CONTEXT-BASED PREDICTIVE LOSSLESS CODING FOR HYPERSPECTRAL IMAGES

A. De Giusti, S. Andriani, G.A. Mian

Department of Information Engineering, University of Padova
Via Gradenigo, 6/B - 35131 Padova - Italy
e-mail: {degiu,mutley,mian}@dei.unipd.it
web: www.dei.unipd.it

ABSTRACT

A cluster-based lossless compression algorithm for hyperspectral images is presented. Clustering is carried out on the original data according to the vectors spectra, and it is used to set up multiple contexts for predictive lossless coding. Low-order prediction is performed using adaptive Linear Least Squares (LLS) estimation which exploits the additional information provided by clustering. Prediction errors are then entropy-coded using an adaptive arithmetic coder also driven by data clusters.

The proposed scheme is used to losslessly code a set of AVIRIS hyperspectral images. Comparisons with the JPEG-LS, JPEG-2000 and the clustered DPCM coding algorithms are given.

1. INTRODUCTION

In recent years, hyperspectral images have grown to hundreds of spectral bands leading a considerable increase of the memory required for data storage. In order to address this problem, many near-lossless compression schemes have been proposed in the literature [1, 2, 3]. However, different practical applications need all the data to be available without any coding distortion, and thus the interest in efficient lossless coding algorithms is steadily increasing. From a statistical point of view, hyperspectral data present both strong spatial and spectral correlations which standard coding algorithms do not exploit efficiently.

In this paper, we present a lossless coding approach based both on data clustering and low-order spectral prediction. Clusters are obtained from a statistical partition of the data vectors according to a mixture of Gaussian densities whose parameters are computed by solving a Maximum Likelihood (ML) problem. Clusters are the contexts used both in prediction and entropy coding. In fact, inside each cluster, Linear Least Squares estimates (LLSE) are used to adaptively compute the optimal prediction coefficients for each pixel. Finally, prediction errors are entropy-coded using a first-order arithmetic coder also based on the contexts provided by clusters.

The paper is organized as follows. Section 2 gives a brief review of the coding algorithms used for comparison. Section 3 introduces our clustering approach. Section 4 presents the LLSE cluster-based prediction. Finally, Sections 5 and 6 report the results obtained by coding a set of AVIRIS hyperspectral images, and conclusions.

2. PREVIOUS CODING SCHEMES

Three algorithms were taken into account for comparisons, namely JPEG-LS, JPEG-2000 and the clustered DPCM.

Both JPEG-LS and JPEG-2000 are popular coding standards for still images, and were adapted to code also hyperspectral data. For their popularity they have been taken into account as reference algorithms. The clustered DPCM, proposed by Mielikainen *et al.* [4], is an algorithm specifically designed for hyperspectral data lossless coding which outperforms other algorithms given in the literature [5, 6, 7, 8].

The JPEG-LS [9] is the JPEG lossless coding standard. It is based on the prediction algorithm provided by the Median Edge Detector (MED) followed by a context-based Golomb-Rice coder. Compression efficiency is possibly improved using Run-Length coding whenever four adjacent pixels have the same value.

The JPEG-2000 [10] is the well-known compression algorithm designed for lossy to lossless data coding of still images. A Discrete Wavelet Transform (DWT) is applied to the data, and the resulting coefficients are coded using a binary bit-plane arithmetic coder (MQ-coder), which leads to a highly scalable coded bitstream.

Finally, the last algorithm taken into account is the clustered DPCM which is a very efficient predictive coder exploiting both clustering and spectral correlation. In this scheme, data are clustered according to vector spectra using the Linde-Buzo-Gray (LBG)[11] vector quantizer. The set of labels generated by the LBG algorithm defines the clusters on which spectral prediction is performed. Labels are entropy-coded and added to the output bitstream as side information. Linear prediction is computed by minimizing the expected value of the squared error inside each cluster. The optimal coefficients of each predictor are uniformly quantized to a 16-bit representation, and outputted to the bitstream. Finally, residual values resulting from the prediction are entropy-coded by an adaptive range coder [12] which uses clusters as contexts.

3. PROPOSED CLUSTERING APPROACH

The proposed algorithm is based on data segmentation which is used to set up context-based prediction and entropy-coding. The statistical segmentation is provided by the SEM algorithm [13] which is a robust stochastic implementation of the Expectation-Maximization (EM) algorithm [14].

Let $X = \{\mathbf{x}_j | j = 1, \ldots, M\}$ be the set of observed hyperspectral vectors. Each K-dimensional vector $\mathbf{x}_j \in X$ is supposed to be drawn by the unknown probability density function (pdf) $f(\mathbf{x})$. Let C be the number of clusters being used in the segmentation. The statistical estimate of $f(\mathbf{x})$ is performed by introducing a set of C multivariate Gaussian pdfs $f(\mathbf{x}|\theta_c)$ of mean μ_c and covariance Σ_c. Each Gaussian has parameters $\theta_c = \{\mu_c, \Sigma_c\}$, and the corresponding mixture parametric model

$$f(\mathbf{x}|\Theta) = \sum_{c=1}^{C} \alpha_c f(\mathbf{x}|\theta_c) \quad (1)$$

is completely specified by the set of parameters $\Theta = \{\alpha_c, \theta_c : c = 1, \ldots, C\}$, where the weighting coefficients α_c are all non-negative and sum up to one.

Let Ω be the space of parameters. Under the assumption of independent and identically distributed (i.i.d.) vectors, the model's optimal parameters Θ^* are estimated by maximizing the log-likelihood function

$$L(\Theta|X) = \ln \prod_{j=1}^{M} f(\mathbf{x}_j|\Theta) = \sum_{j=1}^{M} \ln \sum_{c=1}^{C} \alpha_c f(\mathbf{x}_j|\theta_c) \quad (2)$$

with respect to the observed data X over all the admissible choices of $\Theta \in \Omega$

$$\Theta^* = \arg\max_{\Theta \in \Omega} L(\Theta|X) \; . \quad (3)$$

Given the Maximum-Likelihood (ML) problem (3) some simplifications are in order to make the estimation mathematically tractable. In the SEM algorithm this is achieved considering the observed dataset X as *incomplete*, and assuming the existence of an unobserved dataset W which specifies from which Gaussian density each vector has been drawn. The compound set $Z = \{X, W\}$ is then referred as the *complete* dataset.

Starting from an initial choice of parameters $\Theta^{(0)}$, the ML problem (3) is solved iteratively applying, firstly, the Expectation step which computes the expected value of the complete-data likelihood $L(\Theta|Z)$ given the current parameters $\Theta^{(i)}$ and the observed data X

$$Q(\Theta, \Theta^{(i)}) = E\left[L(\Theta|X, W)|X, \Theta^{(i)}\right] \; , \quad (4)$$

and, secondly, the Maximization step which gives the new parameters $\Theta^{(i+1)}$ by maximizing the functional in (4)

$$\Theta^{(i+1)} = \arg\max_{\Theta \in \Omega} Q(\Theta, \Theta^{(i)}) \; . \quad (5)$$

Robustness to pathological cases is finally achieved by implementing a stochastic assignment of vectors to clusters using random drawings at each iteration. The sequence of parameters $\Theta^{(i)}$ gives non-decreasing complete-data likelihoods $L(\Theta^{(i)}|Z)$ and, consequently, also non-decreasing incomplete-data likelihoods $L(\Theta^{(i)}|X)$ (refer to [13] for further details).

In practice, clustering is not computed directly in the original full-dimensional data space. In order to reduce the computational cost, discrete cosine (DC) transform and basis restriction are used to pack most of the data variance into a small number of components on which the segmentation algorithm is effectively carried out.

Clusters provided by segmentation are represented by a set of labels $\mathscr{C} = \{\ell_j | 1 \leq \ell_j \leq C, j = 1, \ldots, M\}$ specifying which cluster the corresponding vector belongs to. Vectors having the same label are in the same cluster, and each cluster defines a unique context for both *vector prediction* and *entropy-coding* of prediction errors.

The labels in $\mathscr{C}$ are coded using a (JPEG-LS)-like coder and added to the output bitstream as side information.

4. CLUSTER-BASED PREDICTION

In linear predictive coding, the correlation between the current sample and its causal neighborhood is used to code only the unpredictable part of the data.

In hyperspectral imagery, both spatial and spectral correlations can be used for this purpose. Moreover, data clustering may improve prediction by grouping vectors according to their spectral behavior, and providing an effective way to set up different contexts for coding different regions of the image.

Focusing only on the spectral prediction, we may concern that some bands are better predictable than others. Thus, the use of low-order linear predictors may be enhanced by using a band reordering algorithm [15]. Mielikainen *et al.* overcome this issue using high-order predictors[1]. However, this solution yields a lot of side information affecting the achieveable compression ratio, especially when a high number of clusters is used. On the other hand, the prediction error is orthogonal to the vector space generated from prediction vectors. Consequently, high-order predictors usually perform better than low-order ones.

The objective of the proposed algorithm is to define a cluster-based predictive scheme achieving good compression ratios using, however, *low-order* and *strongly-adaptive* predictors which exploit both spectral and spatial correlation.

For each component, spectral prediction coefficients are computed by setting up an LLS estimate whose defining equations are varied band by band according to a distance measure between the current vector and vectors belonging to its causal neighborhood within the same cluster.

Since prediction can only be performed from data already coded (i.e. already available to the decoder), the first band of the hyperspectral data is encoded using only the spatial prediction given by the MED algorithm.

Consider the observed dataset $X = \{\mathbf{x}_j | j = 1, \ldots, M\}$ and the current vector $\mathbf{x}_j = [x_{j,1}, \ldots, x_{j,K}]^T$ whose k-th component $x_{j,k}$ is being predicted. Without loss of generality, we may assume that vectors indexes have been reordered such that all the k-th components of $\mathbf{x}_i$, with $i < j$, have already been predicted.

Let $\ell_j \in \mathscr{C}$ be the cluster label associated to $\mathbf{x}_j$, and consider the N-dimensional vector ($N < K$) formed by the components of $\mathbf{x}_j$ that are used to predict $x_{j,k}$

$$\mathbf{x}_j^{(k)} = [x_{j,(k-1)}, \ldots, x_{j,(k-N)}]^T \; . \quad (6)$$

The linear prediction of $x_{j,k}$ is given by the inner product

$$\hat{x}_{j,k} = < \mathbf{b}_j^{(k)}, \mathbf{x}_j^{(k)} > = \sum_{i=1}^{N} b_{j,i}^{(k)} x_{j,(k-i)} \quad (7)$$

where $\mathbf{b}_j^{(k)} = [b_{j,1}^{(k)}, \ldots, b_{j,N}^{(k)}]^T$ is the N-dimensional vector of prediction coefficients. The prediction error is

$$e_{j,k} = x_{j,k} - \hat{x}_{j,k} \; . \quad (8)$$

In order to estimate the optimal coefficients $\mathbf{b}_j^{(k)}$, let us consider the set I_j of the indexes whose corresponding vectors

[1] In their work, a 16-th order predictor was the optimal trade-off between prediction efficiency and side information overhead.

have already been predicted and belong to the same cluster of $\mathbf{x}_j$

$$I_j = \left\{ i < j \,|\, i,j \in \{1,\ldots,M\} \wedge \ell_i = \ell_j \right\} \ . \tag{9}$$

Given a subset $J_j^{(k)} = \{m_1,\ldots,m_q\} \subseteq I_j$ of vectors in the causal neighborhood of $\mathbf{x}_j$, the LLS estimate of $\mathbf{b}_j^{(k)}$ is the solution of the linear system $\mathbf{X}_j^{(k)}\mathbf{b}_j^{(k)} = \mathbf{c}_j^{(k)}$ given by

$$\begin{bmatrix} \mathbf{x}_{m_1}^{(k)^T} \\ \vdots \\ \mathbf{x}_{m_q}^{(k)^T} \end{bmatrix} \mathbf{b}_j^{(k)} = \begin{bmatrix} x_{m_1,k} \\ \vdots \\ x_{m_q,k} \end{bmatrix} . \tag{10}$$

To explain the system in (10), consider that the indexes $m_1,\ldots,m_q$ identify q causal vectors of $\mathbf{x}_j$ within the same cluster. Thus, the $(q \times N)$-matrix $\mathbf{X}_j^{(k)}$ is formed by the components of these vectors in the preceding N bands. Finally, $\mathbf{c}_j^{(k)}$ is a q-dimensional vector consisting of the known values in the k-th band.

Choosing $q > N$ yields an overdetermined linear system. If the columns of $\mathbf{X}_j^{(k)}$ are linearly independent, then the matrix $\mathbf{X}_j^{(k)^T}\mathbf{X}_j^{(k)}$ is invertible, and the LLS solution is

$$\mathbf{b}_j^{(k)} = \left(\mathbf{X}_j^{(k)^T}\mathbf{X}_j^{(k)}\right)^{-1}\mathbf{X}_j^{(k)^T}\mathbf{c}_j^{(k)} \ . \tag{11}$$

This is known as the *pseudo-inverse* solution of the system [16]. If the $\mathbf{X}_j^{(k)^T}\mathbf{X}_j^{(k)}$ matrix has not full rank, then we may proceed in the same way by adding a fixed small value to its main diagonal. It should be also noted that the prediction coefficients $\mathbf{b}_j^{(k)}$ are estimated from vectors already coded, and thus available to the decoder.

Vectors indexes in $J_j^{(k)}$ used in the LLS equations correspond to a suitable set of vectors belonging to the causal neighborhood of the current vector $\mathbf{x}_j$. Initially, when no information about the spatial correlation is available, this set is simply defined from the clusters labels by taking the nearest q vectors to $\mathbf{x}_j$ according to the Manhattan distance. Once that $x_{j,k}$ has been predicted and coded, the new set $J_j^{(k+1)}$ for the prediction of $x_{j,k+1}$ is determined by taking the q vectors in I_j at minimum distance from $\mathbf{x}_j^{(k+1)}$. This distance can be the usual L_2 norm. However, the L_1 norm can be successfully used to speed up the search.

Since I_j represents the entire causal history of the vector $\mathbf{x}_j$ (inside its cluster), the search for minimum-distance vectors is computationally demanding. From a practical point of view, this search can be limited to a small subset, slightly grater than $J_j^{(k)}$, containing only vectors spatially closer to $\mathbf{x}_j$.

The underlying idea in reordering the prediction indexes is to adaptively adjust the set of vectors that will be used to build the LLSE system. This is accomplished by estimating the spatial correlation between vectors within each cluster, and leads to good performance also when low-order predictors are used.

For the sake of simplicity, the above derivations do not take into account band reordering. As mentioned before, band reordering is a bijective function which maps the bands order as it is collected by hyperspectral sensors to a new order with the purpose of reducing the average prediction error distortion [17]. Clusters can be also used in this kind of operation. Independent ordering functions (one for each cluster) may be defined according to the prediction distortion which occurs within each cluster.

The prediction error $e_{j,k}$ defined in (8) is the unpredictable information that has to be stored for the lossless reconstruction of the data. However, lossless coding requires its integer approximation

$$\hat{e}_{j,k} = x_{j,k} - \lfloor \hat{x}_{j,k} \rfloor \ . \tag{12}$$

Once that the hypercube has been predicted, the last operation is to entropy-code the prediction errors which usually have lower entropy than the original data. To implement entropy-coding, a cluster-based first-order arithmetic coder [18] is implemented. However, when the original bands reach better compression than predicted one, the prediction errors are simply discarded and the original data is encoded.

5. RESULTS

The proposed coding algorithm was tested on a set of hyperspectral images collected by the Airborne Visible/InfraRed Imaging Spectrometer[2] (AVIRIS). The electromagnetic spectrum radiance emitted by the earth's surface is measured in 224 narrow-length frequency bands (each 10 nm wide), starting from $0.42\,\mu m$ to $2.45\,\mu m$. The AVIRIS sensors collect data continuously from an altitude of about 20 km above the sea level. Samples are successively divided into hypercubes of 614×512 spectral vectors, each of which represents a 20×20 m^2 square region to the ground. Finally, each radiance component is numerically represented and stored as a 16-bit number. Test images used in the simulations were Jasper Ridge, Moffett Field, Lunar Lake and Cuprite.

Simulations were conducted using different number of clusters in the parametric model (1). In clustered DPCM, prediction coefficients are stored in the bitstream. Thus, once that the maximum compression ratio has been reached for a certain number of clusters, varying that number results in a rapid growth of the side information size. In the proposed scheme prediction coefficients are *not* stored in the bitstream, and consequently the influence of the number of clusters is not so critical as in the clustered DPCM, and similar compression ratios are achieved for different number of clusters.

The results reported in Table 1 are a comparison between JPEG-LS, JPEG-2000, clustered DPCM and the proposed algorithm using 10 clusters and a 4-th order predictor. As said before, JPEG-LS and JPEG-2000 do not exploit the spectral correlation between bands. Consequently, they achieve an average compression ratio of only 2.03:1 and 1.87:1, respectively. Note that JPEG-LS outperforms JPEG-2000 because it is specifically designed for lossless compression, while JPEG-2000 is not optimized for this kind of coding.

For what concerns the clustered DPCM algorithm, the proposed coding scheme works slightly better reaching an

[2] Free data are available at the AVIRIS website http://aviris.jpl.nasa.gov

Image Sets	JPEG-LS	JPEG-2000	Clustered DPCM	Proposed
Moffet Field	1.99	1.82	3.46	3.45
Jasper Field	1.91	1.78	3.46	3.46
Lunar Lake	2.14	1.96	3.37	3.38
Cuprite	2.09	1.91	3.42	3.43
Average	2.033	1.868	3.428	3.430

Table 1: compression ratios for the AVIRIS test images.

average compression ratio of 3.430:1 instead of 3.428:1 corresponding to an average rate of about 4.66 bits per component (bpc). In practice, the two algorihms achieve the same compression capabilities.

Clustered DPCM and the proposed scheme are similar in the way they use clustering to improve performance. They reach good compression ratios (about 3.4:1) and have about the same coding efficiency. The interesting fact is that the proposed 4-th order LLS adaptive predictor yields the same efficiency used in the 16-th order predictor of the clustered DPCM, demonstrating the effectiveness of the proposed prediction.

For what concerns band reordering, the algorithm proposed in [15] was implemented and tested. In practice, it gave us only small improvements of the coding performance. This is due to the strong adaptivity of the LLSE predictor which successfully exploits both spatial and spectral correlation. Usually, simpler predictors only works on trivial relationships between adjacent bands and components, therefore the straightforward solutions in those cases are either to use band reordering or to increase the predictor order.

6. CONCLUSIONS

A cluster-based predictive coding algorithm for hyperspectral data lossless compression has been presented. Highly-adaptive LLSE predictions exploiting spatial and spectral correlation permitted to achieve high compression ratios (similar to those of clustered DPCM) using very-low order predictors.

Acknowledgements: this work was supported by the Italian Ministry of Education, University and Research (MIUR), under the FIRB PRIMO project “Reconfigurable platforms for wideband wireless communications”.

REFERENCES

[1] Y. Tseng, H. Shih, and P. Hsu, “Hyperspectral Image Compression Using Three-Dimensional Wavelet Transformation,” in *Proc. Asian Conference on Remote Sensing (ACRS 2000)*, (Taipei, Taiwan), pp. 230–233, Dec. 2000.

[2] X. Tang, S. Cho, and W. Pearlman, “3D Set Partitioning Coding Methods in Hyperspectral Image Compression,” in *Proc. IEEE International Conference on Image Processing (ICIP 2003)*, (Barcelona, Spain), pp. 160–163, Sept. 2003.

[3] A. De Giusti and G. A. Mian, “Classification-Based Vector Source Coding,” in *Proc. International Symposium on Information Theory and its Applications (ISITA 2004)*, (Parma, Italy), pp. 380–385, Oct. 2004.

[4] J. Mielikainen and P. Toivanen, “Clustered DPCM for the Lossless Compression of Hyperspectral Images,” *IEEE Trans. Geosci. Remote Sensing*, vol. 41, pp. 2943–2946, Dec. 2003.

[5] J. Mielikainen, P. Toivanen, and A. Kaarna, “Linear Prediction in Lossless Compression of Hyperspctral Images,” *Opt. Eng.*, vol. 42, no. 4, pp. 1013–1017, 2003.

[6] G. Motta, F. Rizzo, and H. Storer, “Compression of Hyperspectral Imagery,” in *Proc. Data Compression Conf.*, pp. 333–342, 2003.

[7] M. Pickering and M. Ryan, “Efficient Spatial-Spectral Compression of Hyperspectral Data,” *IEEE Trans. Geosci. Remote Sensing*, vol. 39, pp. 1536–1539, 2001.

[8] S. Srinivasan and L. Kanal, “Eigenwavelet: Hyperspectral Image Compression Algorithm,” in *Proc. Data Compression Conf.*, 1999.

[9] M. Weinberger, G. Seroussi, and G. Sapiro, “The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS,” *IEEE Trans. Image Processing*, vol. 9, pp. 1309–1324, Aug. 2000.

[10] D. Taubman and M. Marcellin, *JPEG2000: Image Compression Foundamentals, Standards, and Practice*. Kluwer Academic Publishers, Boston, 2002.

[11] Y. Linde, A. Buzo, and R. Gray, “An Algorithm for Vector Quantization Design,” *IEEE Trans. Commun.*, vol. 28, pp. 84–95, Jan. 1980.

[12] G. Martin, “Range encoding: an algorithm for removing redundancy from digitized messages,” in *Proc. Video and Data Recording Conference 1979*, pp. 289–292, Sept. 1979.

[13] G. Celeux, D. Chauveau, and J.Diebolt, “On Stochastic Versions of the EM Algorithm,” *Institut National de Recherche en Informatique et en Automatique*, vol. 1, July 1995.

[14] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, July 1977.

[15] S. Tate, “Band ordering in lossless compression of multispectral images,” *IEEE Trans. Comput.*, vol. 46, pp. 477–479, Apr. 1997.

[16] M. Hayes, *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.

[17] A. Miguel, A. Askew, A. Chang, S. Hauck, R. E. Ladner, and E. Riskin, “Reduced complexity wavelet-based predictive coding of hyperspectral images for fpga implementation,” in *Proc. NASA Earth Science Technology Conference*, pp. 643–668, Sept. 2003.

[18] M. Nelson, *The Data Compression Book*. M&T Publishing, Inc., 1992.