The Hyper-spectral Image Compression Based on K-Means Clustering and Parallel Prediction Algorithm*

Wenbin Wu[1,*], *Yue* Wu[1], and *Jintao* Li[2]

[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]School of mechanical electronic & information engineering, China University of Mining and Technology (Beijing), Beijing, China

Abstract. In this paper, we propose a lossless compression algorithm for hyper-spectral images with the help of the K-Means clustering and parallel prediction. We use K-Means clustering algorithm to classify hyper-spectral images, and we obtain a number of two dimensional sub images. We use the adaptive prediction compression algorithm based on the absolute ratio to compress the two dimensional sub images. The traditional prediction algorithm is adopted in the serial processing mode, and the processing time is long. So we improve the efficiency of the parallel prediction compression algorithm, to meet the needs of the rapid compression. In this paper, a variety of hyper-spectral image compression algorithms are compared with the proposed method. The experimental results show that the proposed algorithm can effectively improve the compression ratio of hyper-spectral images and reduce the compression time effectively.

1 INTRODUCTION

The parallel is a kind of way to improve the processing efficiency, and it uses multiple processing units to deal with the problem. Parallel means that can be calculated or operated at the same time. The parallel computing needs parallel data processors, which can divide an application into multiple sub tasks. They are sent to different processors, and processors work together to accomplish tasks. That will speed up the computation, or that will expand the size of problems.

The hyper-spectral imagery is a three-dimensional image taken by the imaging spectrometer. Because the large size of the hyper-spectral imagery, the traditional compression method is hard to meet the requirements of the high-speed encoding and decoding. Research on the parallel compression technology is not only to satisfy the demands of the engineering develop-ment, but also to resolve practical problems.

The research of hyper-spectral images parallel compression has risen in recent years. Lena Chang proposed a parallel compression method based on group and region in 2011. It contains two algorithms, which are clustering signal subspace projection (CSSP) and the maximum correlation band clus-tering (MCBC). Lucana Santos found a loss compression algorithm by GPU implementation in 2012.

Above methods were about the hyper-spectral imagery parallel compression. We propose a new method to compress hyper-spectral image. We will segment a hyper-spectral imagery, and break up the whole work into parts. Thus we will achieve the goal of the parallel compression of the hyper-spectral imagery. Based on the two order linear prediction compression algorithm, we use K-Means clustering algorithm to divide the original hyper-spectral image to sub image. According to the correlation between the bands of the sub image, we design adaptive predictive compression algorithm, which can provide the basis for the parallel compression of hyper-spectral image.

2 HYPER-SPECTRAL IMAGERY

The hyper-spectral remote sensing technology is a field of geophysical methods in advanced. In recent years, with the development of semiconductor technology, the hyper-spectral imagery caught by imaging spectrometers, as shown in Fig. 1, expanded in the scale of spatial resolution and spectral resolution. Thus, the data size expands constantly, and one scene data is from hundreds of MB to several GB or even a dozen of GB. It makes a lot of challenges to manage and transmit data.

In order to meet the needs of hyper-spectral imagery data management and transmission, hyper-spectral imagery data compression as an effective way has got a wide range of applications. The research of the compression of hyper-spectral imagery has made some progress, but there is still some room for improvement. The main deficiencies are that the better effect it compresses, the longer time it often takes. Therefore, how to improve the compression efficiency is an important part of engineering application. Development of parallel computing provides a useful way to improve the efficiency of hyper-spectral imagery compression. We will focus on the technology and methods of hyper-spectral imagery parallel compression, and gives a

[*] *Wenbin* Wu: zxjun@cumtb.edu.cn

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

parallel algorithm for hyper-spectral imagery compression based on prediction.

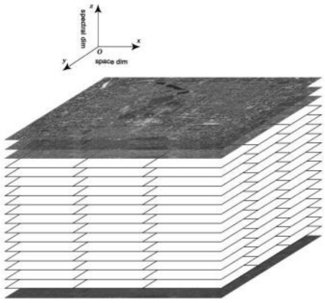

Fig. 1. The structure of hyperspectral image.

3 K-MEANS CLUSTERING ALGORITHM

In 1950s, Steinhaus et al proposed clustering algorithm K-Means independently in their different scientific research. The algorithm was widely used in communication, mechanical, biological, medical, financial, construction and other fields quickly.

For N dimensional data points of the given data set is $X=\{x_1,x_2,\ldots,x_i,\ldots,x_n\}$, where $x_i \in R^N$, and the number of subsets is K. K-Means clustering algorithm gets data organized into K partition $C=\{C_k\}$, $k=1,2,\ldots,N$. Each partition represents a class of C_k, each C_k has a center of category μ_i . We choose Euclidean distance as the similarity and distance criterion, and calculate each point to the cluster center μ_i square distance.

$$J(C_K) = \sum_{X_i \in C_i} \|x_i - \mu_k\|^2 \tag{1}$$

Our goal is to make all kinds of clustering square distance $J(C) = \sum_{k=1}^{K} J(C_k)$ to be minimized.

$$J(C) = \sum_{k=1}^{K} J(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_i} \|x_i - \mu_k\| = \sum_{k=1}^{K} \sum_{i=1}^{n} d_{ki} \|x_i - \mu_k\|^2 \tag{2}$$

In above formula,

$$d_{ki} = \begin{cases} 1, x_i \in C_i \\ 0, x_i \notin C_i \end{cases} \tag{3}$$

According to the least square method and the Lagrange mean value theorem, the clustering center μ_k should be taken as the median value of C_k class for each data point. The K-Means algorithm starts from an initial K category, and then each data point are assigned to each class, in order to reduce the square space. Because with the increasing number of categories K, the square distance of K-Means clustering algorithm tends to diminish. (When K=n, J(C)=0) Thus, the square distance of a category number K, has a minimum value.

K-Means algorithm is an iterative process, and the purpose is to make the square distance to the cluster center to be minimized. The algorithm process contains 4 steps, shown in Fig. 2.

Through the K-Means algorithm, we divide the hyper-spectral image into several sub images in the spatial dimension. We use a two-dimensional matrix to record the position of each element before K-Means clustering algorithm. In order to facilitate the processing, we will compress the sub images at the same time. If the hyper-spectral image is C (i, j, k), where i, j is the spatial dimension and k is the spectral dimension. The Two-Dimensional Processing obeys the order of C(i, j), C(i, j+1),... , C(i+1, j), C(i+1, j+1),... Thus we convert a two-dimensional band of a sub hyper-spectral image into a one-dimensional vector. Then the sub image will be converted into a two-dimensional image, which includes the dimensions of spectrum and space.

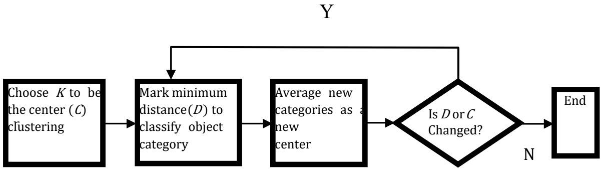
Y
Choose K to be the center (C) clustering
Mark minimum distance(D) to classify object category
Average new categories as a new center
Is D or C Changed?
N
End

Fig. 2. Process of K-Means clustering algorithm

4 ADAPTIVE PREDICTIVE COMPRESSION ALGORITHM

In early twentieth Century, with successive efforts of Kolmogorov, Turing, Shannon et al., probability theory had made great achievements in source coding. The branch of probability theory has cleared the obstacle for future research on data compression. Linear prediction ideas appeared in the nineteenth Century. In 1966, S. Saito and F. Itakura first proposed the linear predictive coding. They put forward a method for automatic phoneme recognition, and this method uses the maximum likelihood estimation to code the speech. Prediction method is one of the oldest and most effective compression methods, which can be used to remove the correlation of hyper-spectral image. It receives errors by coding the restructured values and fore values to achieve compression. Differential pulse code modulation (DPCM) and its improved algorithm are used constantly.

In this paper, we propose an adaptive compression algorithm. At first, we group the spectra of sub high-spectral image according to the correlation. Considering the increase of the reference information, the standard of the grouping should be chosen. If spectral correlation coefficient is greater than 0.9, the adjacent spectrum can be predicted by two spectral bands ahead. If the spectral correlation coefficient is less than 0.9, we calculate the ratio of the absolute sum of residual error and the absolute sum of sub image. We call this ratio as the absolute ratio. When the absolute ratio is larger than 1, we will keep the original data directly, or compress by two spectral ahead.

The traditional two order linear predictive compression is presented in Fig. 3. First of all, we use the band 1 and band 2 to predict band 3, similarly, then we use the band 2 and band 3 to predict band 4.

Hyper-spectral image of the second order predictive compression will be used with constant coefficient of the second order optimal linear predictor to predict, as shown in equation 4. For the critical period, the first two spectral segments are predicted by the first spectral bands.

$$E^2 = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(g(i,j) - a\cdot f_1(i,j) - b\cdot f_2(i,j) - c)^2 \qquad (4)$$

G (i, j) in the formula is the element of the predicted spectrum, f_1 (i, j), f_2 (i, j) as the reference spectral element, b, c and a are the parameters. When the mean square error is minimum, the equation of the formula 5 is satisfied.

$$\begin{cases} \dfrac{\partial E^2}{\partial a} = 0 \\ \dfrac{\partial E^2}{\partial b} = 0 \\ \dfrac{\partial E^2}{\partial c} = 0 \end{cases} \qquad (5)$$

The values of c, b and a can be calculated by the formula 5:

$$\begin{cases} a = \dfrac{\left[r(f_2,f_2) - r(f_2)r(f_2)\right]\left[r(g,f_1) - r(g)r(f_1)\right] - \left[r(f_1,f_2) - r(f_1)r(f_2)\right]\left[r(g,f_2) - r(g)r(f_2)\right]}{\left[r(f_1,f_1) - r(f_1)r(f_1)\right]\left[r(f_2,f_2) - r(f_2)r(f_2)\right] - \left[r(f_1,f_2) - r(f_1)r(f_2)\right]\left[r(f_1,f_2) - r(f_1)r(f_2)\right]} \\ b = \dfrac{r(g,f_1) - r(g)r(f_1) - a\cdot\left[r(f_1,f_1) - r(f_1)r(f_1)\right]}{r(f_1,f_2) - r(f_1)r(f_2)} \\ c = r(g) - a\cdot r(f_1) - b\cdot r(f_2) \end{cases} \qquad (6)$$

Which $r(f_1,g) = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} f_1(i,j)g(i,j)$, and $r(g) = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} g(i,j)$.

The study found that the effect of linear prediction with constant coefficient was better than without constant coefficient. The spectral correlation of hyper-spectral image was better considered, and the entropy of residual error was reduced.

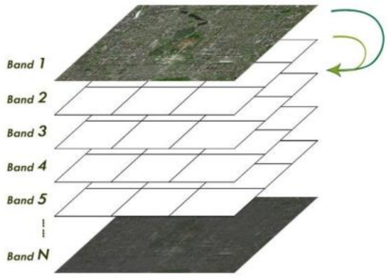

Fig. 3. Hyper-spectral image prediction compression sequence

After the adaptive prediction compression, the high-spectral image includes the residual error and the prediction coefficient and some of the hyper-spectral image. We compress the residual error by the LZW encoding algorithm. In summary, we compress the hyper-spectral images.

5 PARALLEL PREDICTION ALGORITHM

Parallel computing consists of parallel methods between time and space. The parallel of time refers to the computer pipeline, and the parallel of space refers to calculation using multiple processors concurrently. Parallel computing research in science is a parallel problem of space. MATLAB packages MPI into parallel computing toolbox. To the use of the parallel computing toolbox, people do not need to consider the detail of MPI. On the one hand, it is convenient for the user to transplant an existing MATLAB program, on the other hand, advanced users can also develop further program through MPI interface. After continuous improvement, the parallel computing toolbox can be regarded as the core of the multi-processor or the same processor as worker, through the programming control of each worker implementation of the work, through parallel computing to solve complex computing and large data volume.

By the method mentioned above, we can compress each sub image concurrently and save the time of compression. First of all, we use the K-Means clustering algorithm to segment the hyper-spectral image into several sub images. Then, we assign each sub image to different worker. Finally, we will merge the results of different workers.

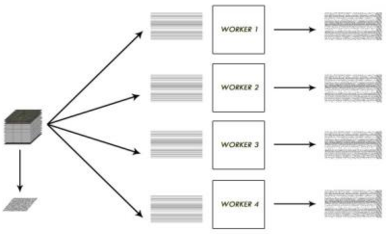

Fig. 4. Parallel compression process of hyper-spectral image

6 RESULTS ANALYSIS

Through the two order linear parallel prediction compression algorithm, we can speed up the efficiency of the hyper-spectral image compression. In the experiment, we use Hyperion hyper-spectral image, which is shown in table 1.

Table 1. Hyperion L1 data description.

Data Name	Hyperion L1
Wavelength	356 ~ 2577 (nm)
Bands Number	242
Image Element	30 (m)
Image Size	256×6460
VNIR Band	1~70 (356 ~ 1058 nm)
SWIR Band	71~242 (852 ~ 2577 nm)
Data Type	2 Bytes
Image Format	BIL
Byte Order	Network (IEEE)
File Size	800,427,520 (bytes)

In parallel compression, we will first use K-Means clustering, and the original hyper-spectral image is divided into 2, 4 or 8 clusters. The compressing time will no more than 50%, 25% or 12.5% of the time without parallel algorithm in theory. Although the implementation of clustering algorithm itself also takes a period of time, the parallel compression will still save time. After clustering, we obtain several two dimensional sub images. In order to judge whether the size of the sub images are close to each other, we put forward the max ratio of sub images. The ratio considers the bytes of sub images, and must be larger than 1.

$$P = \frac{I_{\max}}{I_{\min}} \qquad (7)$$

Among them, $I_{\max}$ is the number of bytes from the largest sub image, $I_{\min}$ is the number of bytes from the smallest sub image, P is the max ratio of sub images.

Costing time of different numbers of clusters:

Table 2. Costing time of different numbers of clusters.

clustering number	compression time (s)	max ratio of sub images
2	2.28	1.89
4	2.76	2.37
8	3.31	2.90

By K-Means clustering, we get some sub images. The sub images can be compressed by the parallel compression algorithm based on prediction.

We compare the parallel compression algorithm in this paper with the optimal recursive bidirectional prediction compression algorithm, LS-JPEG algorithm and JPEG2000 algorithm. The optimal recursive bidirectional prediction compression algorithm is one of the most widely used high-spectral prediction compression algorithms. Comparison of different compression algorithms is shown in Table 3:

Table 3. Comparison of different algorithms.

Compression algorithm	Compression ratio	Compression time (s)
Clustering prediction algorithm (2 Cluster)	3.88	12.10
Clustering prediction algorithm (4 Cluster)	3.91	8.81
Clustering prediction algorithm (8 Cluster)	4.02	6.56
Recursive bidirectional prediction algorithm	3.19	20. 74
JPEG- LS	2.05	35.18
JPEG2000	1.97	42.04

7 CONCLUSION

From the above analysis and processing results, we can see the proposed algorithm based on clustering and parallel prediction can make full use of the correlation coefficient of high-spectral images. Compared with other classical hyper-spectral image compression algorithm, we improve the compression ratio and save the time of compression. In practical engineering, it has a good prospect of application.

ACKNOWLEDGEMENTS

This work was supported by the National Science & Technology Pillar Program during the 12th Five-year Plan Period: Research on the dedicated equipment and demonstration for land consolidation and quality detection. (Grant No: 2011BAD04B05) and Geological survey of the China Geological Survey Bureau: Optimization design technology for remote sensing sensor for mineral and energy exploration applications (Grant No: 1212011120221).

References

1. C. C. Chang and C. J. Lin, LIBSVM, *A library for support vector machines,* http:// www.csie.ntu.edu.tw/ cjlin/libsvm (2001)
2. Ingram, R. N., et al., International Journal of Remote Sensing 25(22), *An automatic nonlinear correlation approach for processing of Hyper-spectral images.* 4981-4998. (2004)
3. Du, Q., et al., IEEE Geoscience and Remote Sensing Letters 6(4), *Segmented Principal Component Analysis for Parallel Compression of Hyper-spectral imagery.* 713-717. (2009)
4. Plaza, A., et al., Remote Sensing of Environment 113, *Recent advances in techniques for Hyper-spectral image processing.* S110-S122. (2009)
5. Plaza, A., et al., Ieee Signal Processing Magazine 28(3), *Parallel Hyper-spectral image and Signal Processing.* (2011)
6. Lena C., et al., IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 4(3), *Group and region based parallel compression method using signal subspace projection and band clustering for hyperspectral imagery.*565-578. (2011)
7. Lucana S., et al., Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing, 4th Workshop on Hyperspectral Image and Signal Processing, *GPU implementation of a lossy compression algorithm for hyperspectral images.* (2012)
8. Zhao, X. and X. Qiao, Spectral Characteristics Research of the Hyperspectral Image Based on the Correlation Matrix. Information Science and Engineering (ISISE), 2012 International Symposium on. (2012)
9. Santos, L., et al., Journal of Applied Remote Sensing 7, *Lossy Hyper-spectral image compression on a graphics processing unit: parallelization strategy and performance evaluation.* (2013)
10. Taher, A., et al., Hyperspectral image segmentation using a cooperative nonparametric approach. Image and Signal Processing for Remote Sensing Xix. L. Bruzzone. 8892. (2013)
11. Santos, L., et al., Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6(2), *Highly-Parallel GPU Architecture for Lossy Hyper-spectral image Compression.* 670-681. (2013)
12. Sanchez, S., et al. PARALLEL HYPER-SPECTRAL IMAGE COMPRESSION USING ITERATIVE ERROR ANALYSIS ON GRAPHICS PROCESSING UNITS. 2012 Ieee International Geoscience and Remote Sensing Symposium: 3474-3477. (2012)