets of bands are similar in all the AVIRIS images. Hence, it is possible a one-time cluster which can be used for the encoding of all the hyperspectral images of the same kind.

The remainder of this paper is organized as follows: Section II shortly reviews SLSQ, CompLearn, and also the on-line correlation-based band ordering. Section III describes our band clustering approach and the experimental results obtained. Section IV presents our conclusions and future research directions.

II. BACKGROUND

A. Lossless and Lossy Compression of Hyperspectral images

In literature there are different approaches for lossless and lossy compression of hyperspectral images.

The lossless approaches are generally based on predictive model (as for example SLSQ, Linear Predictor [16], EMPORDA [18], Fast Lossless [4], etc.), which exploits the spatial and spectral redundancy. In a second stage, the prediction errors are entropy coded. Other approaches are based on differential pulse code modulation (DPCM) [12], improved version of vector quantization [7], dimensionality reduction through principal component analysis [19], etc.

In [1] error-resilient lossless compression methods are proposed and described.

The approaches for lossy compression of hyperspectral images are based on set partitioning in hierarchical trees (SPIHT-2D, SPIHT-3D, etc.), locally optimal vector quantization (LPVQ) [8]. Other methods are based on discrete wavelet transform (DWT) [5], 3-D discrete cosine transform (3-D DCT) [6], etc.

B. The Spectral-oriented Least SQuares (SLSQ) algorithm

Spectral-oriented Least SQuares algorithm is a low-complexity lossless hyperspectral image compression algorithm, which uses least squares to optimize a predictive model. It exploits two forms of redundancy: spatial and spectral correlation. Spatial correlation is based on the hypothesis that neighboring pixel are composed of the same material; spectral correlation derives from the fact that each material has a different spectral signature. Hence, a band can be predicted by another reference band, generally, the previous in the natural ordering.

The SLSQ block diagram is reported in Figure 1. SLSQ uses, for each sample, a 3-D prediction context, in order to perform the prediction through the computation of the optimal least squares coefficients.
The prediction context is obtained by SLSQ by using two distances, based on Euclidean distance: an intra-band distance, defined in (1) and an inter-band distance, defined in (2).

(a) \[ d_{2D}(x_{m,n,k}, x_{p,q,k}) = \sqrt{(m-p)^2 + (n-q)^2} \]

(b) \[ d_{ip}(x_{m,n,k}, x_{p,q,k}) = \frac{1}{4} \left( (m-p)^2 + (n-q)^2 \right) \quad \text{if} \quad i \neq j \]

The notation \( x_{m,n,k} \) refers to the pixel at spatial coordinates \((m, n)\) of the band \(k\). Figures 2 (a) and 2 (b) show respectively the resulting context for intra-band and inter-band context.

\[
\hat{x}(0,0) = \sum_{j=1}^{N} \alpha_j \cdot x(0, j)
\]

The energy of the prediction error is minimized through the coefficients \( \alpha_0 = [\alpha_1, ..., \alpha_N] \). The prediction error is defined as:

\[
P = \sum_{i=1}^{M} (x(i,0) - \hat{x}(i,0))^2
\]

Using the matrix notation, we can write \( P \) as:

\[
P = (C\alpha - X)^T - (C\alpha - X)
\]

where the matrix \( C \) and \( X \) are defined as following:

\[
C = \begin{bmatrix}
  x(1,1) & \cdots & x(1,N) \\
  \vdots & \ddots & \vdots \\
  x(M,1) & \cdots & x(M,N)
\end{bmatrix},
X = \begin{bmatrix}
  x(1,0) \\
  \vdots \\
  x(M,0)
\end{bmatrix}
\]

The solutions of the linear system \( (C^T C) \cdot \alpha = C^T X \), obtained by taking the derivate with respect to \( \alpha \) of \( P \) and setting it to zero, are the optimal predictor coefficients.

A predictor selector structure performs the prediction through the 2-D approach Median Predictor [2] (see the block “Predictor Selector” in Figure 1), if the band \( k \) belongs to the Intra-Band Set (IB), and through the 3-D predictive model, otherwise. The prediction error \( e = x - \hat{x} \) is entropy coded by using an arithmetic coder.

C. The CompLearn Toolkit

Clustering is the task of assigning a set of objects into clusters, i.e. into homogenous groups, so that the objects in the same cluster are more similar with each other than to those in other clusters with respect to a given distance metric.

In clustering by compression [3] the distance metric is based on the compressibility of the data and does not include any explicit semantic knowledge.

To intuitively understand why compression can be used as a distance metric, let us suppose that we have two digital files \( A \) and \( B \). If we compress \( A \) and \( B \) with a general-purpose, lossless, data compressor (for example gzip or bzip) we can indicate with \( L(A) \) and \( L(B) \) the compressed lengths (in bits) of \( A \) and \( B \).
If we need to compress together A and B then we can first compress A and then B and we have as resulting length of the two compressed files: \( L(A) + L(B) \). Another option we have is to append file B to file A and compress the resulting file AB.

The resulting length of the new compressed file shall be \( L(AB) \).

Experimentally it is possible to show that if and only if A and B are “similar”, then \( L(AB) << L(A) + L(B) \).

**D. Band Ordering**

**Test Set Description:** In the rest of this paper we have tested our approaches on a test set composed by five hyperspectral images freely provided by NASA Jet Propulsion Laboratory (JPL).

The images are: Cuprite (5 scenes), Lunar Lake (3 scenes), Moffett Field (4 scenes), Jasper Ridge (6 scenes) and Low Altitude (8 scenes). Each scene (except the last) of each hyperspectral image has 614 columns and 512 lines and 224 bands, each sample is represented as a signed integer on 16 bits.

Cuprite was acquired from the site of the Cuprite mining distinct, located in Nevada (USA). Lunar Lake was acquired from the site of Lunar Lake in Nye Country, also in Nevada (USA). Figure 3 reports a graphical representation of all the three scenes of Lunar Lake. Moffett Field was acquired from Moffett Federal Airfield, which is a civil-military airport located in California (USA) between Mountain View and Sunnyvale. Jasper Ridge covers a portion of the Jasper Ridge Biological Preserve, also located in California (USA).

Finally, the hyperspectral image denoted as Low Altitude was collected with a high spatial resolution, in which each pixel cover an area 4x4 meters, instead of the 20x20 meters area of the other hyperspectral images.

**Correlation-based Band Ordering:** When we try to exploit the inter-band redundancy with a three-dimensional predictive model, we have to assume that there is a strong spectral correlation between the current band and the previous, already coded, bands.

If we consider the standard, natural band ordering of a hyperspectral image, it is possible to prove that there are bands for which it would be possible to improve compression by using as prediction context a band that is different from the previous band in the natural order.

In [14] we considered Pearson’s Correlation [11], as measure of similarity between two bands. The value of the correlation is in range \([-1, 1]\), where a positive value indicates a direct relation and a negative value indicates an inverse relation.

Mathematically, the correlation is defined as:

\[
\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \cdot \sigma_y}
\]

where \( \sigma_{x,y} \) is the covariance between two random variables X, Y; \( \sigma_x \) and \( \sigma_y \) are the respectively the standard deviations for the two variables X and Y.

The approach we proposed in [14] can be sub-divided into three steps: the Graph creation, the Minimum Spanning Tree (MST) computation and the Depth-First Search (DFS) visit.

The first step consists of generating a graph \( G = (V, E) \), where each vertex denotes a band and each vertex \( i \) is connected, by a weighted edge, with each other vertex \( j \). The weight indicates the Pearson’s correlation:
\[ w(i, j) = -\text{Pearson Correlation}(\text{band}_i, \text{band}_j). \]

\[ w(\text{band}_a, \text{band}_b) = -0.2 \]
\[ w(\text{band}_b, \text{band}_c) = 0.1 \]
\[ w(\text{band}_a, \text{band}_c) = -0.3 \]

**Figure 4.** Incorrect band ordering given by DFS visit on MST M

The minus sign is required because the MST algorithm identifies the minimum possible as the best choice, in contrast to the correlation.

When the graph is created, MST is computed on the graph \( G \). With MST algorithm, each couple of bands \((i, j)\) is associated with the minimum distance (maximum Pearson’s correlation).

Finally, a Depth-First Search (DFS) visit, is performed on \( M \). This gives a band ordering, where \( M_i \) is a MST on graph \( G \).

In some situations DFS visit does not work correctly for our purposes. Figure 4 shows an example where DFS visit obtains unexpected results.

The band ordering given by DFS on MST \( M \) is \( a, b, c \). In this simple example SLSQ could predict the band \( b \) from the band \( a \) and the band \( c \) from the band \( b \), but the band \( b \) and the band \( c \) are very low related with respect to the band \( b \) and the band \( a \). Therefore, it is evident that this band ordering is incorrect.

For these cases, we need to modify our approach and introduce band ordering based on DFS and pairs, where a pair has this definition: \( \langle \text{reference band}, \text{band to predict} \rangle \).

In the previous example the band ordering given by a modified DFS visit is therefore \( \langle a, b \rangle, \langle a, c \rangle \), this means that SLSQ predicts the band \( b \) from the band \( a \) and the band \( c \) also from the band \( a \), and this is now a correct band ordering.

We tested our approach on the previous described test set. The achieved results are reported in Table 1.

The results show a compression improvement, and it is meaningful to consider this approach also with other distances.

However, there are some bands that are not related with any others and cannot be used as reference prediction bands. Section IV describes a possible approach for the off-line identification of these bands.

### III. Band Clustering For Lossless Compression

In every hyperspectral image there is a sub-set of bands, which have no relation with any others and cannot be intra-band predicted. We named these bands NR-bands. We suppose that the same kind of hyperspectral images \((i.e.\) acquired from the same kind of sensor), could present a similar sub-set of NR-bands \((NR-set)\).

Figure 5 shows for each image of the test set a portion of representation \((150 \times 150 \text{ pixel})\) of the bands number 80, 111, 161 and 222 (the band number 1 is the first).

### Table I: Experimental results for lossless compression of the test set.

<table>
<thead>
<tr>
<th>Hyperspectral Images / C.R.</th>
<th>SLSQ</th>
<th>Band Ordering+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lunar Lake</td>
<td>3.19</td>
<td>3.24</td>
</tr>
<tr>
<td>Moffett Field</td>
<td>3.17</td>
<td>3.20</td>
</tr>
<tr>
<td>Jasper Ridge</td>
<td>3.19</td>
<td>3.22</td>
</tr>
<tr>
<td>Cuprite</td>
<td>3.19</td>
<td>3.24</td>
</tr>
<tr>
<td>Low Altitude</td>
<td>3.01</td>
<td>3.04</td>
</tr>
<tr>
<td>Average</td>
<td>3.15</td>
<td>3.19</td>
</tr>
</tbody>
</table>

The bands number 111, 161 and 222 are NR-bands (as it is possible to see these bands are affected by noise in all the hyperspectral images of the test set).

The band number 80, instead, present strong relation with other bands, in all the hyperspectral images.

From the compression point of view, an NR-band is a band that is not related with (any possible) previous band and cannot be used as a reference prediction band for other bands.

Therefore, on the NR-bands a 3-D predictive model, such as SLSQ, fails because the spectral correlation is not strong enough. Hence, it is reasonable to compress the NR-bands with a different prediction approach. The Median Predicto, for example, achieves better results on these bands with respect to the SLSQ predictor.

Our goal is the extraction of a complete NR-set \((NR-set)\) from our small test data set of AVIRIS images, in order to use it for the compression of all the hyperspectral images of the same kind.

The problem of the identification of the NR-bands can be seen as a data clustering problem. Therefore, we used the CompLearn data clustering tool, in order to find the differences between the NR-bands and the other bands.

### A. Band Clustering

We divided the first scene of each hyperspectral image of our test data set in separated bands \((\text{for each band we produced a file})\), and then, for each first scene, these files are used as input to CompLearn, and it clusters these bands and produces an unrooted tree.

Figures 6, 7, 8 and 9 show respectively the resulting tree of the first scene of each hyperspectral images: Lunar Lake, Moffett Field, Jasper Ridge and Cuprite. The images, when printed, are not easily readable, therefore, to improve the legibility, we report a zoomed portion for each clustering trees.

Each leaf of the tree represents a band and the labels of the leaves identify the bands, as it is possible to see in the zoomed areas. An internal node has no label and represents a relation between two sub-trees.

CompLearn clusters the bands by putting bands that are similar in the same sub-tree. Sub-trees that are closer represent bands that are more related.

For example in Figure 6, the related bands 48 and 49 are in the same sub-tree, as it is possible to see in the zoomed area of the clustering tree.
B. Cuts and NR-Set

From the analysis of the clustering trees it is possible to see that all the NR-bands are grouped together in at most two sub-trees.

If we define a cut in the clustering tree as the elimination of a link between two nodes of the tree (i.e. the elimination of a sub-tree of the clustering tree), then we need at most two cuts in the AVIRIS clustering trees to take care of the NR-bands.

Figures 10, 11, 12 and 13 show the clustering tree respectively for Lunar Lake, Moffett Field, Jasper Ridge and Cuprite, and the black circles identify the cuts.

<table>
<thead>
<tr>
<th>Hyperspectral Image / Band</th>
<th>80-th</th>
<th>111-th</th>
<th>161-th</th>
<th>222-th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lunar Lake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moffett Field</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jasper Ridge</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cuprite</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Altitude</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table II. The number of cuts needed for the computation of the primitive NR-set for each hyperspectral image of the test set

Table II shows the number of the cuts for each first scene of the hyperspectral images.

In each NR-set we added also the first eight bands and all the bands where the prediction reference band is in the original, primitive, NR-set, i.e. if for the band \( i \) the compression algorithm uses as prediction reference band the band \( i - 1 \) and \( i - 1 \) belongs to the original, primitive, NR-set then we consider also the band \( i \) in NR-set. Moreover, the obtained trees propose also a possible relations, based on similarity, among the bands, and may be used also for other applications.

C. NRF-Set

We have obtained the final \( NR_{F} \)-Set by performing an intersection of all computed NR-sets on each test image, as follow:

\[
NR_{F} \sim Set_{\bigcap_{Img \in Test Set} NR - set_{Img}}
\]

where \( NR - set_{Img} \) indicates the NR-set on the first scene of the hyperspectral image \( Img \).

It is possible to experimental prove that the bands in this \( NR_{F} \)-Set are effectively not related with any other bands in the AVIRIS images, and that the shown in Table 3, which contains effectively the bands, with no relations with any others.

\( NR \)-bands belonging to \( NR_{F} \)-Set must be compressed through the Median Predictor, instead of the SLSQ predictor.

Table 4 reports the achieved results, in terms of compression ratio, by using the classical approach (second column) and the achieved results by predicting the NR-bands though the Median Predictor (third column).

As it is possible to see, \( NR_{F} \)-Set+SLSQ improves the SLSQ compression performances. Our one-time approach improves the compression performances but maintain unchanged the low-complexity nature of the SLSQ algorithm.
Figure 6. Overview of the resulting tree produced by the analysis of the first scene of Lunar Lake image.
Figure 7. Overview of the resulting tree produced by the analysis of the first scene of Moffett Field image
Figure 8. Overview of the resulting tree produced by the analysis of the first scene of Jasper Ridge image
Figure 9. Overview of the resulting tree produced by the analysis of the first scene of Cuprite image
\[ NR_{set_{Lunar\ Lake}} = \left\{ 1,2,3,4,5,6,7,8,29,107,108,109,110,111,112,113,114,153,154,155,156,157,158,159,160,161,162,163,164,165,166,167,222,223,224 \right\} \]

Figure 10. The two cuts on the resulting tree produced by the analysis of the first scene of Lunar Lake.
Figure 11. The two cuts on the resulting tree produced by the analysis of the first scene of Moffett Field.

$$\text{NR set}_{\text{Moffett Field}} = \{1,2,3,4,5,6,7,8,107,108,109,110,111,\}
\{112,113,114,115,153,154,155,156,\}
\{157,158,159,160,161,162,163,164,\}
\{165,166,167,169,221,222,223,224\}$$
Figure 12. The two cuts on the resulting tree produced by the analysis of the first scene of Jasper Ridge.

\[ NR - set_{\text{Jasper Ridge}} = \{1,2,3,4,5,6,7,8,108,109,110,111,112,113,114,154,155,156,157,158,159,160,161,162,163,164,165,166,167,168,222,223,224\} \]
Figure 13. The two cuts on the resulting tree produced by the analysis of the first scene of Cuprite.

\[ NR - set_{\text{Cuprite}} = \{1, 2, 3, 4, 5, 6, 7, 8, 107, 108, 109, 110, 111, 112, 113, 134, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 222, 223, 224\} \]
In this paper we proposed an off-line band clustering strategy for AVIRIS images that identifies a recurrent sub-set of bands that are not spectrally related with any other bands (NR$_B$-Set).

The bands belonging to this NR$_B$-Set must be compressed by exploiting only the spatial correlation, as for example with the Median Predictor.

Future research will include the improvement of this approach, in order to make it more robust, and the design of other off-line approaches, which have as target the improvement of lossless compression of hyperspectral images without changing the complexity of the compression algorithm, even with different kinds of hyperspectral images.

IV. Conclusions and Future Work

In this paper we proposed an off-line band clustering strategy for AVIRIS images that identifies a recurrent sub-set of bands that are not spectrally related with any other bands (NR$_B$-Set).

The bands belonging to this NR$_B$-Set must be compressed by exploiting only the spatial correlation, as for example with the Median Predictor.

Future research will include the improvement of this approach, in order to make it more robust, and the design of other off-line approaches, which have as target the improvement of lossless compression of hyperspectral images without changing the complexity of the compression algorithm, even with different kinds of hyperspectral images.

ACKNOWLEDGEMENT

We would like to thank our student Filippo Alfano for testing a preliminary version of this work.

REFERENCES


