TREE IDENTIFICATION USING A DISTRIBUTED K-MEAN CLUSTERING ALGORITHM

K. T. Fan[1]*, Y. C. Tzeng*[1]*, Y. J. Su*[1]*, Y. F. Lin*[1]*, M. H. Hsu*[1]*, and K. S. Chen*[2]

[1]Department of Electronic Engineering, National United University, Maio-Li, Taiwan
john@nuu.edu.tw
[2]Center for Space and Remote Sensing Research, National Central University, Chung-Li, Taiwan
dkschen@csrsr.ncu.edu.tw

1. INTRODUCTION

Trees play an important role in maintaining environmental conditions suitable for life on the earth [1-2]. It is very important to classify the tree type for the forest maintenance. Details of forest can be obtained by field investigation; however it is almost impossible to constantly measure large area, especially on mountain area where field investigations are usually costly and laborious. On the other hand, remote sensing is helpful for investigation of wide area of forest. With the advent of high spatial resolution remote sensing sensors, our ability has greatly increased for tree species identification. However, high-resolution imagery presents a new challenge over, in that a huge amount of data must be analyzed to identify tree species [3]. Considering the amount of data in need of processing and the high computational costs required by image processing algorithms, conventional computing environments are simply impractical. Therefore, it is necessary to develop techniques and models for efficiently processing large volume of remote sensing images.

2. THE EXPERIMENTAL ENVIRONMENT

Cluster computing is a form of distributed computing which provides an advanced computing and sharing model to solve large and computationally intensive problems [4]. In contrast to grid computing, the nodes on a cluster are homogeneous and are networked in a tightly-coupled fashion. Generally, the nodes are configured identically and are all on the same subnet of the same domain. Only the cluster application run on a cluster node, so each node on a cluster is a dedicated resource. One advantage available to cluster computing is the Message Passing Interface (MPI) which is a programming interface that allows the distributed application instances to communicate with each other and share information. To identify tree species from high resolution remote sensing images, a cluster computing environment was established in this study. Our cluster computing environment was achieved by Windows HPC Server 2008 on a 4-node system. HPC server supports a cluster of servers that includes a single head node and one or mode compute nodes. The head node controls and mediates all access to the cluster resources and is the single point of management, deployment, and job scheduling for the compute cluster. Each node was a multi-core PC which contained a 64-bit Intel Celeron Core 2 Quad CPU. Multi-threading technique was adopted to fully explore its multi-core capability. Besides, a 64-bit programming environment and MPI development were supported in Visual Studio 2008 Professional.

3. RESULTS AND DISCUSSION

The test site, covered 2.548 km in width and 30.825 km in length, is located about Tai-Chuang City in Taiwan. Forest covers a large area of the test site. A long-term field investigation of the test site was used to validate the classification results. A test image was acquired by the ADS40 (Airborne Digital Sensor) system over the test site. The test image had four optical bands (R, G, B, and IR). With 20.455 cm pixel spacing, an image with a huge size of 12458 samples and 150831 lines was obtained. Because the test image was so huge that it was very difficult to process the whole image at a time. Therefore, the test image was first partitioned into hundreds of manageable sub-images. Scheduled by the head node, the sub-images were then distributed to compute nodes for processing. A distributed K-mean clustering algorithm with undetermined number of class was applied to each compute node [5]. Started from a single class, for each sub-image, the classes were continuously divided until the cluster validity index reached its minimum value. The cluster validity index [6] is defined as

$$XB = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n_i} \left\| x_{ik} - \mu_i \right\|^2}{n\left[\min_{i \neq j} \left\{ \left\| \mu_i - \mu_j \right\|^2 \right\} \right]} \tag{1}$$

where c is the total number of classes, μ_i is the mean center of class i, n_i is the number of pixels in class i, x_{ik} is the k-th pixel of class i, and n is the total number of pixels. In order to get better classification accuracy, not only its spectral information but also the derived NDVI and texture information were applied. Also, in each sub-image, non-vegetation area was masked according to its NDVI value to relieve the computational burden. Since K-mean clustering algorithm is an unsupervised algorithm, the same class id classified in each sub-image may represent different class type. To ensure the same class type having the same class id, a merging process was hence followed to combine the classification results from all of the sub-images [7]. The similarity between two classes i and j can be defined as

$$S_{ij} = \frac{d_i + d_j}{\left\| \mu_i - \mu_j \right\|} \tag{2}$$

$$d_i = \sqrt{\sum_{k=1}^{n_i} \left\| x_{ik} - \mu_i \right\|^2} \tag{3}$$

in which d_i indicates the extent that class i spreads. If the similarity is greater than a predetermined threshold then the two clusters should be merged.

A promising result was obtained. Compared to the field investigation, tree species of the test site were properly identified. Besides, the merging process works successfully, no abrupt change was observed at the boundary between adjacent sub-images. In addition, great improvement in computation time was obtained. The distributed K-mean clustering algorithm on our cluster computing environment performed about ten times faster than stand-alone alternatives. By adding more compute nodes to our cluster computing environment, further improvement in computation time is expected.

4. REFERENCES

[1] R. Komura and K. Muramoto, "Classification of Forest Stand Considering Shapes and Size of Tree Crown Calculated from High Spatial Resolution Satellite Image," *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, (IGARSS 2007), July 23-27, Barcelona, Spain, pp. 4356-4359, 2007.

[2] P. Meyer, K. Staenz, and K. I. Itten, "Semi-automated Procedures for Tree Species Identification in High Spatial Resolution Data from Digitized Colour Infrared-Aerial Photography," ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 51, pp. 5-16, 1996.

[3] A. Cheriyadat, E. Bright, D. Potere, and B. Bhaduri, "Mapping of Settlements in Hige-Resolution Satellite Imagery Using High Performance Computing," *GeoJournal*, Vol. 69, pp. 119-129, 2007.

[4] Z. Shen, J. Luo, G. Huang, D. Ming, W. Ma, and H. Sheng, "Distributed Computing Model for Processing Remotely Sensed Images Based on Grid Computing," *Information Science*, Vol. 117, pp. 504-518, 2007.

[5] H. S. Rhee and K. W. Oh, "A Validity Measure of Fuzzy Clustering and Its Use in Selecting Optimal Number of Clusters," *Proceedings of the fifth IEEE International Conference on Fuzzy Systems*, September 8-11, New Orleans, LA, USA, pp. 1020-1025, 1996.

[6] X. L. Xie, and G. Beni, “A Validity Measure for Fuzzy Clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, pp. 841-847, 1991.

[7] X. Xiong, K. L. Chan, and K. L. Tan, “Similarity-Driven Cluster Merging Method for Unsupervised Fuzzy Clustering,” *Proceedings of the 20th International Conference on Uncertainty in Artificial Intelligence*, (UAI’2004), July 7-11, Branff, Canada, pp. 611-618, 2004.