Band Selection for Dimension Reduction in Hyper Spectral Image Using Integrated Information Gain and Principal Components Analysis Technique

Kitti Koonsanit, Chuleerat Jaruskulchai, and Apisit Eiumnoh

***Abstract*—Nowadays, hyper spectral image software becomes widely used. Although hyper spectral images provide abundant information about bands, their high dimensionality also substantially increases the computational burden. An important task in hyper spectral data processing is to reduce the redundancy of the spectral and spatial information without losing any valuable details. In this paper, we present band selection technical using principal components analysis (PCA) and information gain (IG) for hyper spectral image such as small multi-mission satellite (SMMS). Band selection method in our research not only serves as the first step of hyper spectral data processing that leads to a significant reduction of computational complexity, but also a invaluable research tool to identify optimal spectral for different satellite applications. In this paper, an integrated PCA and IG method is proposed for hyper spectral band selection. Based on tests in a SMMS hyper spectral image, this new method achieves good result in terms of robust clustering.**

***Index Terms*—Band selection, principal components analysis, PCA, satellite image, information gain, IG.**

I. Introduction

Satellite application in which such fast processing is needed is the dimension reduction of hyper spectral remote sensing data. Dimension reduction can be seen as a transformation from a high order dimension to a low order dimension. Principle Component Analysis (PCA) is perhaps the most popular dimension reduction technique for remotely sensed data [1]. The growth in data volumes due to the large increase in the spectral bands and high computational demands of PCA has prompted the need to develop a fast and efficient algorithm for PCA. In this work, we present on an implementation of Information Gain with PCA dimension reduction of hyper spectral data. For this paper, hyper spectral data was obtained from the Small Multi-Mission Satellite (SMMS)[2] which has a ground pixel size of 100m x 100m and a spectral resolution of 115 channels, covering the range from 450 nm to 950 nm. We focus on a collection of data taken in June 12, 2010 in the northern part of Amnat Charoen province, Thailand. The data consists of 200 x 200 pixels by 115 bands of a total size of 8.86 Mbytes. Fig. 1 shows the colour composites image.

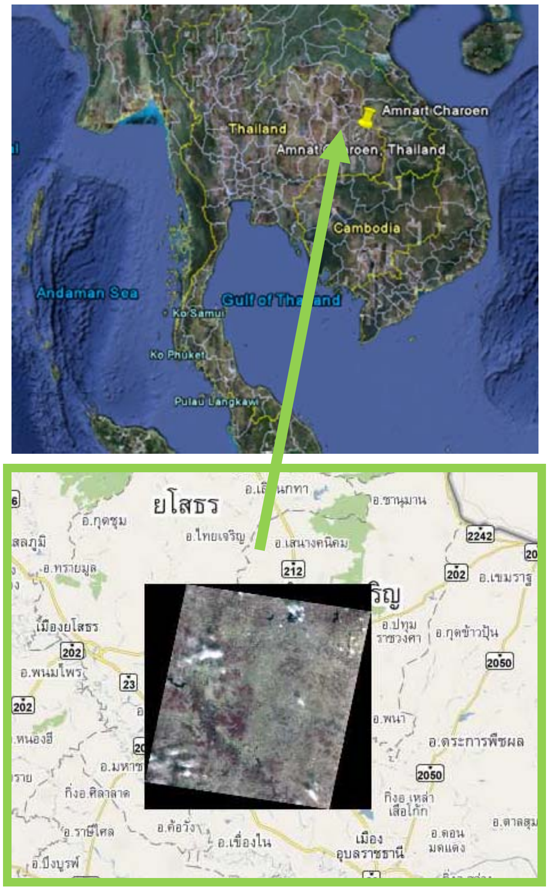


Fig. 1. The color composite image

In this paper, the research interest is focused on comparing the effects of integrated Information Gain (IG) and principal components analysis method (PCA) of band selection on the final clustering results for hyper spectral imaging application. In the following sections, the proposed integrated PCA-IG, are described, and the experimental results are presented later.

II. Methods

A. Dimensionality Reduction

Nowadays, hyper spectral image software becomes widely used. Although hyper spectral images provide abundant information about bands, their high dimensionality also substantially increases the computational burden. An important task in hyper spectral data processing is to reduce

Manuscript received April 9, 2012, revised May 29, 2012. This work was supported by Thailand Graduate Institute of Science and Technology (TGIST) is gratefully acknowledged. The scholar student ID is TG-22-11-53-005D and the grant number is TGIST 01-53-005.

Kitti Koonsanit and Chuleerat Jaruskulchai are with Kasetsart University, Bangkok, Thailand (e-mail: sc431137@hotmail.com).

Apisit Eiumnoh is with National Center for Genetic Engineering and Biotechnology, Patumthani, Thailand.

the redundancy of the spectral and spatial information without losing any valuable details. Therefore, these conventional methods may require a pre-processing step, namely dimension reduction. Dimension reduction can be seen as a transformation from a high order dimension to a low order which eliminates data redundancy. Principal Component Analysis (PCA) [1][3][4][5] is one such data reduction technique, which is often used when analyzing remotely sensed data. The collected hyper spectral image data are in the form of three dimensional image cube, with two spatial dimensions (horizontal and vertical) and one spectral dimension (from SMMS spectrum 1 to spectrum 115 in this study). In order to reduce the dimensionality and make it convenient for the subsequent processing steps, the easiest way is to reduce the dimensions by PCA

B. Background on PCA

PCA is a widely used dimension reduction technique in data analysis. It is the optimal linear scheme for reducing a set of high dimensional vectors into a set of lower dimensional vectors. There are two types of methods for performing PCA, the matrix method, and the data method. In this work, we will focus on the matrix methods. To compute PCA, we follow the general 4 steps given below[1]:

1) Find mean vector in x-space
2) Assemble covariance matrix in x-space
3) Compute eigenvalues and corresponding eigenvectors
4) Form the components in y-space

It has been previously shown that only the first few components are likely to contain the needed information [4]. The number of components that hold the majority of the information is called the intrinsic dimensionality and each data image may have a different intrinsic dimensionality. PCA condenses all the information of an "N" band original data set into a smaller number than "N" of new bands (or principal components) in such a way that maximizes the covariance and reduces redundancy in order to achieve lower dimensionality as shown in the Fig. 2.

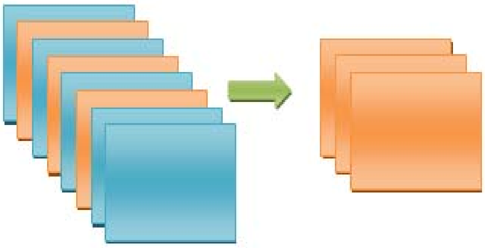


Fig. 2. PCA is a dimension reduction technique

C. Information Gain

One more technique was integrated in our research, Information Gain [6]is a measure of dependence between the feature and the class label. It is one of the most popular feature selection techniques as it is easy to compute and simple to interpret. Information gain of a feature or band X and the class labels Y is calculated as

$$IG(X,Y) = H(X) - H(X \mid Y) \quad (1)$$

Entropy (H) is a measure of the uncertainty associated with a random variable. H(X) and H(X|Y) is the entropy of band X and the entropy of band X after observing Class Y , respectively calculated as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \quad (2)$$

$$H(X \mid Y) = -\sum_i P(y_i)\sum_i P(x_i \mid y_i)\log_2(P(x_i \mid y_i)) \quad (3)$$

The maximum value of information gain is 1. A feature with a high information gain is relevant. Information gain is evaluated independently for each feature and the features with the top-k values are selected as the relevant features. Information Gain does not eliminate redundant features.

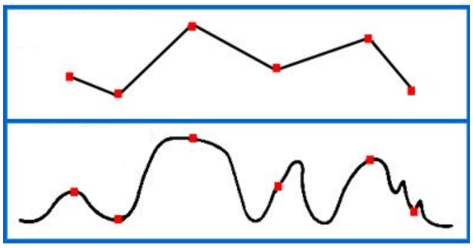


Fig. 3. Band selection by IG.

Result from band selection by IG methods suspected to be notoriously redundant (much data, but not much information).Therefore our methods propose IG methods which was integrated PCA (PCA-IG) to transform into a reduced representation set of features. Transforming the input data into the set of features is called feature extraction. We expect that if the features extracted are carefully chosen, it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.

Main features that classify the different objects should be extracted and preserved. The optimal bands is accordingly defined as the bands that not only maintains the major representation of the original data cube, but also maximally preserves features that separate different object classes [6]. Since the PCA method does not necessarily guarantee that the resulting transformation will preserve the classification information among different object classes, an information gain method is proposed in this study to achieve a better performance to satisfy the optimal band selection criteria. We think that information gain value can use preserved features that separate different object classes.

D. PCA-IG

At the band selection stage, several projection-based methods are studied, including integrated PCA-IG methods. We integrated PCA-IG method which is follow as

$$\text{X Band Selected} = \text{PCA of Band} \cap \text{IG of Band} \quad (4)$$

Let we take two sets PCA of band and IG of band, (∩ = intersection) PCA of band ∩ IG of band, is the set of band selected such that the behind statement is true for the entire element x:

X is a band member of band selected if one means if X is a band member of PCA and X is a band member of IG

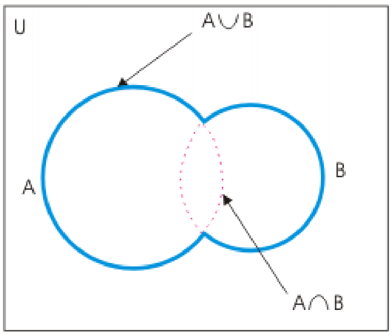


Fig. 4. The intersection set consists of elements common to two sets is a PCA-IG methods.

This paper presents a PCA-IG method, which can effectively reduce the hyper spectral data to intrinsic dimensionality as shown in the Fig. 5.

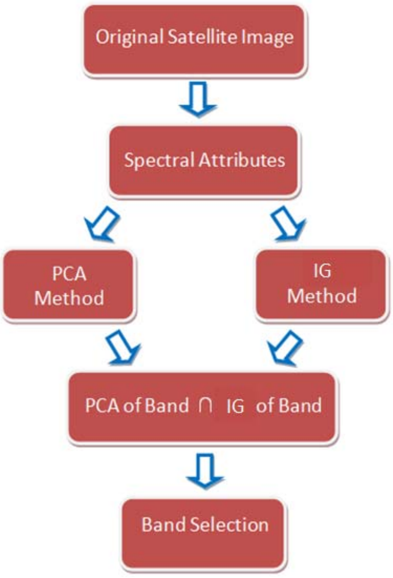


Fig. 5. Overall of our band selection technical

In the PCA-IG, we divide step into 2 parts and combine the results by intersection. In the following sections, the clustering results over the original and the resulting reduced data have been compared.

III. Results

In this section, we present and compare the experimental results obtained by applying each of the techniques, and then evaluating its effectiveness in clustering.

A. Experimental Setup

1) Hyper spectral Data

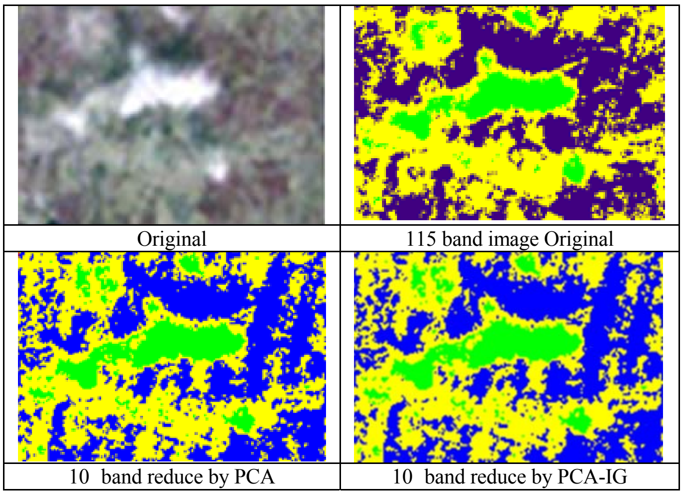
For this paper, hyper spectral test data was obtained from the SMMS imaging. For experiments we focus on a collection of data taken in June 12, 2010 in the northern part of Amnat charoen province, Thailand. The data consists of 200 x 200 pixels by 115 bands.

2) Unsupervised Classifcation Method

The experiments performed in this paper use the simple K-mean from the Weka software package [7][8]. The simple K-Mean is the common unsupervised classification method used with remote sensing data. The effectiveness of the K-Mean depends on reasonably accurate estimation of the k cluster for each spectral class.

B. Experimental Results

In this paper, the PCA-IG is performed using the two techniques described above. While performing the PCA, we have performed 3 experiments as shown in the Fig. 6.


Original

115 band image Original

10 band reduce by PCA

10 band reduce by PCA-IG

Fig. 6. Result image

TABLE I: Show Comparison PCA, PCA-IG, Original Image by Simple K-Mean Clustering

	4 Cluster 115 band Original image	4 Cluster 10 band reduce by PCA	4 Cluster 10 band reduce by PCA-IG
Cluster1	22677 (34%)	22677 (35%)	22677 (34%)
Cluster2	1974 (3%)	1398 (2%)	2093 (3%)
Cluster3	20294 (30%)	20302 (31%)	20185 (30%)
Cluster4	20590 (31%)	21158 (32%)	20580 (31%)
All cluster	100%	100%	100%

From the experiment, it was found that clustering using PCA combined with information gain functional gives the nearest result with 115 band original image and can reduces numbers of attribute from 115 down to 10 attributes. These selected attributes were used as input for clustering algorithms. Table 1 shows the percent clustering obtained for various classes for an example experiment. It can be noticed that the differences in clustering between original image and PCA-IG techniques are very closed. Because features from information gain functional can use preserved features that separate different object classes and, both the representation information and class specific information are included.

One more experiment was tested a proposed band selection on the statlog (landsat satellite) data set from UCI databases [9]. The database consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the classification associated with the central pixel in each neighbourhood. One frame of Landsat MSS imagery consists of four digital images of the same scene in different spectral bands. Two of these are in the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the (near) infra-red. Each pixel is an 8-bit binary word, with 0 corresponding to black and 255 to white. The spatial resolution of a pixel is about 80m x 80m. Each image contains 2340X3380 such pixels.

These data contain 6435 instances. Each instance consists of 36 band attributes. The proposed process was implemented on java environment, and tested on CPU 2.80 GHz Intel(R) Core two duo processor with 1 GB of RAM. From table 2, it can be noticed that the differences in clustering between original image and PCA-IG techniques are very closed. Following the results of the experiment are shown in Table. II.

TABLE II: SHOW COMPARISON PCA, PCA-IG, ORIGINAL IMAGE BY SIMPLE K-MEAN CLUSTERING.

	7 Cluster 36 band Original image	7 Cluster 7 band reduce by PCA	7 Cluster 7 band reduce by PCA-IG
Cluster1	489 (8%)	675 (10%)	608 (9%)
Cluster2	1299 (20%)	1343 (22%)	1362 (21%)
Cluster3	927 (14%)	760 (12%)	742(12%)
Cluster4	572 (9%)	598 (9%)	586 (9%)
Cluster5	1315 (20%)	1504 (23%)	1504 (23%)
Cluster6	1043 (16%)	627 (10%)	881 (14%)
Cluster7	790 (12%)	928 (14%)	752 (12%)
All cluster	100%	100%	100%

IV. RESULTS

Hyper spectral image software becomes widely used. Although hyper spectral images provide abundant information about bands, their high dimensionality also substantially increases the computational burden. Dimensionality reduction offers one approach to hyper spectral image (HSI) analysis. In this paper, we present band selection technical using principal components analysis (PCA) and information gain functional for hyper spectral image such as small multi-mission satellite (SMMS). We tested the proposed process on satellite image data such as small multi-mission satellite (hyper spectral) for unsupervised classification. We compared this classification results between original images and PCA-IG by clustering. The experimental results show that the differences in clustering between original image and PCA-IG techniques are very closed. Because features from information gain functional can use preserved features that separate different object classes and, both the representation information and class specific information are included. A result of this research was developed to provide users have been selected band for hyper spectral image. The outcome of this research will be used in further steps for analysis tools in hyper spectral image processing.

ACKNOWLEDGMENT

This presented paper was supported by budget for overseas academic conference from the faculty of science, Kasetsart University and the graduate school Kasetsart University. The authors would like to thank TGIST. The financial support from Thailand Graduate Institute of Science and Technology (TGIST) is gratefully acknowledged. The scholar student ID is TG-22-11-53-005D and the grant number is TGIST 01-53-005.

REFERENCES

[1] J. A. Richards, Remote Sensing Digital Image Analysis: An introduction, *Spring-Verlag, Berlin*, Heidelberg, 1986.

[2] Small Multi-Mission Satellite (SMMS) Data Available to: http://smms.ee.ku.ac.th/index.php

[3] Agarwal, A.; El-Ghazawi, T.; El-Askary, H.; Le-Moigne, J.; , "Efficient Hierarchical-PCA Dimension Reduction for Hyperspectral Imagery," *Signal Processing and Information Technology, 2007 IEEE International Symposium on* , vol., no., pp.353-356, 15-18 Dec. 2007

[4] Kaewpijit S., Le-Moige J., El-Ghazawi T., "Hyperspectral Imagery Dimension Reduction Using Pricipal Component Analysis on the HIVE," In Science *Data Processing Workshop, NASA Goddard Space Flight Center*, Feb. 2002

[5] P_adraig Cunningham,"Dimension Reduction," Technical Report on Dimension Reduction, University College Dublin, 2007

[6] T. M. Cover and J. A. Thomas."Information Gain," *Elements of Information* Theory. Wiley, 1991.

[7] Remco R. Bouckaert , "WEKA Manual," WAIKATO University, pp.1-303, January 2010.

[8] Rechard Kirkby and Eibe Frank, Weka Explorer User Guide, University of Waikato, New Zealand, 2005.

[9] Frank, A. & Asuncion, A. "*UCI Machine Learning Repository,*" [http://archive.ics.uci.edu/ml/support/Statlog]. Irvine, CA: University of California, School of Information and Computer Science, 2010

[10] Z. Zhao et al., "Advancing Feature Selection Research," Retrieved: Sep 2, 2010, [online]. Available: http://featureselection.asu.edu/featureselection_techreport.pdf

[11] X. Cheng, Y. R. Chen, Y. Tao, C. Y. Wang, M. S. Kim, A. M. Lefcourt. A novel integrated PCA and FLD method on hyperspectral image feature extraction for cucumber chilling damage inspection. *ASAE Transactions, Vol. 47(4),* 1313-1320, 2004.

[12] J. E. Jackson, "A User Guide to Principal Components", New York: John Wiley and Sons, 1991.

[13] I. T Jolliffe I.T, Principal Component Analysis. *Springer- Verlag,* 1986

Kitti Koonsanit received his M.S. degree in computer science from Kasetsart University in 2008. He is currently a Ph. D. candidate in computer science, Kasetsart University, Thailand. His fields of interest include clustering, image processing, image segmentation, band selection, multispectral image, and medical imaging.

Chuleerat Jaruskulchai received her D.Sc. degree in computer science from George Washington University, School of Engineering and Applied Science, USA in 1998. She is currently an Associate Professor and lecturer in the Department of Computer Science, Kasetsart University, Thailand. Her fields of interest and research areas include information retrieval, clustering, text classification, and statistic modeling.

Apisit Eiumnoh received his Ph.D. degree in Soil Genesis & Classification from North Carolina State, UK. He is currently an Associate Professor in National Center for Genetic Engineering and Biotechnology, Patumthani, Thailand.