Journal of the Korean Society of Remote Sensing, Vol.16, No.2, 2000, pp.145~156

Clustering Algorithm Using Hashing in Classification of Multispectral Satellite Images

Sung-Hee Park*, Hwang-Soo Kim, and Young-Sup Kim*****

Computer & Software Technology Laboratory, ETRI*

Dept. of Computer Science, Kyungpook National University**

School of Computer Science and Electronic Engineering, Handong University***

Abstract : Clustering is the process of partitioning a data set into meaningful clusters. As the data to process increase, a faster algorithm is required than ever. In this paper, we propose a clustering algorithm to partition a multispectral remotely sensed image data set into several clusters using a hash search algorithm. The processing time of our algorithm is compared with that of clusters algorithm using other speed-up concepts. The experiment results are compared with respect to the number of bands, the number of clusters and the size of data. It is also showed that the processing time of our algorithm is shorter than that of cluster algorithms using other speed-up concepts when the size of data is relatively large.

Key Words : Classification, Clustering, Hashing, Multispectral Satellite Images.

1. Introduction

As satellite images have higher spatial resolution, their application fields are expanded, such as, environmental monitoring, wood fire watching, national defence, urban planning, transportation and geographic information systems(GIS), etc. To extract information from these images, classification process is used. Multispectral classification is one of the most frequently used methods for extraction of information(Jensen, 1996). The general steps required to extract land cover information from satellite images are summarized in Fig. 1.

Input data are preprocessed and enhanced images. The preprocessing such as geometric and radiometric correction is a procedure of correcting distortion, degradation, and noise introduced during imaging process. The next step is to determine a purpose of classification, such as, surveying degree of water pollution, vegetation classification, geological investigation, land use/land cover classification, and so on. Classification methods can be categorized into supervised classification and unsupervised classification depending on clustering method. In

Received 1 April 2000; Accepted 15 May 2000.

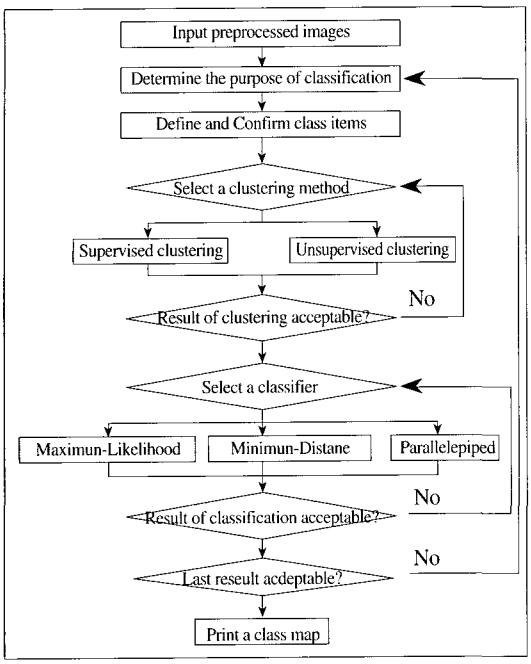


Fig. 1. Step of Classification of Satellite Imagery

selecting appropriate clustering method in supervised classification, appropriate values of parameters are to be selected, such as number of desired cluster, number of initial cluster, size of seed, size of sampling and minimum size of pixel. In unsupervised classification, a user selects one of ISODATA algorithm, k-means algorithm, fuzzy c-means algorithm, statistical clustering algorithm, clustering algorithm using neural network, specifying appropriate values of parameters.

Clustering accuracy is evaluated by covariance, Euclidean distance, Jeffries-Matusita distance, or scatter plot analysis.

Classifiers can be categorized into parallelepiped algorithm, minimum-distance algorithm, maximum- likelihood algorithm and fuzzy algorithm. The classifier is selected according to the purpose of classification and the classification result may show much difference.

Classification accuracy can usually be evaluated by means of aerophotos, maps and especially, GPS(Global Positioning Systems) for accurate position estimation. It is desirable to compare the results of classification to in situ information. The accuracy can be improved using image tracking algorithm using database and the result can be compared with database to assess accuracy indirectly. If the result of classification is different from the result of interpretation, the class items need to be redefined.

The former half of classification, is performed iteratively, while the latter half of classification is processed at one pass. Therefore, an iterative clustering is time consuming and many efforts have been made to reduce computation time for Euclidean-based minimum distance clustering algorithm(Hodgson, 1988).

In this paper, we make the survey of conventional speed-up techniques used for Euclidean minimum distance clustering algorithm in order to find the weak point of them. We propose a computationally efficient algorithm using hashing. We examine it with respect to the number of bands, the number of clusters and the size of data. It is showed that the processing time of our algorithm is shorter than that of cluster algorithms using other speed-up concepts when the size of data is large.

2. Conventional Speed-Up Techniques for Clustering

The conventional techniques for clustering are partial sum, the nearest-neighbour distance, auto-correlation and sorting(Bryant, 1989; Venkateswarlu and Raju, 1992).

1) Partial Sum

The partial sum is the partial distance of total

distance between a data point and a cluster center. While we compute the distance from an unknown pixel to a cluster, the partial distance is compared to the current minimum distance from the pixel to any cluster centers. If the partial sum is greater than the current minimum distance, we can stop the iteration without computation any more for the cluster currently considered to reduce time. This concept is useful for the multispectral satellite images (for example, hyper-spectral images have fifty to one hundred bands). However, this technique have problem with the overhead of computing distances to all the cluster centers.

2) NND: Nearest Neighbor Distance

A nearest neighbor distance(NND) of a cluster is the minimum distance to any cluster in n-dimensional feature space(Venkateswarlu and Raju, 1992). When calculating the distance from an unknown pixel to a cluster, one-half of NND of the cluster is compared to the distance from the pixel to the cluster. If the pixel-to-cluster distance is less than the one-half of NND of the cluster, we can stop the iteration without consideration any other clusters because this pixel is surely assigned to the currently considered cluster. Therefore, the concept of NND is to assign an unknown pixel to a near cluster without further comparison.

3) Auto-Correlation Property

The auto-correlation property in an image means that two spatially neighbouring pixels may be similar in brightness, hence the probability that these pixels are assigned to the same cluster is high(Tou and Gonzalez, 1974). Therefore, we can exploit this property to guess that a currently considered pixel belong to the same cluster to which previously considered pixel belongs with high possibility. When the auto-correlation property is combined with partial sum and NND techniques mentioned earlier, the computational efficiency increases.

4) Sorting

The principle of sorting is to sort clusters in order of the possibility for an unknown pixel to belong to one of them[1]. However, since the possibility is not known, an alternative criterion for sorting is needed. A commonly used criterion is the number of cluster members. In the case of image data, clusters may be sorted by the number of their members for immediately preceding scan line. This technique also can be combined with partial sum, NND, and auto-correlation property to increase computational efficiency.

3. Proposed Technique: Hashing Technique

The weakness of the conventional techniques to reduce computation time for distance is that the distance to all clusters need to be calculated to decide which cluster is the nearest to the current data, which it is time consuming process. To solve this problem, the hashing technique is introduced in this paper.

1) Clustering algorithm using hashing

Hashing is a method of searching appropriate bucket by hashing function using the key obtained from the data under consideration. The advantage of hashing is that it does not need to search the whole table, but it directly obtains the location via hashing function. When hashing is applied to the clustering algorithm, this property allows the nearest cluster to be found without computing the distance to all clusters. However, we need to

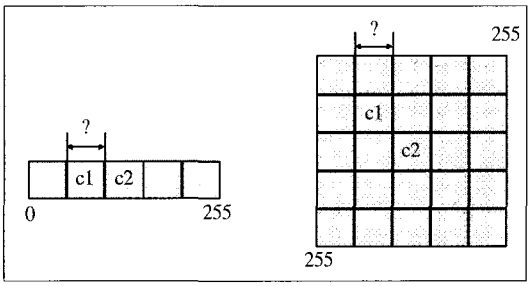
divide the feature vector space into bins of appropriate size. The bins are visualized in Fig. 2. Since the size of bin plays very important role in this process, determining its size is more closely examined in the next section.

The clustering procedure using hashing is presented in Fig. 3. In the following subsections, hashing is applied to one-dimensional data first, then extended to the multi-dimensional data.

(1) Application of hashing for one-dimensional data

The algorithm given in Fig. 3 can be applied to one-dimensional data as follows.

a. Read in the N data points and C (1<C≪N) cluster centers.
b. Obtain the bin indices from cluster centers. An index of bin is obtained by positioning each cluster center in a bin whose vector space contains the center. We must consider for each bin not to include more than one cluster center, since the hashing function can not determine the possible cluster which the data belongs to. Thus, the maximal bin size should be smaller than the smallest NND. However, To make sure that each bin contains only the center, the bin size should be smaller than half of the minimum NND.
c. Compute the hashing function using the centers.
d. Position the cluster center in the bin pointed by the function.
e. Get the key from data.
f. Compute the hashing function to get the bin index. Seventh, check if the bin contains a cluster center. If it does, put the data into the cluster; otherwise, compute the distances to all the other clusters as usual. We could compute them only for more possible clusters by storing the neighboring clusters in each bin; but this is not implemented in this study.



(a) 1-band data (b) 2-bands data

Fig. 2. Formation of bin

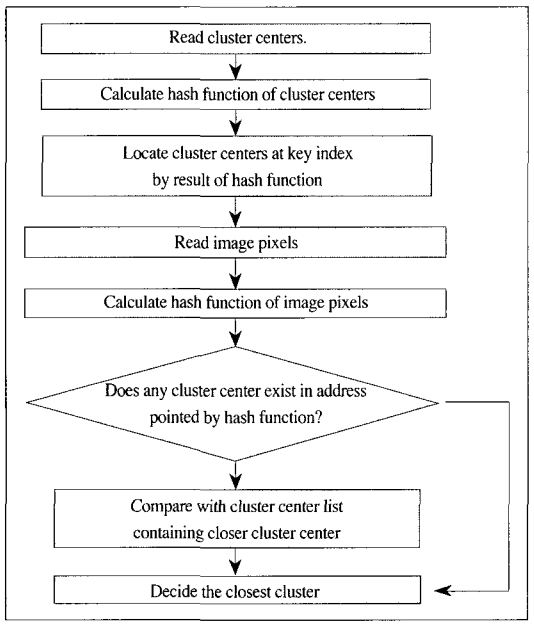


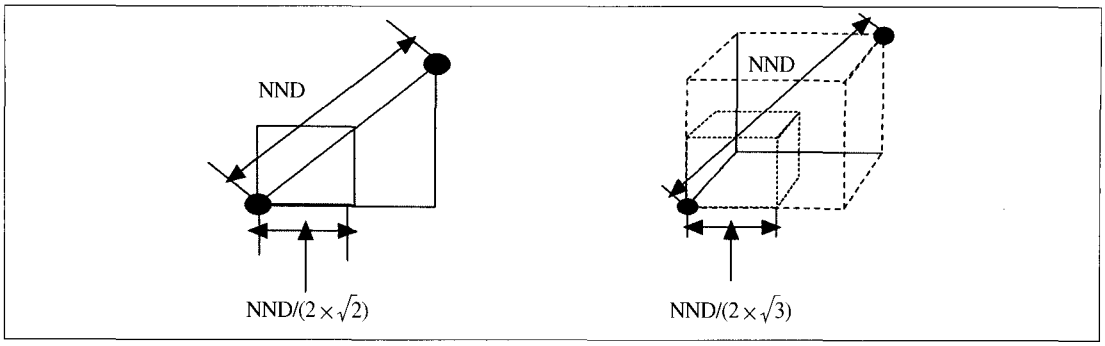
Fig. 3. Clustering procedure using hashing

(2) Application of hashing for multi-dimensional data

In this case, the size of bins should also be carefully chosen. For instance, consider 2- and 3-dimensional cases. Fig. 4. shows the maximal bin sizes for 2D and 3D data.

Thus, for n-dimensional data, the bin size should be

$$(\min.\ NND/2) \times (\frac{1}{\sqrt{n}}). \qquad (3)$$



(a) 2D data: (mim. NND) /(2×√2) (b) 3D data: (min. NND) / (2×√3)

Fig. 4. Maximal bin sizes

2) Speed-up Technique for Clustering by Hashing

The following pseudo-code shows the algorithm for reducing distance computation in clustering using hashing.

```
{ Make hash table }
  iclass = 0
  for each pixel(i, j) {
    if (the bucket indicated by hash(pixel(i,j))
    contains a cluster center)
    then iclass = hash(pixel(i,j)) and go to
    next pixel iteration
    sum = 0.0;
    for each band(k) {
        sum = sum + (intensity[i, j, k] -
        mean[iclass, k])^2
        if (sum is greater than NND[iclass])
     then go to label a save iclass as cluster of
     current pixel exit current cycle of 'for each
     pixel loop'
  label a :
    for class l = 0 to c-1 {
    if (l is equal to iclass) then skip current loop
    sum = 0.0;
    for each band(m) {
      sum = sum + (intensity[i, j, m] - mean[l, m])^2
      if sum is greater than NND[l] then exit
      current loop
    } // end of for-loop for 'band'
    iclass = l;
    exit current cycle of 'for each pixel loop'
  } // end of for-loop for 'class'
} // end of for-loop for 'pixel'
```

4. Experiments and Results

1) Experiments

Seven multispectral images obtained from Landsat-TM(Thematic Mapper) optical sensor on board of Landsat 4, 5 satellites are used as input data. The sizes of images are 512 x 512, 1024 x 1024 and 2048 x 2048. Fig. 5. shows 7 bands of sample images which we used for experiments.

Techniques used for experiment are various combinations of speed-up techniques which are partial sum, NND, auto-correlation, and Hashing, as shown in Table 1.

For each method, the processing times are compared with the variation of variables as shown in Table 2. Three experiments are carried out. In experiment 1, while dimension varies,

Table 1. Experimented techniques

Experimented techniques	Speed-up techniques			
	Partial sum	NND	Auto-correlation	Hashing
method 1	×	×	×	×
method 2	○	×	×	×
method 3	○	○	×	×
method 4	○	×	○	×
method 5	○	○	○	×
method 6	○	○	○	○

○ : used technique × : unused technique

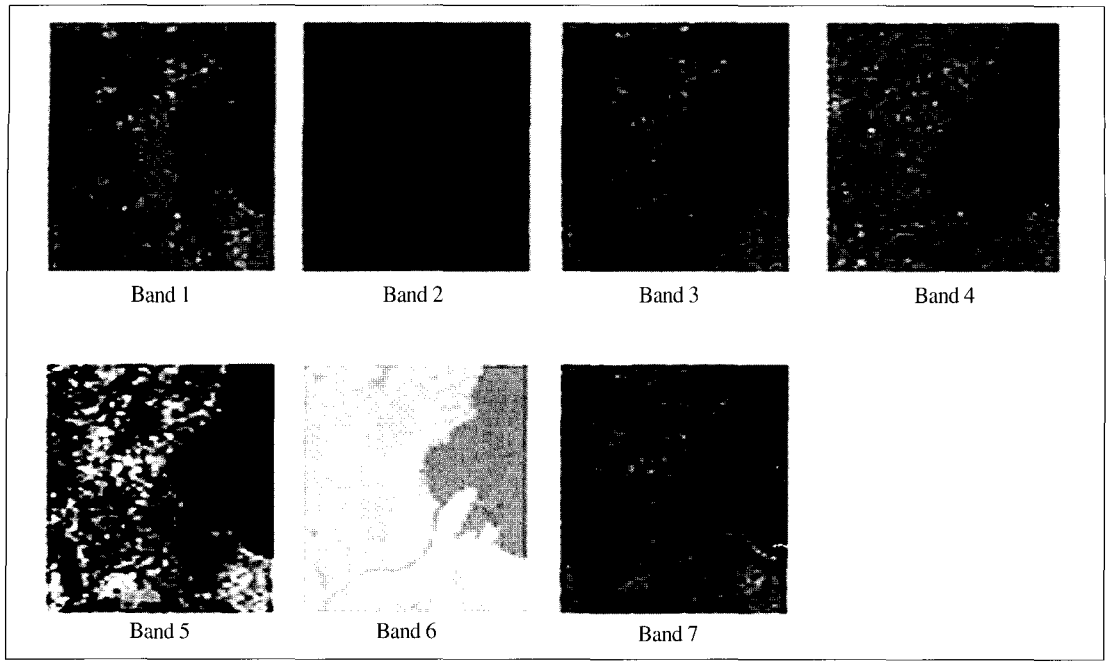


Fig. 5. Landsat TM 7 bands images (512x512)

Table 2. Experiments and Variables used in Experiments

Experiments	Variables										Results given in
	Dimension			Number of clusters			Data size				
Experiment 1	4	6	8	Fixed[2)]			Fixed[3)]				Figure 6, 7
Experiment 2	Fixed[1)]			5	10	15	Fixed[3)]				Figure 8, 9
Experiment 3	Fixed[1)]			Fixed[2)]			256 ×256	512 ×512	1024 ×1024	2048 ×2048	Figure 10, 11

[1)] 4 bands [2)] 5 Clusters [3)] 512 ×512 Pixels

other variables are fixed. In experiment 2, while number of clusters varies, other variables are fixed. In experiment 3, while size of data varies, other variables are fixed.

The experiments are performed on Hyundai Axil-245 Unix computer, and the processing times are measured using CPU clocks.

2) Comparison on time when applied to ISODATA

As the second way of evaluating the effect of proposed technique, we included it in ISODATA and compared with other techniques in the following ways:

(a) Conv: conventional method including none of speed-up techniques.
(b) NND: NND techinique included.
(c) Hashing: proposed techinque of hashing included.

Data used in the experiments are (1) images of band 3 and 4 of size 1024x1024 and (2) image of band 7 of size 512×512.

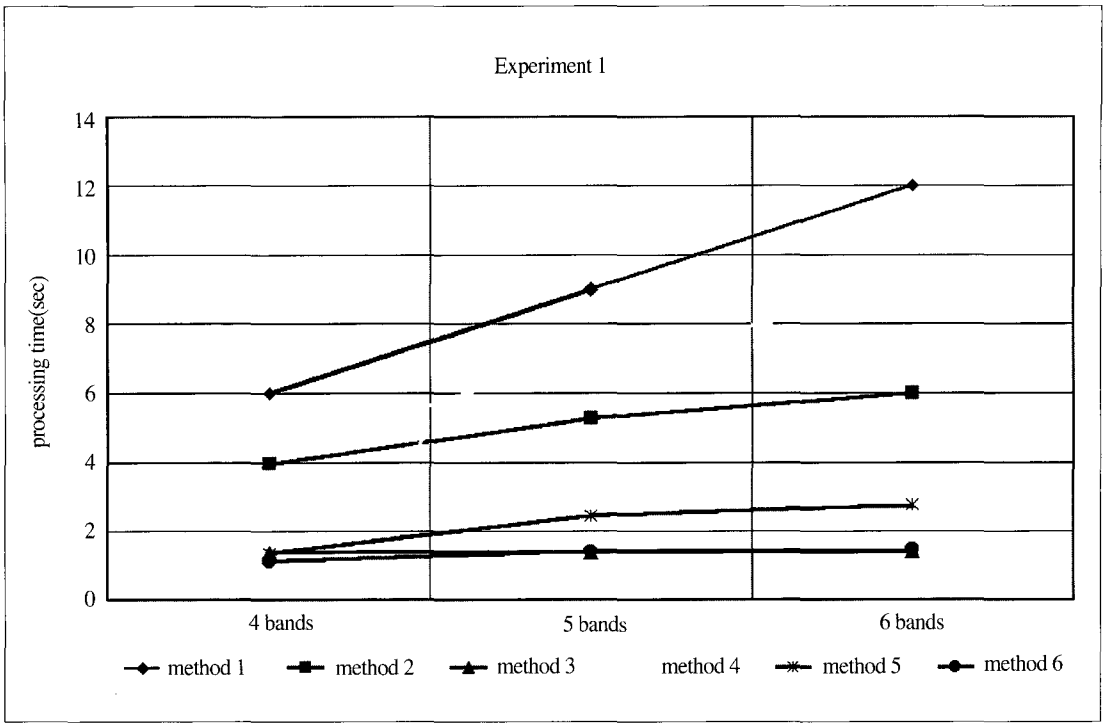


Fig. 6. Comparison with respect to dimension

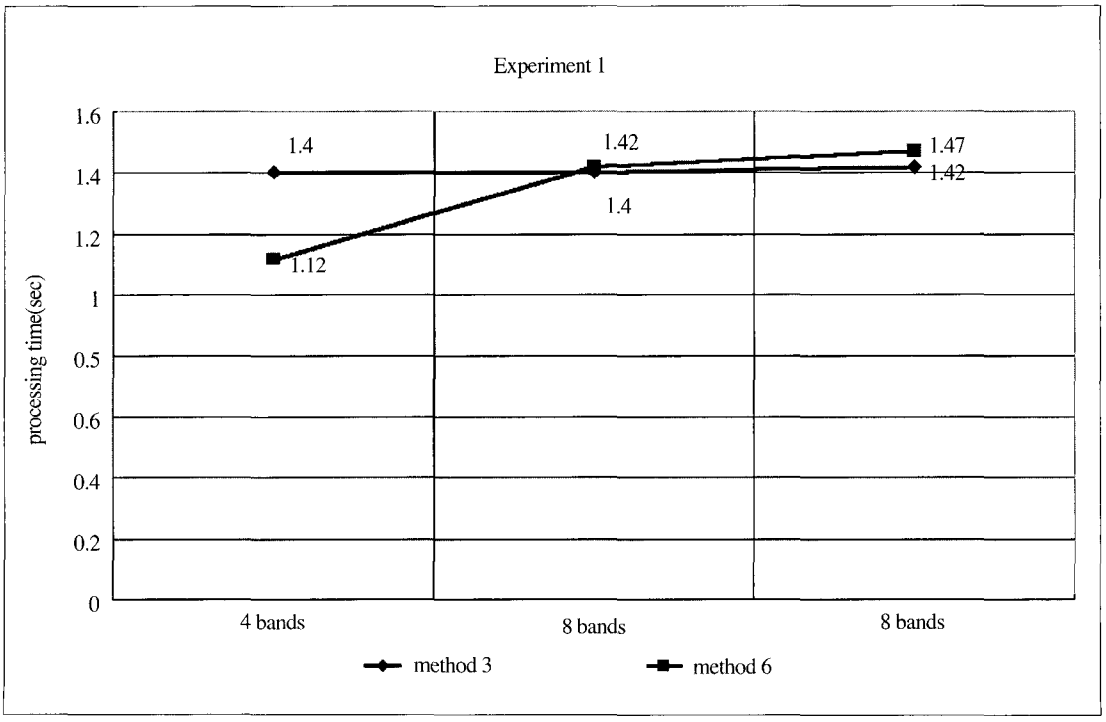


Fig. 7. Detail of Fig. 6 showing method 3 and 6.

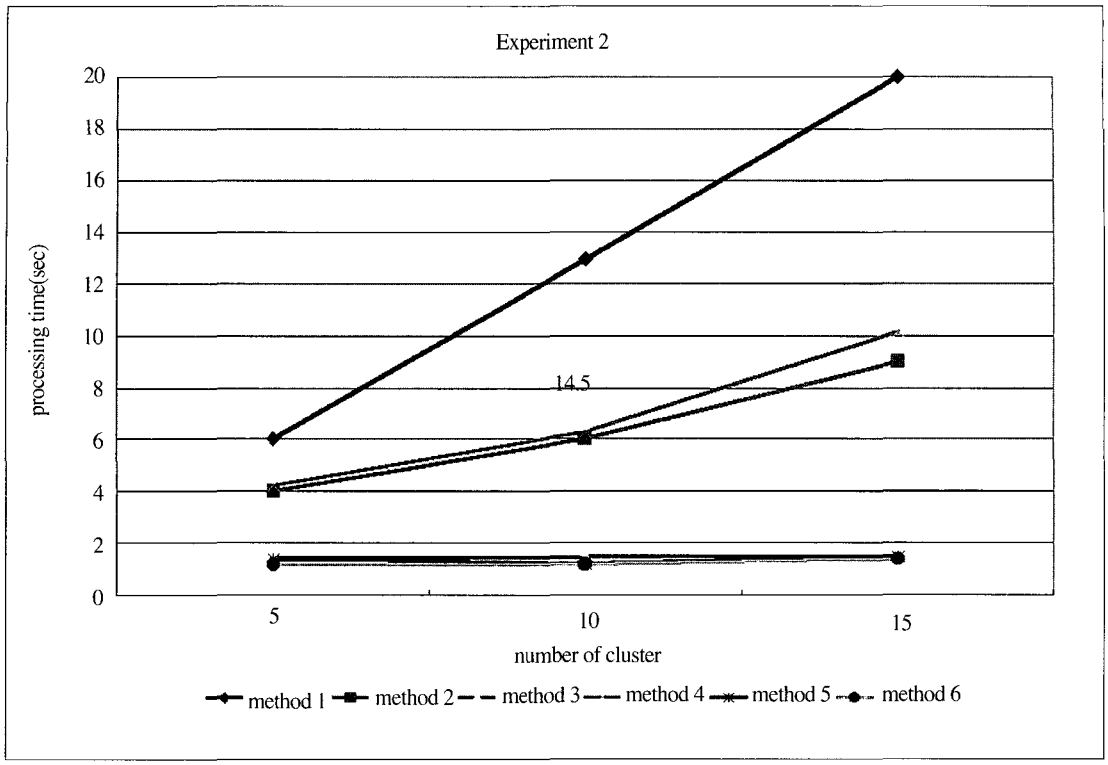


Fig. 8. Comparison with respect to the number of clusters

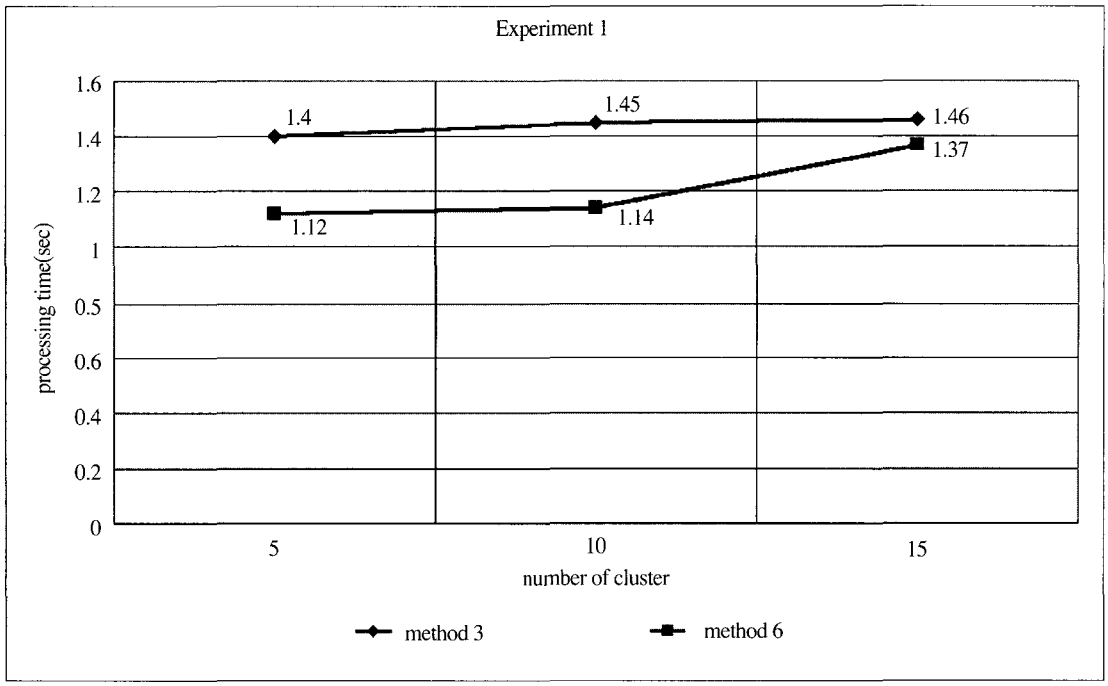


Fig. 9. Detail of Fig. 8 showing method 3 and 6

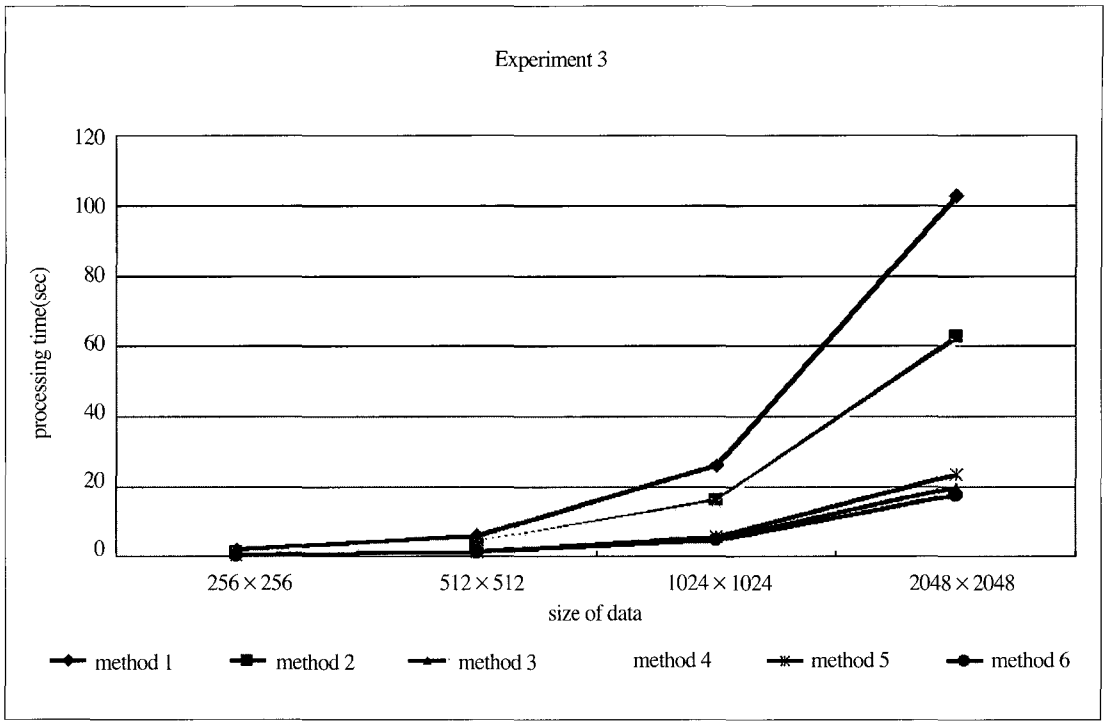


Fig. 10. Comparison with respect to the size of data

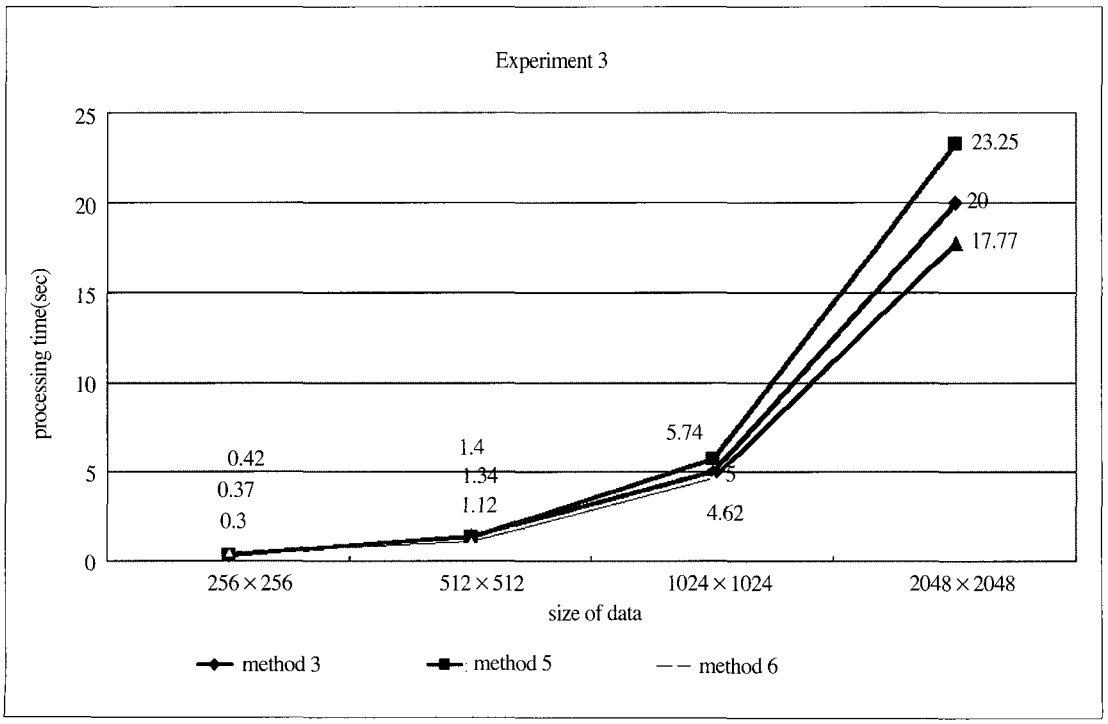
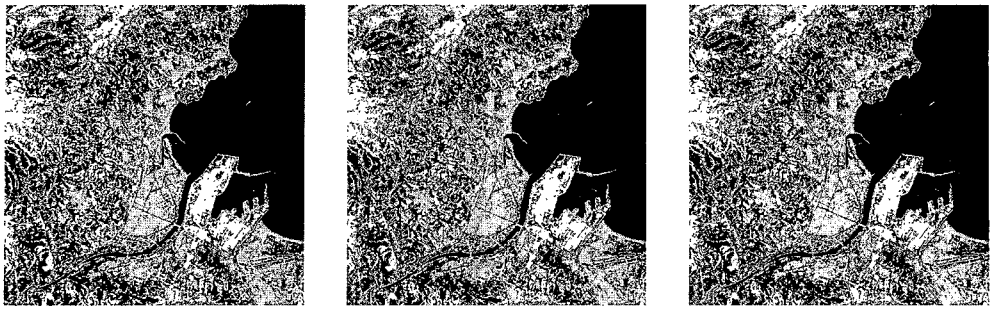


Fig. 11. Detail of Fig. 10 showing method 3,5,6



(a) conventional method (b) NND method (c) Hashing method

Fig. 13. Classification result of each method.

The classification result are shown in Fig. 13. As shown in Fig. 13, the results are the results are almost identical regardless of which technique is used. Only processing time is affected.

The processing times shown in Fig. 14 and Fig. 15 is obtained by averaging 50 runs.

The horizontal axis represents iteration number, while the vertical axis represents cumulative time in Intel Pentium 200 MHz processor with 128MB RAM.

3) Analysis of Results

The results presented in the previous section shows that we could find out that the proposed hashing technique improves performance almost equally with partial sum and NND techniques when the number of dimension or the number of clusters are varied.

However, it performs better than partial sum, NND and auto-correlation techniques when the

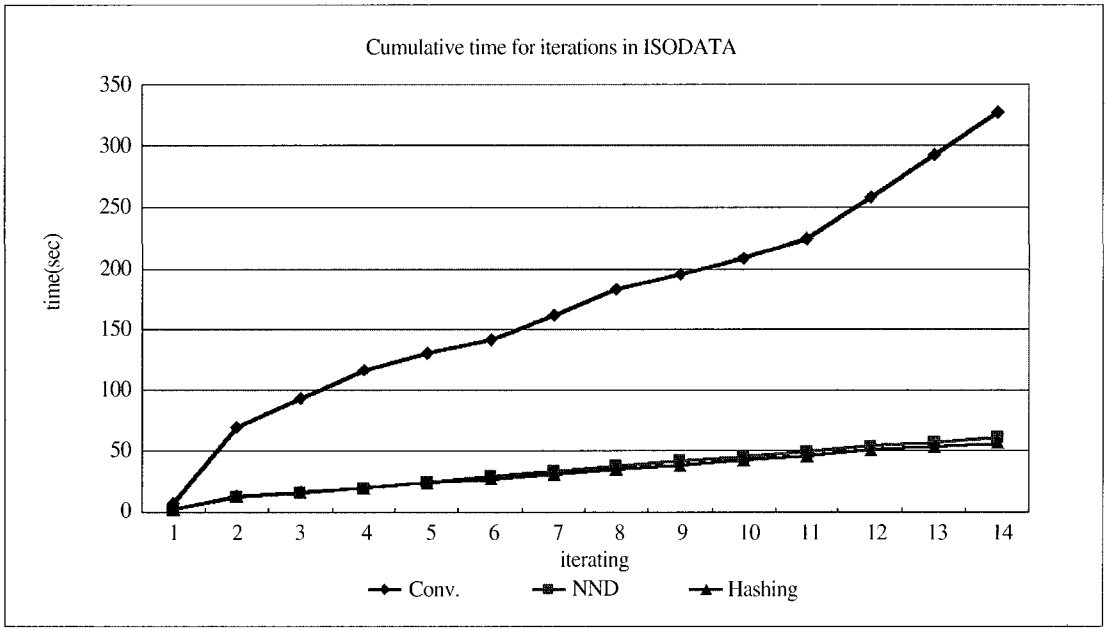


Fig. 14. Cumulative time for iterations in ISODATA (band 3,4 of 1024×1024)

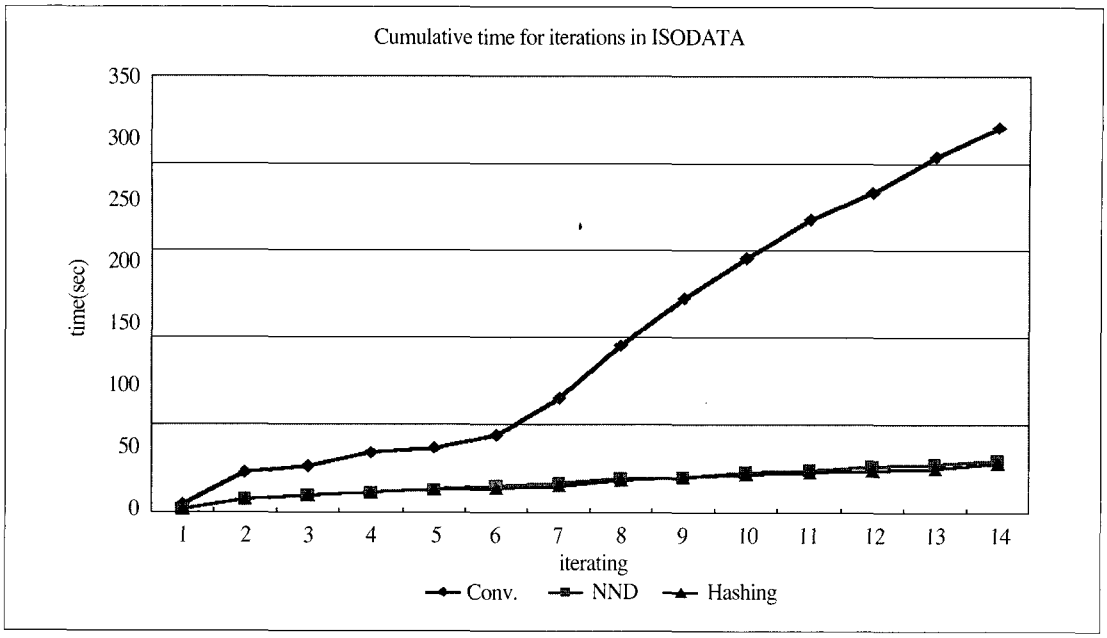


Fig. 15. Cumulative time for iterations in ISODATA (band 7 of 512×512)

data size is varied. This is meaningful result, as the data size is usually quite big for satellite images.

The result in section 4.2 shows that the proposed technique requires more time than NND in early iterations, but it requires less time in later iterations. The reason owe to the facts that the frequency for a pixel to find its cluster at once increases as many pixels are concentrated around the cluster centers in later iterations.

5. Conclusion and Future Research

This paper proposes a new technique based on hashing to reduce the time in clustering for satellite images. The technique avoids computing the distance to all cluster centers: Instead of computing the distance to all cluster centers, hash function is computed once.

The experiment results show that the proposed technique performs better than other techniques when the data size is large. The technique becomes more effective as iterations go on.

As future research, hashing technique used for clustering can be directly applied to the classifier to improve the whole classification process.

Acknowledgments

The authors wish to acknowledge the financial support of the Korea Research Foundation made in the Program Year 1999. They also express their gratitude to Mr. Sang-Ik Lee for Ministry of Environment of Korea to provide Landsat TM satellite imagery for the purpose of research.

References

Hodgson, M. E., 1988, Reducing the Computational

Requirements of the Minimum-distance Classifier, *Photogrammetric Engineering & Remote Sensing*, 55(5): 613-619.

Jack Bryant, 1989, A fast classifier for image data, *Pattern Recognition*, 22(1): pp. 45-48.

Jensen, J. R. 1996, *Introductory Digital Image Processing A Remote Sensing Perspective*, Prentice Hall.

Tou, J. T., Gonzalez, R. C., 1974, *Pattern Recognition Principles*, Addison-Wesley Publishing Company.

Venkateswarlu, N. B. and Raju, P. S. V. S., 1992, Fast ISODATA clustering algorithms, *Pattern Recognition*, (25)3: pp. 335-343