N89-22344

DATA COMPRESSION EXPERIMENTS WITH LANDSAT THEMATIC MAPPER AND NIMBUS-7 COASTAL ZONE COLOR SCANNER DATA

James C. Tilton and H. K. Ramapriyan
Space Data and Computing Division
Goddard Space Flight Center

ABSTRACT

The variety of remote sensing instruments expected to be deployed in the last decade of this century and early 21st century, their resolutions and the anticipated data collection rates imply requirements for on-board reduction of data volumes in order to maximize the scientific return from space in the face of limited transmission bandwidth. Such data reductions can be achieved either through lossless or lossy data compression or through on-board "analysis and information extraction" and transmission of results. Several recent and potentially anticipated advances in computer science and hardware technology make it feasible to consider the development of on-board computer systems with sufficient capability to accomplish the above tasks. It is obvious that compression techniques which are shown to be feasible for on-board implementation can also be implemented for on-ground data compression thus helping reduce the archival storage costs, increase the on-line availability of data, and reduce times needed for browsing data for a given region or time interval.

Studies evaluating image entropies treating images pixel by pixel or considering differences between adjacent pixels indicate that lossless compression ratios of 1.5 to 3 can be achieved (of course, depending on the data) by optimal encoding of pixel values or differences. It is to be noted, however, that the entropies so defined do not represent the theoretical performance limit on reversible (i.e., lossless) data compression. Lossy compression

PRECEDING PAGE BLANK NOT FILMED

techniques such as predictive encoding, discrete transforms, cluster coding and vector quantization can achieve greater compression ratios which are a function of the acceptable level of loss. A common objection to these data compression techniques is that with significant compression (factors greater than 10) the data cannot be exactly recovered in their raw form.

For any lossy technique to be acceptable for a given application, it is necessary to demonstrate that the most of the relevant information remains in the compressed data. Therefore, to prove the utility of a compression technique for a scientific application, it is necessary to perform case studies with remotely sensed data in selected disciplines, use well accepted analysis techniques, and demonstrate that compressed data result in very nearly the same analysis results as the original data. With sufficient interaction between the scientific community and developers of data compression techniques it should be possible to define such case studies, and in fact, arrive at techniques which will not only reduce the data volume using criteria tailored to the analysis techniques, but also facilitate data analysis by direct use of compressed data.

In this paper, we present a case study where an image segmentation based compression technique is applied to Landsat Thematic Mapper (TM) and Nimbus-7 Coastal Zone Color Scanner (CZCS) data. The compression technique, called Spatially Constrained Clustering (SCC), can be regarded as an adaptive vector quantization approach. The SCC can be applied to either single or multiple spectral bands of image data. The segmented image resulting from SCC is encoded in small rectangular blocks, with the "codebook" varying from block to block. Lossless compression potential (LCP) of sample TM and CZCS images are evaluated. For the TM test image, the LCP is 2.79. For the CZCS test image the LCP is 1.89, even though when only a cloud-free section of the image is considered the LCP increases to 3.48. Examples of compressed images are shown at several compression ratios ranging from 4 to 15. In the case of TM data, the compressed data are classified

using the Bayes' classifier. The results show an improvement in the similarity between the classification results and ground truth when compressed data (with compression ratios of up to 13.8) are used, thus showing that compression is, in fact, a useful first step in the analysis. Future work in this case study will include the use of SCC-compressed CZCS data to obtain chlorophyll concentrations using the algorithm currently in use at GSFC for the production of global chlorophyll maps.

INTRODUCTION

The resolutions and anticipated data collection rates of the variety of remote sensing instruments expected to be deployed in the last decade of this century and the early 21st century imply requirements for on-board reduction of data volumes in order to maximize the scientific return from space in the face of limited down-link transmission bandwidth. Such data reductions can be achieved either through lossless or lossy data compression or through on-board analysis and information extraction and transmission of results. Several recent and anticipated advances in computer science and hardware technology make it feasible to consider the development of on-board computer systems with sufficient capability to accomplish the above tasks. It is obvious that compression techniques which are shown to be feasible for on-board implementation can also be implemented for on-ground data compression for the purpose of reducing archival storage costs, increasing the online availability of data, and reducing the time needed for browsing data for a given region or time interval.

Studies evaluating image entropies treating images pixel by pixel or considering differences between adjacent pixels indicate that lossless compression ratios of 1.5 to 3.0 can be achieved (of course, depending on the data) by optimal encoding of pixel values or differences (Chen[1], Ramapriyan[7], Wharton[11]). The entropies so defined do not necessarily represent the theoretical performance limit on reversible

(i.e., lossless) data compression. However, the actual theoretical performance limit is likely to be less than twice the compression ratios indicated. Lossy compression techniques such as predictive encoding, discrete transforms, cluster coding and vector quantization can achieve greater compression ratios subject to an acceptable level of loss. A common objection to these data compression techniques is that with significant compression (factors greater than 10) the data cannot be exactly recovered in their raw form.

For any lossy technique to be acceptable for a given application, it is necessary to demonstrate that most of the relevant information is retained in the compressed data. To prove the utility of a compression technique for a scientific application we must perform case studies with remotely sensed data in selected disciplines, use well accepted analysis techniques, and demonstrate that the use of compressed data produces very nearly the same analysis results as with the original data. There are several case studies where effects of data compression on multispectral classification have been studied (Kauth[5], Hilbert[2], Ramapriyan[7]). A common characteristic among these studies is that the compression technique is really a precursor to analysis and information extraction. A variety of such case studies are needed for several scientific disciplines and applications with sufficient interaction between the scientific community and developers of data compression techniques. Through such case studies it should be possible to arrive at techniques which will not only reduce the data volume using criteria tailored to the analysis techniques, but also facilitate data analysis by direct use of compressed data.

In this paper we present a case study where an image segmentation based compression technique is applied to Landsat Thematic Mapper (TM) and Nimbus-7 Coastal Zone Color Scanner (CZCS) data. When accompanied by an encoding of the resulting segmentation, the Spatially Constrained Clustering (SCC) segmentation approach can be regarded as an adaptive vector quantization approach to data compression. The SCC

data compression approach can be applied to either single or multiple spectral bands of image data. The segmented image resulting from SCC is encoded in small rectangular blocks, with the "codebook" varying from block to block.

ALGORITHMS AND ERROR MEASURES

In this section, we define the Lossless Compression Potential (LCP) of an image and the error measures used to evaluate the compressed data. We also describe the SCC segmentation algorithm and the block cluster coding method used to encode the segmented image to obtain a compressed image.

Lossless Compression Potential

Ideally, the LCP of a given image would be defined as the maximum factor by which the image can be reversibly (that is, losslessly) compressed. If every pixel of an image is totally uncorrelated with other pixels in the image, this ideal LCP could be easily calculated from the zeroth order entropy of the image. However, image pixels are generally highly correlated, causing the zeroth order entropy to underestimate the LCP of the image. To compensate partially for the correlation between image pixels, we define our LCP based on the zeroth order entropy of a difference image in which each pixel is represented as a function of three neighboring pixels and itself (Rosenfeld and Kak[8]):

$$d'(x,y) = d(x,y) - d(x-1,y) - d(x,y-1) + d(x-1,y-1) \qquad (1)$$

where d(x,y) represents the original image value at pixel (x,y), and d'(x,y) represents the difference image value at pixel (x,y). This is a special case of two-dimensional Differential Pulse Code Modulation (DPCM) (Jain[3]). (Note: the first row and column of d'(x,y) are generated by assuming that the "0^{th}" row and column of d(x,y) are equal to some "average" value. This average value must be stored

separately. For convenience, we can take this to be the median value of the first row of the image.) In the case of multiband images, each band is transformed separately according to equation (1) above.

The entropy (Shannon[9]) of an image is obviously dependent on the definition of the "source alphabet" and statistics of the reception of symbols from it. In the one extreme, the image could be considered to be obtained from a binary source (i.e., as a serial bit stream). In the other extreme, the source could be an "image generator" which produces images of a given size and the given image would then be regarded as an instance from the ensemble of all possible images (i.e., a single symbol from a very large alphabet!). In practical applications, the source alphabet is defined to consist of single or multiband n-bit pixel values (e.g., for 7-band Landsat TM data the source alphabet could either be all possible 8-bit pixel values or all possible 56-bit pixel values).

The zeroth order entropy, H0, of an image, d, is given by

$$H_0\ (d) = -\sum_{i=1}^{\beta} P_i\ (d) \times \log_2\ (P_i\ (d)) \qquad (2)$$

where Pi(d) is the probability of a pixel in image, d, having value i. $\beta = 2$, where b is the number of bits per pixel in the image. We calculate LCP of image d(x,y) by first finding the difference image d'(x,y) through the process defined by equation (1). We then estimate the pixel value probabilities, Pi(d'), from the histogram of d'(x,y) and calculate the zeroth order entropy, $H_0(d')$, through equation (2). To complete the definition of the LCP we note that in order to decode the compressed image reconstructing the original image we need to know the code used to encode the image. We assume that a variable length Huffman code is used to encode the image to achieve close to ideal. Thus, we define our LCP as

$$LCP = \frac{bxN}{H_0xN + B_C} \qquad (3)$$

where b is the number of bits per pixel, N is the total number of pixels in the image, B_C is the total number of bits needed to describe the Huffman Code, and H_0 is the zeroth order entropy of the difference image, d'(x,y).

In the case of multiband images, the encoding can be done by treating each band separately (band-by-band compression) or by treating them all together (across-band compression).

From equation (3), we see that as N becomes sufficiently large, the overhead of storing the code becomes negligible. Since the variable length Huffman code is uniquely determined by the ranking of frequencies of the grey levels in the image, a means of storing the code is to store the rank order table derived from the image histogram. In the case of a single band image with b-bit pixels, the number of bits required to store the compression code is bounded by

$$B_C \leq b + 4x2^b x(b+2) \qquad (4)$$

assuming b bits to store the median of the original image d(x,y) and (b+2) bits per entry in the rank order table of the transformed image d'(x,y) with $4x2^b$ possible entries.

In the case of multiband image taken together (across-band compression), the number of possible entries in the histogram becomes quite large and, if we store all entries (including those with zero frequencies) the overhead for code storage becomes considerable. One way around this is to store the histogram as a paired table, with entries (viz, multiband image value, frequency or rank order) only when the frequency is nonzero.

However, unique vector counting experiments (Wharton[11]) have shown that for moderately sized (say 512x512) images, the number of entries

in the histogram of a 7-band TM image are comparable to the image size itself. In these cases there is nothing to be gained by attempting across-band compression as defined above.

Error Measures

A widely used method for evaluating the quality of a compressed and reconstructed image relative to the original image is the Mean Squared Error (MSE). The MSE of band "i" of a multiband image is defined as

$$MSE_i = E[(D_i - D^r_i)^2] \simeq 1/N-1 \sum_{p=1}^{N} (D_{ip} - D^r_{ip})^2 \qquad (5)$$

where D_i and D^r_i are the data values of the i^{th} band of the original and reconstructed images, respectively; D_{ip} and D^r_{ip} are the values of the p^{th} pixel of the i^{th} band of the original and reconstructed images, respectively; E denotes the expected value; and N is the total number of pixels in the image.

The MSE_i as defined above is a single-band error measure. One could define a multiband MSE by simply summing the MSE_i over the bands. However, this definition does not account for the differences in variance between individual bands, and the values that would be obtained do not correspond to a direct conceptual notion of error. We prefer an error measure we call the multiband Root Normalized RNMSE, which we define as follows:

$$RNMSE = 1/m \sum_{i=1}^{m} \sqrt{MSE_i/VAR_i} \qquad (6)$$

where VAR_i is the variance of the i^{th} band and m is the number of bands in the multiband image. In addition to accounting for the differences in variance between individual bands, the RNMSE carries an intuitive interpretation: The RNMSE is the band average of the single-band RNMSE, which can be regarded as the mean deviation of a

reconstructed image pixel value from the corresponding original image pixel value per standard deviation of the band.

Spatially Constrained Clustering (SCC)

SCC is an iterative parallel segmentation approach that performs the "globally best merge" among spatially adjacent regions at each iteration. The globally best merge is the merge with the best similarity criterion value over all pairs of spatially adjacent regions. As implemented here, the SCC algorithm starts by initializing each pixel as a separate region. The globally best pair of regions are then merged at each iteration. The algorithm is considered to have converged when either a desired number of regions remain, or when no pair of adjacent regions is similar enough to be merged according to a predefined bound on the similarity criterion. A key aspect of any region growing approach is the similarity criterion used to determine whether or not a region should grow by merging with a neighboring region or pixel. The best similarity criterion depends upon the application. To fully explore the utility of the general SCC approach, we need to devise and test several different similarity criteria for different types of scientific image data and for various analysis procedures performed on each type of scientific image data. In the experiments reported here, the similarity criterion used is based on minimizing variance normalized mean squared error.

In the previous section we defined the mean squared error for band "i", MSE_i (see equation 5). The variance normalized mean squared error for band "i" ($NMSE_i$) is defined as

$$NMSE_i = \frac{MSE_i}{VAR_i} \qquad (7)$$

where VAR_i is the variance of band "i", as before. The similarity criterion used in our tests is the $MAX(\Delta NMSE_i)$ for each pair of spatially adjacent regions, where the maximum is taken over all bands

($1 \leq i \leq m$). For a particular pair of spatially adjacent regions, $\Delta NMSE_i$ is the change in $NMSE_i$ when the pair of regions is merged and the reconstructed image is formed by substituting the mean grey level of each region for the grey level for each pixel in the region. The globally best merge is then the pair of regions, out of all spatially adjacent regions, that minimizes the similarity criterion.

The change in $NMSE_i$, or $\Delta NMSE_i$, is calculated as follows. Define

$$\Delta NMSE_i = \frac{MSE^c{}_i - MSE_i}{VAR_i} \tag{8}$$

where $MSE^c{}_i$ is the mean squared error when regions j and k are merged, while MSE_i is the mean squared error before regions j and k are merged. Using the definitions of MSE_i, and the region mean, it is easy to derive a more fundamental version of equation (8), viz

$$\Delta NMSE_i = \frac{n_j\,(\bar{D}_{ij} - \bar{D}_{ijk})^2 + n_k(\bar{D}_{ik} - \bar{D}_{ijk})^2}{(N-1)VAR_i} \tag{9}$$

where n_j and n_k are the number of points in regions j and k, respectively, before combining, and N is the number of points in the image. D_{ij} and D_{ik} are the mean values of band i for regions j and k, respectively, before combining, and D_{ijk} is the mean value of band i for the region that would result from combining regions j and k.

We have implemented the SCC algorithm on the Massively Parallel Processor (MPP) at the NASA Goddard Space Flight Center. The MPP is a Single Instruction, Multiple Data stream (SIMD) computer containing 16,384 bit serial microprocessors logically connected in a 128-by-128 mesh array with each microprocessor have direct data transfer interconnections with its four nearest neighbors. With this massively parallel architecture, the MPP is capable of billions of operations per second.

Block Cluster Coding

The SCC segmentation is encoded for storage or transmission by a block encoding technique. The current implementation of the SCC algorithm performs segmentations on relatively small blocks of image data (from 32x32 to 128x128) because of memory limitations on the MPP. (NOTE: This limitation has been lifted in a more recent implementation in which image data and intermediate results are stored temporarily in a "staging buffer memory.") This restricts the block sizes to be used for encoding to sizes that can be evenly divided into the SCC segmentation block size. The SCC segmentation block size for the 7-band Landsat TM data used in this study was 42-by-42 pixels. This restricted the encoding block sizes to 42-by-42, 21-by-21, 14-by-14, 7-by-7, 6-by-6, 3-by-3, or 2-by-2 pixels. (The 6-by-6, 3-by-3 and 2-by-2 block sizes were not used because encoding becomes inefficient at very small block sizes.) The optimal encoding block size must be determined empirically for each application.

In performing the block cluster coding, two files are created. The region labels in each block are renumbered to use the minimal number of bits and stored as the region map file, and the mean vectors for each region in each block are stored in a region feature file. The region map file is further losslessly compressed by an appropriate method. A method we found to be effective is run-length coding along bidirectional scan lines (odd lines scanned left to right, even lines scanned right to left) with maximum run length equal to the number of samples in each line of the coding block.

This compression scheme of segmentation followed by block cluster coding was inspired by the Cluster Compression Algorithm (CCA) developed by Hilbert[2]. The main difference between CCA and our approach is the segmentation algorithm used to define the regions.

EXPERIMENTAL EVALUATION

In this section, we describe the data sets used for the evaluation of data compression procedure, the quality criteria considered, and the experimental results.

The Data Sets

A 468-by-368, 7-band subset of a Landsat TM image is one of the data sets used in this study. The subset is registered to a U.S. Geological Survey topographic map of the Ridgely Quadrangle (7.5 minute quad sheet) in Maryland. This particular image subset was chosen because the data had a sufficient variety of classes and it had a digitized ground truth map that was registered and rectified to the topographic map. In addition, this data set was used in an earlier data compression study (Ramapriyan[7]).

The other data set used in this study is a 486-by-1968 section of a 5-band Nimbus-7 CZCS image. This image was collected on October 25, 1980 over a section of the equatorial Pacific Ocean, and contained substantial numbers of scattered clouds in the western half of the image, and some very heavy clouds in the far eastern quarter of the image. The remaining quarter of the image was almost completely cloud free. The CZCS data contrasts sharply with the TM data set in that, except for the clouds, the CZCS has no obvious spatial features, while the TM data set has numerous, very obvious, spatial features. The CZCS image has no "ground truth" file.

Quality Criteria

The complexity of each dataset is measured using the band-by-band LCP. We measured the effects of data compression on both the TM and CZCS data sets by calculating the RNMSE. However, the RNMSE does not necessarily measure how much scientifically relevant information is

retained in the compressed image. For the TM data we used the scientifically relevant quality measure of classification accuracy using two different classification approaches. We have as yet not developed a scientifically relevant quality measure for the CZCS data.

The original and SCC compressed and reconstructed TM images were classified using the Maximum Likelihood Classifier (MLC) (Swain and Davis[10]). For comparison, the original TM image was also classified using the Supervised Extraction and Classification of Homogeneous Objects (SECHO) classifier (Kettig and Landgrebe[6]). The spectral classes required by these classifiers were selected from clusters generated by the ISOCLASS algorithm (Kan, Holley and Parker[4]) from NASA GSFC's Land Analysis System (LAS). The ISOCLASS algorithm was used on the entire TM test image to produce 64 clusters. It was also used on areas of the image with a high proportion of the residential and water/marshland ground cover classes to produce 16 and 32 additional clusters, respectively. The resulting 112 clusters were reduced to 31 spectral classes based on visual inspection and suppression of the most overlapping classes (within each ground cover class). For our tests, four informational classes were de-fined: Water/Marshland, Forest, Residential, and Agricultural/Domestic Grass. The means and covariance matrices of the spectral classes were then used to perform "supervised" classifications of the image using both MLC and SECHO. The classified images were mapped into the four information classes, and the resulting label images were compared pixel by pixel with the ground truth label image to obtain the classification accuracies.

Experimental Results

The LCPs for each of the seven bands of the TM test image are given in Table 1. The average LCP across all bands is 2.76. This means that the test image could be compressed by a factor of at least 2.76 (but probably not much more) without loss of any information. The large value of the average LCP is primarily due to the large LCP for

band 6 (6.47), the low resolution thermal band. The average LCP over the six reflective (full resolution) bands is 2.14.

In Table 2 we show compression factors for the TM test data set resulting from the SCC segmentation followed by encoding with blocks of various sizes and run-length coding. The compression factor tends to peak for an encoding block of size 21x21 or 42x42, with block size 14x14 trailing close behind. The "Threshold" shown in the table is the maximum NMSEi allowed in the SCC algorithm.

Table 1. Lossless Compression Potential (LCP) of the 7-Band Thematic Mapper Test Image.

Band	1	2	3	4	5	6	7	Band Ave.
LCP	2.11	2.66	2.28	2.09	1.72	6.47	1.96	2.76

Table 2. Compression Factors for Varying Encoding Block Size for the 7-Band Thematic Mapper Test Image.

Threshold	Encoding Block	CF	CF/ LCP*
0.1	42x42	6.21	2.25
"	21x21	6.36	2.30
"	14x14	6.13	2.22
"	7x7	5.02	1.82
0.2	42x42	13.6	4.93
"	21x21	13.6	4.93
"	14x14	12.5	4.53
"	7x7	8.77	3.18
0.3	42x42	23.3	8.44
"	21x21	22.5	8.15
"	14x14	19.8	7.17
"	7x7	12.2	4.42

*LCP here is the Band Average LCP = 2.76.

The RNMSE image quality measure for three $NMSE_i$ thresholds is given in Table 3 for the TM test data set. Classification accuracy evaluations are given in Table 4 for the MLC algorithm for the original and three cases of compressed TM data. For comparison, the classification accuracy is also given for the SECHO classifier on the original data. As can be seen by inspecting the accuracy figures in Table 4, for the MLC algorithm the classification accuracies are consistently as good or better for the compressed data than they were for the original data. For most cases, the MLC classification accuracies are better for the compressed data than the classification accuracies for the SECHO classifier on the original data. In fact, the classification accuracies obtained by running the MLC algorithm

on the data that was compressed by a factor of 23.3 are consistently better than the accuracies obtained by running either the MLC algorithm or the SECHO classification algorithm on the original data! We hypothesize that the SCC segmentation is behaving like a more sophisticated homogeneous object extraction procedure than that used in the SECHO classification algorithm.

Table 3. Reconstructed Image Quality for the 7-Band Thematic Mapper Test Image.

Threshold	RNMSE	CF*	CF/LCP*
0.1	0.23	6.36	2.30
0.2	0.32	13.6	4.93
0.3	0.38	23.3	8.44

*This is the maximum CF and CF/LCP over various encoding block sizes.

Table 4. MLC and SECHO Classification Accuracy Comparisons for the 7-Band Thematic Mapper Test Image (% correct classification).

Class	MLC on Original Image	MLC on CF*= 6.36	MLC on CF*= 13.6	MLC on CF*= 23.3	SECHO on Original Image
Water/Marsh	58.6	58.6	61.0	61.8	58.4
Forest	67.3	68.3	69.2	68.7	65.5
Residential	54.4	60.2	60.5	70.8	67.5
Ag./Dom. Grasses	84.9	85.9	86.6	85.9	83.5
Overall	80.0	81.1	81.8	81.3	78.7

*This is the maximum CF over the various encoding block sizes.

A subjective evaluation of the reconstructed TM images shows that areas in the original image which are relatively homogeneous, but not necessarily uniform, become completely uniform in the reconstructed images. Low contrast spatial features are often lost in the reconstructed images, but the higher contrast spatial features are retained very precisely. Even very small spatial features are retained if they have sufficient contrast relative to the surrounding area. Further experiments are needed to verify whether the SCC compression approach effectively retains all relevant scientific information in Landsat TM data. The above results seem to indicate, however, that this compression approach retains much of what would seem to be the relevant scientific information.

The LCPs for each of the five bands of the CZCS test image are given in Tables 5a and 5b. The average LCP across all bands of the entire image is 1.89. However, the LCP across all bands for a 486-by-504 pixel cloud-free section of data is 3.45.

Table 5a. Lossless Compression Potential (LCP) of the 5-Band Coastal Zone Color Scanner Test Image (Full scene - 486 lines by 1968 columns).

Band	1	2	3	4	5	Band Ave.
LCP	1.90	1.78	1.71	1.50	2.57	1.89

Table 5b. Lossless Compression Potential (LCP) of the 5-Band Coastal Zone Color Scanner Test Image (Cloud free section - 486 lines by 504 columns).

Band	1	2	3	4	5	Band Ave.
LCP	3.00	3.18	3.04	2.35	5.69	3.45

In Tables 6a and 6b (full scene and cloud-free section, respectively) we show compression factors for the CZCS test data set resulting from the SCC segmentation followed by encoding with blocks of various sizes and run-length coding. The compression factor tends to peak for an encoding block of size 21x21 or 42x42, with block size 14x14 trailing close behind.

The RNMSE image quality measure is given in Table 7 for the CZCS test data set. These results are inconclusive, but a visual inspection of the mean images produced by the SCC algorithm shows that the algorithm behaves poorly only in the vicinity of the clouds. The very high variance of the clouds cause the algorithm to segment very

coarsely in the vicinity of clouds compared to elsewhere in the image. Modifying the algorithm to use the variance of the whole image rather than just the variance of the individual segmentation blocks in calculating the variance normalized mean

Table 6a. Compression Factors for Varying Encoding Block Size for the 5-Band Coastal Zone Color Scanner Test Image (Full scene - 486 lines by 1968 columns).

Threshold	Encoding Block	CF	CF/ LCP*
0.3	42x42	8.53	4.51
"	21x21	8.94	4.73
"	14x14	8.66	4.58
"	7x7	6.93	3.67

*LCP here is the Band Average LCP = 1.89.

Table 6b. Compression Factors for Varying Encoding Block Size for the 5-Band Coastal Zone Color Scanner Test Image (Cloud free section - 486 lines by 504 columns).

Threshold	Encoding Block	CF	CF/ LCP*
0.3	42x42	3.92	1.14
"	21x21	4.14	1.20
"	14x14	4.11	1.19
"	7x7	3.64	1.06
0.5	42x42	9.86	2.86
"	21x21	10.5	3.04
"	14x14	10.9	3.16
"	7x7	7.58	2.20
0.7	42x42	24.2	7.01
"	21x21	23.7	6.87
"	14x14	21.0	6.09
"	7x7	13.0	3.77

*LCP here is the Band Average LCP = 3.45.

Table 7. Reconstructed Image Quality for the 5-Band Coastal Zone Color Scanner Test Image

Scene	Threshold	RNMSE	CF*	CF/ LCP*
Full scene	0.3	0.28	8.94	4.73
Cloud-free sec.	0.3	0.11	4.14	1.20
Cloud-free sec.	0.5	0.14	10.9	3.16
Cloud-free sec.	0.7	0.17	24.2	7.01

*This is the maximum CF or CF/LCP over the encoding block sizes.

squared error (equation 7), should improve the behavior of the algorithm. However, going back to our original premise of tailoring our compression approach to the characteristics of the data, we question the utility of pursuing this approach further for CZCS data.

Except in the vicinity of clouds and land masses, the CZCS image data generally has very little spatially variability. In the case of ocean images, there are no distinct boundaries as seen in the case of the land images (e.g., between a forested area and an agricultural field as found in TM image data). Since the forte of the SCC approach is the preservation of boundaries between contrasting regions, it may make little sense to apply it to data, such as CZCS data, where such boundaries aren't important. The only contrasting boundaries found in CZCS data are between clouds and ocean, and land and ocean. The users of CZCS data routinely mask out and discard cloudy data and data collected over land using simple thresholding schemes. A more appropriate compression approach may be to mask out the cloudy data and data collected over land in the same way done routinely now by the users of the data and use some variation on run-length encoding to compress the remaining data.

CONCLUDING REMARKS

Given the resolutions and data rates expected from the remote sensors to be flown during the next two decades, it will be necessary to consider both lossless and lossy data compression techniques to keep the transmitted data rates and archived data volumes within manageable limits. Lossy techniques, wherein the raw data bits cannot be exactly reconstructed, require careful studies in coordination with scientific users to determine whether most of relevant information for a given application is retained. Several such studies are needed in selected disciplines and application areas. In this paper, we have presented one such study using a compression technique which can be a precursor to analysis and information extraction. The compression technique is based on SCC and subsequent local encoding of regions. This is an

adaptive vector quantizer where very short codebooks are needed and are developed "on the fly" for local rectangular regions.

Since the SCC is a segmentation technique, it is a very useful precursor to the analysis of image data with significant amount of detail. The algorithm can be controlled with a single parameter to obtain different degrees of segmentation (retaining different levels of detail) and corresponding compression ratios. In our case study, we have explored the use of the SCC with two types of data. The first, a land image from the Landsat TM, has considerable spatial detail while the second, an ocean image from the CZCS has no recognizable features except clouds (which are usually suppressed in performing any analyses).

For the case of the TM data, we found that land cover classification accuracies are higher with compressed data than with raw data even up to compression ratios over 20. This agrees with results from earlier studies with other compression techniques such as the Cluster Coding Algorithm (CCA). Further experiments are needed to verify whether all relevant scientific information is retained by such compression techniques. However, the present study confirms that for land cover classification applications significantly compressed data can be used directly, and in many cases, more usefully than raw data. In the case of CZCS data, the image distortion measures and subjective image evaluation show that compression ratios of 4 to 24 can be achieved with relatively small distortions. Further experiments on derived geophysical parameter data are needed to examine the impact of compression on the analysis of CZCS data. However, given the nature of the CZCS data it is probably more fruitful to consider other compression techniques such as linear prediction and run length encoding.

REFERENCES

1) Chen, T.M., D.H. Staelin and R.B. Arps, "Information Content Analysis of Landsat Image Data for Compression,", IEEE Transactions on Geoscience and Remote Sensing, vol. GE-25, #4, pp 499-501, July 1987.

2) Hilbert, E. E., "Cluster Compression Algorithm: A Joint Clustering/Data Compression Concept", JPL Publication 77-43, Jet Propulsion Laboratory, Pasadena, CA, 1977.

3) Jain, A. K., "Image Data Compression: A Review", Proc, IEEE, Vol. 69, pp. 349-389, 1981.

4) Kan, E. P., W. A. Holley and H. D. Parker, Jr., "The JSC Clustering Program ISOCLS and its Applications", Proceedings of the 1973 Machine Processing of Remotely Sensed Data Symposium, (IEEE Catalog No. 73 CH 0834-2 GE), Oct. 16-18, 1973, pp. 4B-36 thru 4B-45.

5) Kauth, R.J., et al, "BLOB: An Unsupervised Clustering Approach To Spatial Preprocessing of MSS Imagery", Proceedings of 11th International Symposium of Remote Sensing of the Environment, Environmental Institute of Ann Arbor, Michigan, 1977.

6) Kettig, R. L. and D. A. Landgrebe, "Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects", IEEE Trans. on Geoscience Electronics, pp. 19-26, 1976.

7) Ramapriyan, H. K., J. C. Tilton and E. J. Seiler, "Impact of Data Compression on Spectral/Spatial Classification of Remotely Sensed Data", Advances in Remote Sensing Retrieval Method, Deepak, Fleming and Chahine, Eds., pp. 687-706, 1985.

8) Rosenfeld, A. and A. C. Kak, "Digital Picture Processing", 2nd Ed., Acad. Press, NY, pp. 181-182, 1982.

9) Shannon, C. E., "The Mathematical Theory of Communication", Bell System Technical Journal, Parts I and II, pp. 379-423, 623-656, 1948.

10) Swain, P. H. and S. M. Davis (Eds.), "Remote Sensing: the Quantitative Approach", McGraw-Hill, NY, 1978.

11) Wharton, S. W., "An Algorithm for Computing the Number of Distinct Spectral Vectors in Thematic Mapper Data", IEEE Trans. on Geoscience and Remote Sensing, pp. 294-302, 1985.