Korean Journal of Geomatics
Vol. 5, No. 1 / December 2005
pp. 27~34

A Study on the Unsupervised Classification of Hyperion and ETM+ Data Using Spectral Angle and Unit Vector

Dae Sung KIM*, Yong Il KIM and Ki Yun YU*****

Abstract

Unsupervised classification is an important area of research in image processing because supervised classification has the disadvantages such as long task-training time and high cost and low objectivity in training information. This paper focuses on unsupervised classification, which can extract ground object information with the minimum "Spectral Angle Distance" operation on be behalf of "Spectral Euclidian Distance" in the clustering process. Unlike previous studies, our algorithm uses the unit vector, not the spectral distance, to compute the cluster mean, and the Single-Pass algorithm automatically determines the seed points. Atmospheric correction for more accurate results was adapted on the Hyperion data and the results were analyzed. We applied the algorithm to the Hyperion and ETM+ data and compared the results with K-Means and the former USAM algorithm. From the result, USAM classified the water and dark forest area well and gave more accurate results than K-Means, so we believe that the "Spectral Angle" can be one of the most accurate classifiers of not only multispectral images but hyperspectral images. And also the unit vector can be an efficient technique for characterizing the Remote Sensing data.

Keywords : *Hyperspectral Image, USAM, Single-Pass Algorithm, Unit Vector, Hyperion Data, FLAASH*

1. Introduction

Hyperspectral Remote Sensing (also known as Imaging Spectrometry, Imaging Spectroscopy or Hyperspectral Imaging) is defined as "the acquisition of images in hundreds of registered, contiguous spectral bands such that for each pixel of an image it is possible to derive a complete reflectance spectrum (Freek, 2001)". Recently, the number of researches related to Hyperspectral image classification has increased, but because of too many bands and unstable training information, the accuracy of hyperspectral data classification is lower than that of classical classification, and thus does not give confident results. As a measure, Dimensionality Reduction methods like Feature Extraction or Feature Selection have been used (Anil, 2003), and new algorithms for Hyperspectral data classification have been proposed.

The fundamental premise of the remote sensing of land cover/land use is that every surface object has its own unique pattern of reflected, emitted, and absorbed radiation across the spectral band and the same types of surface objects show similar spectral response patterns (James, 1996). SAM (Spectral Angle Mapper) is a new algorithm based on the fact that the spectra of the same type of surface objects in RS data are approximately linearly scaled variations of one another due to atmospheric and topographic effects (Youngsinn, 2002). In this paper, we applied the spectral angle technique to unsupervised classification based on previous research and obtained a more efficient algorithm by using the mean of the angle based on the unit vector (also called angle mean) in the calculation of a cluster center. To obtain more accurate results, we applied atmospheric correction (FLAASH) and automatic seed point selection (Single-Pass algorithm).

The following three hypotheses are tested in this study;

*Ph. D. Candidate, School of Civil, Urban, and Geosystem Eng., Seoul National University (E-mail : mutul94@empal.com)
**Associate Professor, School of Civil, Urban, and Geosystem Eng., Seoul National University (E-mail : yik@snu.ac.kr)
***Assistant Professor, School of Civil, Urban, and Geosystem Eng., Seoul National University (E-mail : kiyun@snu.ac.kr)

The manuscript for this paper was submitted for review on May 16, 2005 / accepted on September 30, 2005

Hypothesis 1: SAM algorithm needs to apply the atmospheric correction to give more accurate results.

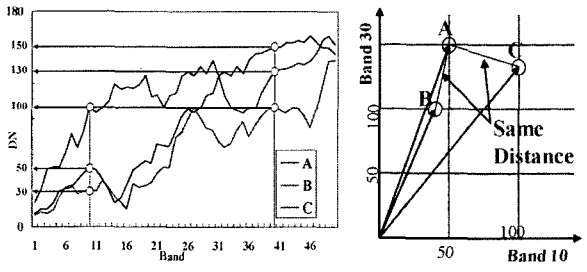
Hypothesis 2: Processing time and accuracy of algorithm vary according to the method of determining seed points.

Hypothesis 3: SAM algorithm using the spectral angle is more accurate than K-Means with spectral distance and angle is more efficient than the distance in the calculation of cluster mean.

This paper is organized as follows. Section 2 describes the fundamental principle of the proposed approach. Section 3 presents our method for unsupervised SAM. Section 4 tests the proposed hypotheses. Finally, Section 5 provides some conclusions and comments on future research.

2. Unsupervised Spectral Angle Mapper

The SAM classifies the image pixels based on the minimum "Angular Distance" rule and does not require the training data to be normally distributed, and it is insensitive to the data variance and the size of the training data set (Kruse *et al*, 1993). When the unsupervised SAM is used, image pixels that have similar shape patterns will be classified together into the same cluster. For example, consider the three spectra in Figure 1-(a). The spectral profile A is the spectrum of the seed point; B and C are the unknown spectrums of pixel in the data. In the 10^{th} band, the Digital Numbers of A, B, and C have 50, 30, and 100 and in the 40^{th} band, respectively, and A, B, and C are 150, 100, and 130, respectively. Figure 1-(a) is transformed to a 2D feature space like Figure 1-(b). The Euclidian distance between A and B and the distance between A and C are the same, so the cluster can be decided only from the Angle distance. A is more similar to B than C.



(a) Spectral Profile (b) 2D Plot

Fig. 1. Concept of Spectral Angle Mapper.

The USAM performs unsupervised classification based on the minimum spectral angular distance rule using the iterative data analysis technique. Its processing is as follows;

1. Select the seed points
2. Assessment the similarity through the angle distance between seed points and unknown pixels
3. Calculate each cluster mean
4. Repeat steps 2 and 3 until the iteration reaches the user specified threshold

The processing about USAM is described in the following sub-section.

2.1 Selection of Seed Points

First, the seed points must be selected. It is an N-dimensional mean data vector with N being the number of bands used in the clustering. In general, the pixels in an image are pointed to seed vectors (1) at random, (2) by equal division of the range of DN each band, (3) through single-pass algorithm and (4) hierarchical clustering (Ward *et al*, 1963) and so on. In this paper, the single-pass technique and equal division of the range of DN each band were applied and compared the results.

Single-pass algorithm is one of the fast clustering techniques, which require only one pass through the data. It uses only sample pixels, not all the pixels of image, and the samples are arranged into a two dimensional array. The first row of the samples is then used to obtain the starting set of cluster centers, and the user specified critical distance is used to form another cluster center (Richard, 1999).

2.2 Assessment of Similarity

Contrary to other conventional clustering algorithms, SAM computes the similarity of an unknown spectrum to a reference spectrum by using the spectral angle rule. The spectral angle $\theta_{i,c}$ between every pixel i in the image and every cluster mean μ_c is found by using equation (1).

$$\theta_{i,c} = \cos^{-1}\left[\frac{\sum_{k=1}^{m} x_{i,k}\,\mu_{c,k}}{\sqrt{\sum_{k=1}^{m} x_{i,k}^2 \sum_{k=1}^{m} \mu_{c,k}^2}}\right] \quad (1)$$

$$If \quad \theta_{i,cmin} = Min(\theta_{i,c}),\ then\ x_{i,c} \rightarrow x_{i,k,c} \quad (2)$$

Small values of θ indicate that the two spectra are quantitatively similar (Arel *et al*, 1999, see the

equation (2)), where, m is the number of band, and $x_{i,k}$ is one of the pixels that were randomly selected.

2.3 Determination of Cluster Means

After seeking the similar elements of the clusters, the means of elements are generally recalculated by using "Distance Mean". But, we produced a more efficient algorithm at the spectral angle hypothesis with "Unit Vector" (or Angle Mean) in this study. The related equation is as follows,

$$\theta_{i,k} = \cos^{-1}\left(\frac{x_{i,k}}{\sqrt{\sum_{k=1}^{m} x_{i,k}^2}}\right) \quad (3)$$

$$\mu_{\theta_{i,k}} = \frac{1}{N_c}\sum_{i=1}^{N_c}\theta_{i,k} \quad (4)$$

$$\mu_{c,k} = unitvector \times \cos(\mu_{c,k}) \quad (5)$$

$$unitvector = (p_{i,1}^2 + \cdots + p_{i,m}^2) = \sqrt{\sum_{k=1}^{m} p_{i,k}^2} = 1 \quad (6)$$

The angle $\theta_{i,k}$ between seed points and each band is above equation (3), and the mean angle $\mu_{\theta_{i,k}}$ (equation (4)) is calculated by dividing the sum of each angle. Here, N_c is the number of elements of any cluster. Using the mean angle $\mu_{c,k}$, the unit vector which the length of vector is one (called as unit) is used for determining the cluster mean (equation (5)). In consequence, the spectral angle is calculated again through the equation (1).

As shown in Figure 2, if unknown pixels 1 and 2 are included in the same cluster, the former USAM algorithm calculates the cluster mean with distance at each band but it is not suitable for some cases like pixel existing near 0. We believe that the use of the unit vector is a more efficient technique for remotely sensed data because RS data generally exist in the middle of the space plane, and the unit vector is not sensitive to the outlier.

Finally, the above steps are repeated until the iterations reach the user specified convergence threshold. In this study, 10 seed points for clustering were selected, and the iterative calculations just stopped when the number of elements changed by less than 1%.

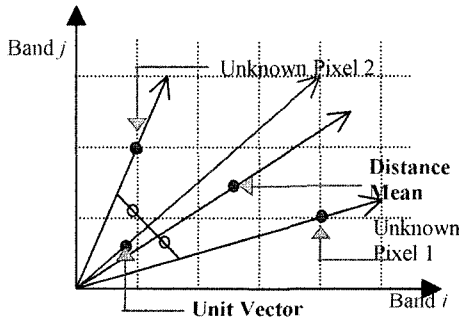


Fig. 2. Comparison of Mean Pixel and Unit Vector.

3. Implementation

3.1 Data

The USAM algorithm was applied to two different sets of data, that is, the Hyperion data, which are loaded on the EO-1 satellite and ETM+ on Landsat-7. The Earth Observing (EO-1) Satellite was launched on November 21, 2000 and contains three observing instruments, Advanced Land Imager (ALI), LEISA Atmospheric Corrector (LAC), and Hyperion Imaging Spectrometer, supported by a variety of newly developed space technologies. The Hyperion provides a high spectral resolution hyperspectral imager capable of resolving 242 spectral bands (from 0.4 to 2.5 μm) with a 30-meter resolution (same as ETM+ spatial resolution). The specification of data used for testing is arranged in Table 1. Each image was sensed at the same day, April 3, 2002, and test area was situated in the southern part of Seoul, Korea (Figure 3).

3.2 Preprocessing

Band selection, atmospheric correction, and co-registration techniques were pre-processed before

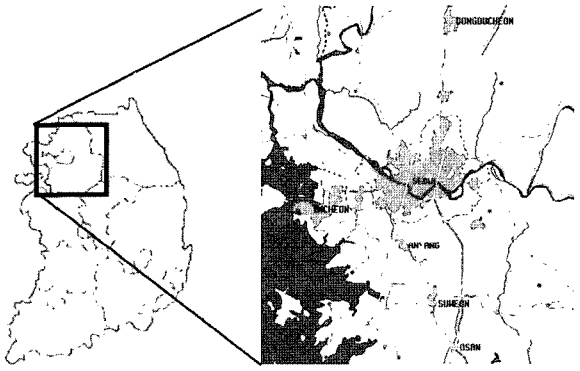


Fig. 3. Study Area.

Table 1. Specifications of Data.

Sensors	Swath Width	Spatial Resolution	Channels	Spectral Range
Hyperion	7.6 km	30 m	220	0.40~2.5 ㎛
ETM+	185 km	30 m	8	0.45~12.5 ㎛

applying the algorithm to the Hyperion data and ETM+.

Ninety-three bands having the same band width as the ETM+ wavelength were selected, and each band corresponded to the following: band 1 (ETM+): band 10-17 (Hyperion), band 2: 18-26, band 3: 27-34, band 4: 39-55, band 5: 140-161, and band 7: 192-220. Figure 4 explains the selected Hyperion band and the simple preprocessing for USAM is arranged in Figure 5.

FLAASH (for demonstration of hypotheses I) which used as atmospheric correction was developed by air by Spectral Science, Inc. under the sponsorship of the U. S. Air Force Research Lab as the first-principles atmospheric correction modeling tool for retrieving spectral reflectance from hyperspectral radiance images (Berk *et al*, 1989). Unlike many other atmospheric correction techniques, FLAASH incorporates the MODTRAN4 (Matthew *et al*, 2000) radiation transfer code and its basic concept is based on equation (7) below.

$$L=\left(\frac{A\rho}{1-\rho_e S}\right)+\left(\frac{B\rho}{1-\rho_e S}\right)+L_a \qquad (7)$$

where ρ is the pixel surface reflectance, ρ_e is an average surface reflectance for the pixel and a surrounding region, S is the spherical albedo of the atmosphere, L_a is the radiance back scattered by the atmosphere, and A and B are the coefficients that depend on atmospheric and geometric conditions but not on the surface.

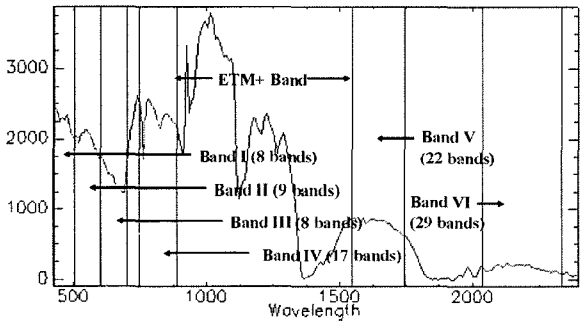


Fig. 4. Selected Hyperion Band (same as ETM+).

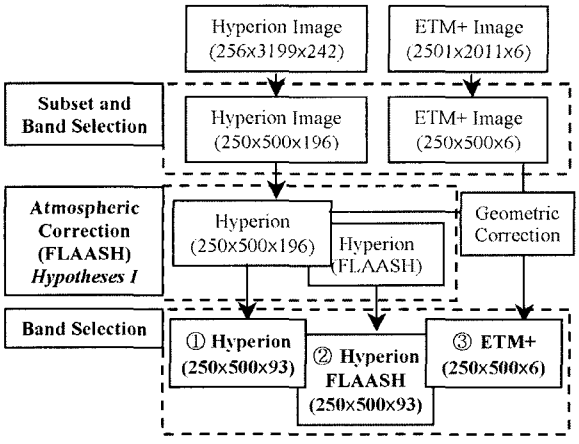


Fig. 5. Flow Chart of Preprocessing.

Because each image must be in same spot for fair terms, image-to-image registration and image subset were applied. ETM+ was registered on the basis of the Hyperion image, and each image was made by re-sampling as the same size (200 pixels by 500 pixels). The false color composite images, which were preprocessed for implementation, are shown in Figure 6.

3.3 Validation

Three data (Hyperion, Hyperion-FLAASH, and ETM+), which were preprocessed, passed the process with selected ten seed points. This processing is needed to demonstrate hypotheses II.

Three data were branched off with six data through the Single-Pass algorithm and the range dividing. On the basis of the six data, two techniques, spectral distance (K-Means) and angle (USAM), were applied to similarity calculation, and the results were comprised of twelve products. Finally, six data out of the twelve results from the USAM were tested for recalculating the cluster centers (for demonstration of hypotheses III). Therefore, finally eighteen result images were produced, and we assessed the result. You can find the indices, ⓐ to ⓡ, in Figure 7. The eighteen results, which were classed through each clustering algorithm, are arranged in the Figure 8 (You can find the results at the end of this paper).

Each case result appears as a series of gray-scale images, one for each cluster; especially, the USAM results generated the angle value file. Angle value file

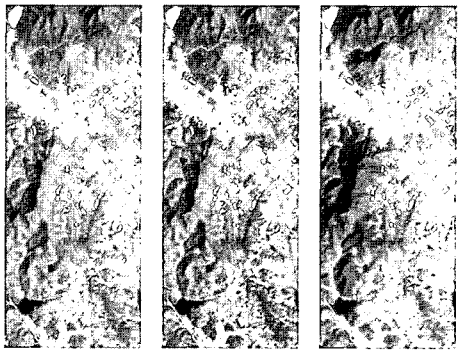


Fig. 6. Preprocessed Data.
[From left to right, Hyperion (R: 30, G: 20, B: 10), Hyperion-FLAASH (30, 20, 10), and ETM+ (4, 3, 2)]

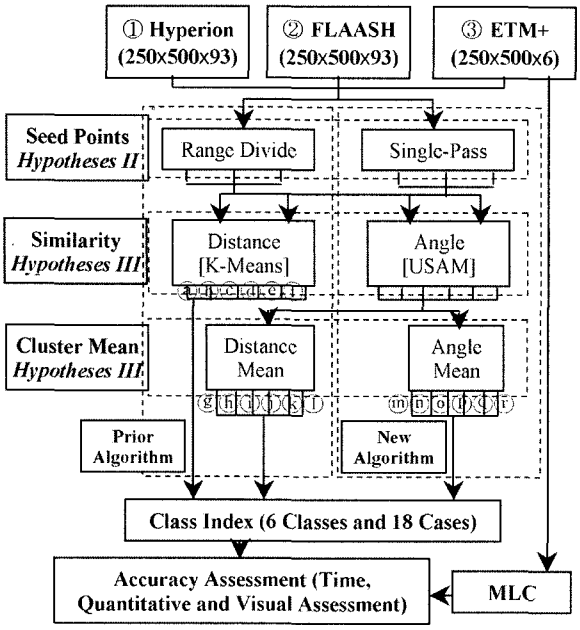
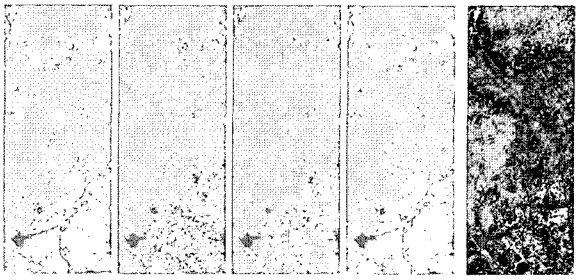


Fig. 7. Flow Chart for Application.

Table 2. The Results of MLC, MDC, ECHO, and SAM.

Methods	MLC	MDC	SAM	ECHO
Accuracy	93.0%	81.8%	80.9%	97.1%
Kappa Co.	91.8%	78.7%	77.6%	96.6%



(a) MLC, MDC, SAM, ECHO from right (b)

Fig. 10. Thematic Map and Sampled Reference Data. (Note: The classes, Water, Forest, Soil, Urban and Grass, are corresponded with the colors, which are Blue, See Green, Yellow, Red, and Green, at (a). The black color in the image means the un-sampled pixels (b).)

represents the angle between a cluster center and an unknown pixel when the unknown pixel is embedded in any cluster. Higher angle is represented by the brighter pixels (larger floating-point numbers) and means that probability (or weight) of inclusion at any cluster is low. Figure 9 shows the value image in each case applied to USAM (at the end of this paper).

In addition, quantitative assessment, time, and stability of astringency were considered for the criteria of assessment. For thematic reference map to assess the clustering results, the MLC (Maximum Likelihood Classifier) and the supervised SAM, MDC (Minimum Distance Classifier), and ECHO (Extraction and Classification of Homogeneous Objects) classification techniques were applied and the pixels included in the same area of each result were selected for organizing the error matrix. Five classes, Water, Forest, Soil, Urban and Grass, were trained as a test set, and the training set were processed with the Multispec program. The classification accuracy and Kappa statistics (x 100) which calculated through test set show in Table 2. The each result of MLC, MDC, ECHO, and SAM is arranged in Figure 10-(a).

For visual approach, the result images were represented as a series of binary image (black and white image) on the basis of ten clusters. Similar clusters were merged from ten to five for quantitative approach and the merged clusters were indexed as the same class as the thematic reference map from supervised classification. Figure 10-(b) is an example of a thematic map composed of five classes. The accuracy and iteration number of each classification result are filled in Table 3 (at the end of this paper). From the above results, we verified 3 hypotheses which were established in the introduction. The demonstration of 3 hypotheses is described in the following section.

4. Demonstration of Hypotheses

4.1 Hypothesis I

Hypothesis I proves that the atmospheric correction method needs to be applied to the SAM algorithm to obtain a better product. The atmospheric corrected Hyperion data (ⓑ, ⓔ, ⓗ, ⓚ, ⓝ, and ⓠ in Fig. 8) and not (ⓐ, ⓓ, ⓖ, ⓙ, ⓜ, and ⓟ) were compared to demonstrate hypotheses I.

The results can not be found for any other difference through visual approach and also the classification accuracy of Hyperion-FLAASH data shows a similar or somewhat lower result in comparison with Hyperion. Time and stability of astringency are much alike in the iteration processing. Therefore, we can conclude that hypotheses I is wrong. You can find the graphs about classification accuracies, iteration number, and astringency of Hyperion and Hyperion-FLAASH in Figure 11.

4.2 Hypothesis II

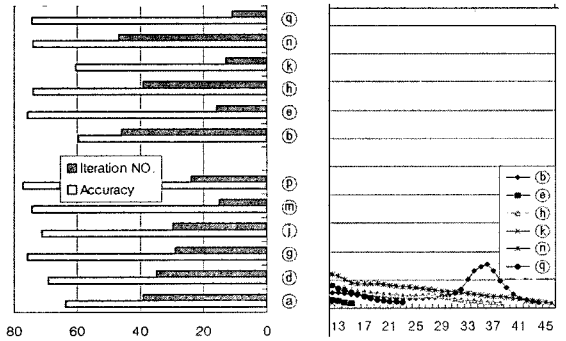
The fact that processing time and accuracy of algorithm give different results according to determination of seed points is tested by hypotheses II. The

Single-Pass algorithm (in cases of ⓓ, ⓔ, ⓕ, ⓙ, ⓚ, ⓛ, ⓟ, ⓠ, and ⓡ) and the division of the range method (ⓐ, ⓑ, ⓒ, ⓖ, ⓗ, ⓘ, ⓜ, ⓝ, and ⓞ) are compared with iteration number. The Single-Pass has faster processing time and slightly higher classification accuracy than the division of the range method does. In this case, we can find that good seed points have an effect on the accuracy and processing time. Therefore, determination of the seed point will be researched in the future. Figure 12 shows the graphs about the classification accuracies, iteration number, and astringency of the Single-Pass algorithm and the dividing range method.

4.3 Hypothesis III

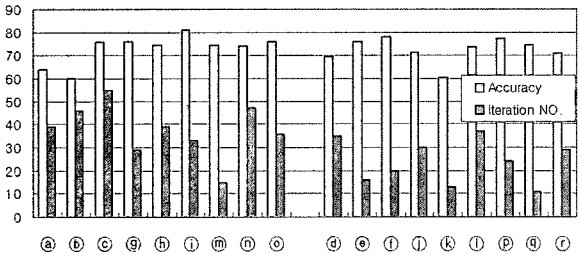
The SAM algorithm using the spectral angle (ⓖ, ⓗ, ⓘ, ⓙ, ⓚ, and ⓛ) and distance (ⓜ, ⓝ, ⓞ, ⓟ, ⓠ, and ⓡ) is more accurate than K-Means (ⓐ, ⓑ, ⓒ, ⓓ, ⓔ, and ⓕ) and can be one of the most efficient classifier of not only multispectral images but hyperspectral images in this sub-section. This hypothesis was already demonstrated by Youngsinn *et al* (Youngsinn, 2002), and we also confirmed this result by visual and quantitative tests. In the visual estimation, the result of K-Means clustering can not distinguish the dark forest from water but USAM can (see the Figure 13).

We can also find the particular result image in case of ⓐ, that is, the five clusters only appeared, and the rest of the cluster was not classified. This result is due to the fact that K-Means is sensitive to the outlier. Figure 14 shows the particular result with a binary image and 6^{th}-10^{th} images look black because of one or two cluster elements. The accuracy of USAM also is higher than the results of K-Means. In this perspective, the spectral angle is an effective method for the hyperspectral data in addition to multispectral data. Figure 15 presents the comparison graph about hypothesis III.

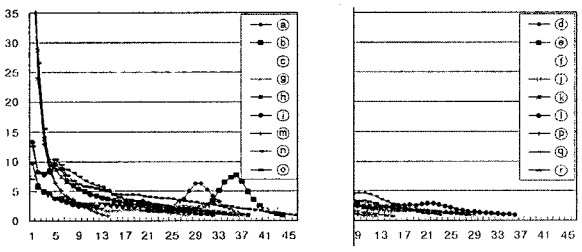


(a) Accuracy and Iteration Number (b) Astringency

Fig. 11. Comparison of Hypothesis I.

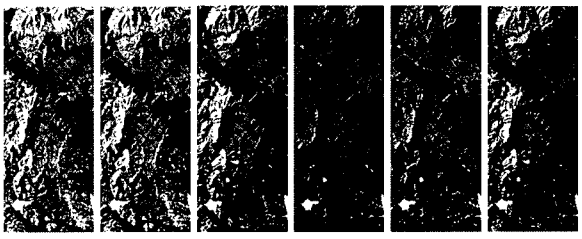


(a) Accuracy and Iteration Number

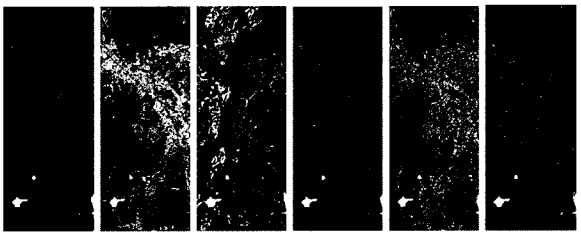


(b) Astringency

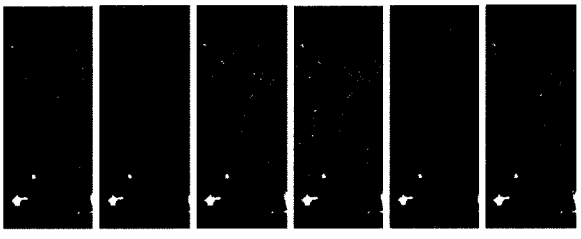
Fig. 12. Comparison of Hypothesis II.



(a) K-Means (from left, ⓐ, ⓑ, ⓒ, ⓓ, ⓔ, and ⓕ)



(b) USAM Distance (from left, ⓖ, ⓗ, ⓘ, ⓙ, ⓚ, and ⓛ)



(c) USAM Angle (from left, ⓜ, ⓝ, ⓞ, ⓟ, ⓠ, and ⓡ)

Fig. 13. Critical Difference of Water Class in Each Case (Note: The white color means water class).

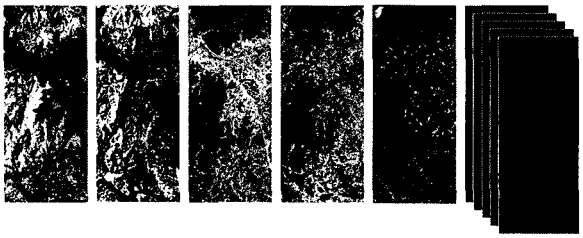


Fig. 14. The Particular Result Image in Case of ⓐ.

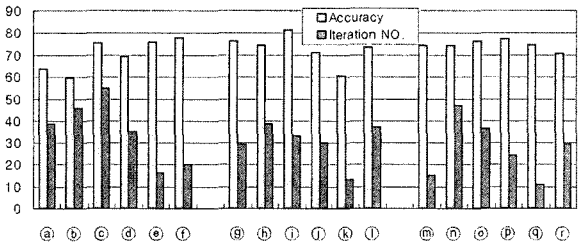


Fig. 15. Comparison of Hypothesis III.

5. Conclusion

In this paper, we examined one of the classification algorithms, Unsupervised SAM, and also discussed the test performed on the Hyperion and ETM+ data. We applied the Single-Pass algorithm for selecting the seed points. Atmospheric correction, FLAASH, was used for the pre-processing of the Hyperion data. Contrary to the previous SAM algorithm, the angle mean for the cluster center was used in this study and compared with results of other algorithms. Using the eighteen cases which were branched off from three data, 3 hypotheses were demonstrated through the test like classification accuracy, processing time, astringency of elements change and so on.

A few essential conclusions from the result may be stated more simply:

1. Atmospheric correction has few influences on the result of the clustering algorithms.

2. The selection of good seed points can reduces the cost of iterative computation.

3. The spectral angle can be one of the most accurate classifiers and a valuable tool for clustering.

In this study, we detected some limitations of the USAM model. First, the accuracy of the classification result depends on the selected seed points. Second, SAM is insensitive to near dark point because it uses only the direction of the spectra not the length.

Based on the understanding of the limitation of the USAM algorithm, we plan to conduct further studies on the following topics:

1. The calculation of the cluster mean by applying the split and merge technique.

2. Study on the combination of the angle and distance concept as a solution of the above second limitation.

3. The research on the SAM value file.

Acknowledgements

The authors thank the Research Institute of Engineering Science, Seoul National University for their support and the ETRI (Electronics and Telecommunications Research Institute) for providing the Hyperion for this study.

References

1. A. Berk, L. S. Bernstein and D. C. Robertson (1989), MODTRAN: a Moderate Resolution Model for LOWTRAN7, MA, Air Force Geophys. Lab., Hanscom AFB, pp. 38.
2. Anil Cheriyadat (2003), Limitations of Principal Component Analysis for Dimensionality-Reduction for Classification of Hyperspectral Data, MA, Mississippi State University, pp. 8-24.
3. Arel Weisberg, Michelle Najarian, Brett Borowski, Jim Lisowski, Bill Miller (1999), Spectral Angle Automatic cLuster rouTine (SAALT): An Unsupervised Multispectral Clustering Algorithm, Aerospace Conference, IEEE, Vol. 4, pp. 307-317.
4. ENVI Tutorials (2002), Research Systems.
5. Freek D. Van Der Meer, and Steven M. De Jong (2001), Imaging Spectrometry - Basic Principles and Prospective Applications, Kluwer Academic Publishers, Netherlands.
6. James B. Campbell (1996), Introduction to Remote Sensing, Second Edition, The Guilford Press, New York, pp. 5.
7. John R. Jensen (2005), Introductory Digital Image Processing-A Remote Sensing Perspective, Third Edition, Prectice Hall, the United States of America, pp. 450-453.
8. Kai-Yi Huang (2002), A Synergistic Automatic Clustering Technique (SYNERACT) for Multispectral Image Analysis, Photogrammetric Engineering & Remote Sensing, ASPRS, Vol. 68, No. 1, pp. 33-40.
9. Kruse FA, Lefkof AB, Boardman JB, Heidebrecht KB, Shapiro AT, Barloon PJ, Goetz AFH (1993b), "The Spectral Image Processing System (SPIS)-Interactive Visualization and Analysis of Imaging Spectrometer Data", Remote Sensing of the Environment, No. 44, pp. 309-336.
10. Mazer AS, Martin M, Lee M, Solomon JE (1988), "Image Processing Software for Imaging Spectrometry Data Analysis", Remote Sensing of the Environment, No. 24, pp. 201-210.
11. M. W. Matthew, S. M. Adler-Golden, A. Berk, S. C.

Richtsmeire (2000), Status of Atmospheric Correction Using a MODTRAN4-based Algorithm. SPIE Proceeding, Algorithm for Multispectral, Hyperspectral, and Ultra-spectral Imagery VI, 4049, pp. 199-207.

12. Richard J. (1999), Remote Sensing Digital Image Analysis (An Introduction)-Third Edition, Springer-Verlag, Germany, No. 4, pp. 230-236.
13. Ward, J.H., Jr. (1963), Hierarchical Grouping to Optimize an Objective function, Journal of the American Statistical Association, Vol. 53, No. 301, pp. 236-244.
14. Youngsinn Sohn and N. Sanjay Rebello (2002), Supervised and Unsupervised Spectral Angle Classifiers, PE & RS, ASPRS, Vol. 68, No. 12, pp. 1271-1280.

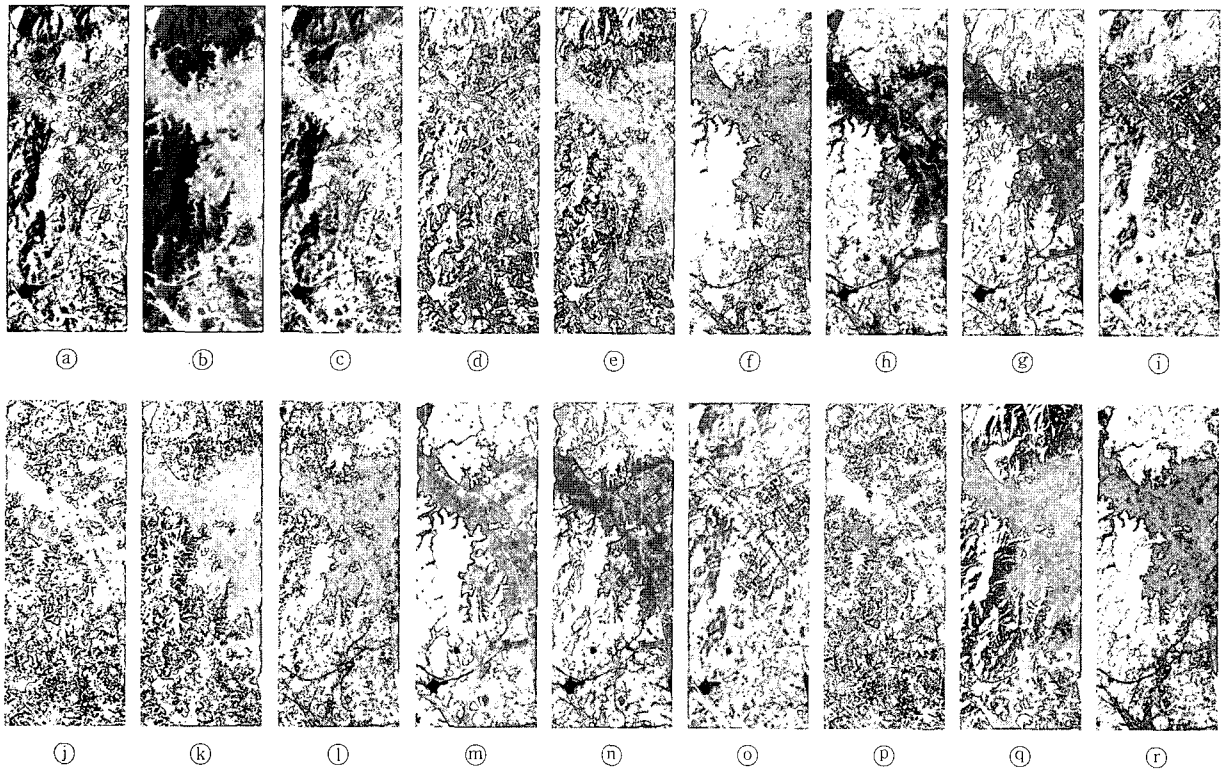


Fig. 8. Eighteen Results which were classed through Each Clustering Algorithm (same as indices in Figure 7).

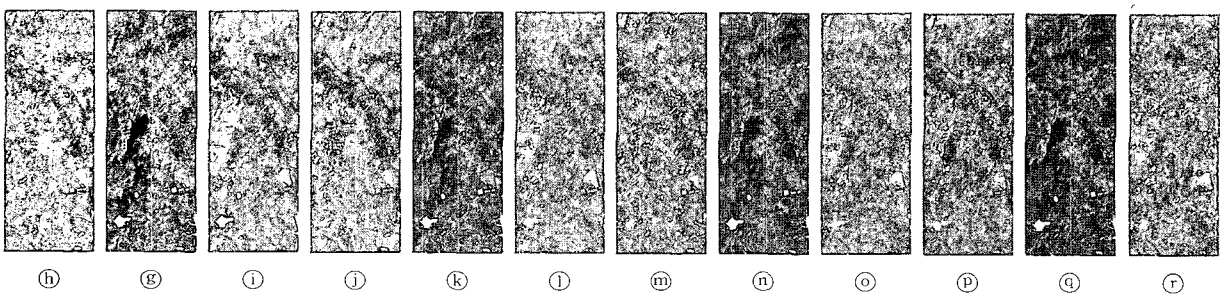


Fig. 9. Value Image about Each Case.

Table 3. Accuracy and Iteration Number of Each Classification Result.

Cases	ⓐ	ⓑ	ⓒ	ⓓ	ⓔ	ⓕ	ⓖ	ⓗ	ⓘ
Accuracy	63.77%	59.81%	75.41%	69.31%	75.98%	78.01%	76.08%	74.33%	81.16%
Iteration NO.	39	46	55	35	16	20	29	39	33

Cases	ⓙ	ⓚ	ⓛ	ⓜ	ⓝ	ⓞ	ⓟ	ⓠ	ⓡ
Accuracy	71.32%	60.58%	73.64%	74.36%	74.21%	76.12%	77.33%	74.55%	70.88%
Iteration NO.	30	13	37	15	47	36	24	11	29