Dimension Reduction of Hyperspectral Images for Classification Applications*

Pai-Hui Hsu[1], Yi-Hsing Tseng[1] and Peng Gong[2]

[1]Department of Surveying Engineering, National Cheng Kung University, No.1, University Road, Tainan 701, TAIWAN
[2]Department of Environmental Science, Policy and Management, University Of California, 151 Hilgard Hall, Berkeley, CA 94720-3110

Abstract

Hyperspectral images contain rich and fine spectral information, an improvement of land use/cover classification accuracy is expected from the use of such images. However, due to the high dimensionality of data and high correlation between adjacent spectral bands, the classification process may involve a large amount of training samples, result in low efficiency and been hard to improve classification accuracy. In this paper, we tested some feature extraction methods based on wavelet transform to reduce the high dimensionality with losing much discriminating power in the new feature space. An AVIRIS data set with 220 bands and an EO-1 data set with 193 bands were tested to illustrate the performance of the wavelet based methods and be compared with the existing methods of feature extraction.

I. INTRODUCTION

Multispectral sensors have been developed and widely used to observe the earth surface since 1960's. However, traditional sensors can only collect spectral data less than 20 bands due to the inadequate storage space of sensor technology. In recent years, imaging sensors have been improved to collect spectral data with several hundred bands, called imaging spectrometers. The images acquired from imaging spectrometers are also called as hyperspectral images. For example, the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) scanner developed by Jet Propulsion Laboratory (JPL) of NASA provides 224 contiguous spectral channels. The first spaceborne hyperspectral imager in the world, Hyperion, which is one of the three instruments on board of Earth Observing-1 (EO-1) satellite, provides 242 spectral bands (calibrated from 0.4 to 2.5 μm wavelength). Theoretically, using hyperspectral images should increase our abilities in classifying land use/cover types. However, the data classification approach that has been successfully applied to multispectral data in the past is not as effective as for hyperspectral data. The major problem is the high dimensionality of hyperspectral images. As the dimensionality of the feature space increases with the number of bands, the number of training samples needed for image classification must be increased as well. If training samples are insufficient for the need, which is common for the case of using hyperspectral data, the estimation of statistical parameters becomes inaccurate and unreliable. The result is that the classification accuracy first grows and then declines as the number of spectral bands increases while training samples are kept the same, which is often referred to as the Hughes phenomenon (Hughes, 1968).

In general, classification performance depends on four factors: class separability, the training sample size, dimensionality, and classifier type (Hsien, 1998). The focus of our study is the dimension reduction. There are two approaches to reduce the data dimension (Young and Fu, 1986). The first approach is to select a small subset of features directly from the original feature space according to their contribution to the class separability or classification criteria. This dimension reduction process is referred to as feature selection or band selection. The other approach referred as feature extraction is to use all the data from the original feature space and map the effective features and useful information to a lower-dimensional subspace. In other words, the goal of employing feature extraction is to remove the redundant information substantially without sacrificing significant information.

In this paper, we introduce wavelet-based methods of feature extraction which will transfer the spectral data from the original feature space to a time-scale space. The wavelet transform decomposes a signal into a series of shifted and scaled versions of the mother wavelet function. Thus the local energy variation of a hyperspectral signal in different spectral bands at each scale (or frequency) can be detected automatically and provide useful information for hyperspectral data analysis. The local characteristics of wavelet transform will provide information on the oscillation of the spectral curve for each pixel. Different type of materials can be distinguished on the basis of the differences in the time-scale plain. In order to compare with the new proposed methods, we review some existing feature extraction methods. A set of AVIRIS data and a Hyperion data set were tested to illustrate our method and to show the effectiveness of the new feature extraction methods. Graphical diagrams will show an overall comparison of the

* Best paper award at Geoinformatics 2002

1082-4006/02/0801-1$5.00
©2002 The Association of Chinese Professionals in Geographic Information Systems (Abroad)

classification results from existing feature extraction methods and the proposed new methods with respect to the dimensionality and sample size.

II. DIMENSION REDUCTION

The goal of dimension reduction is to reduce the number of features substantially without sacrificing significant information. The process of dimension reduction is to project the data from the high dimensional space to a low dimensional subspace which is formed with fewer but effective features. The benefits of performing feature reduction for remote sensing applications are twofold which are to circumvent the Hughes phenomenon and to reduce the computation time required for classification. Dimension reduction methods can be roughly divided into two categories: (1) feature selection and (2) feature extraction. The basic ideas of these two approaches and their difference are shown in Figure 1. In feature selection, an optimal subset of features is selected directly form the original data by assessing some criteria. In feature extraction, a linear or nonlinear operator T is used to project the data from the original feature space to a lower-dimensional subspace. A brief review on the methods of feature selection and extraction will be conducted below.

Feature selection

The process of feature selection is to choose a best subset of spectral features that will emphasize the discrimination of classes by assessing some criteria. Thus features of no help to classification can be removed from the original data. Feature selection should be conducted based on the test of class discrimination. A method widely used in remote sensing is to determine the separability of different classes (Richards, 1986). Separability is a measurement of probabilistic distance between or within two different classes. The commonly used measurements of separability in feature selection are the Mahalanobis, Divergence, Transformed Divergence, Bhattacharyya, and Jeffries-Matusita, etc (Schowengerdt, 1997). In this method, separability analysis is performed on the training data to estimate the expected error in the classification for various feature combinations (Swain and Davis, 1978).

Feature Selection using statistical distance measures has been widely studied and successfully applied. However, as the dimension of data increases, the combination of features to be examined increases exponentially, resulting in unacceptable computational cost [Lee and Landgrebe, 1993]. Suppose that the number of spectral bands is n, the problem of feature selection is to select an optimal subset of m with $m<n$. The number of feature combinations that need to be considered equals $n!/((n-m)!m!)$. This number will be large for hyperspectral data and leads to low efficiency in computation. Some algorithms such as Branch-and-Bound algorithm, Sequential Forward, Sequential Backward and Max-Min Feature Selection that could determine an optimal or sub-optimal feature set were proposed to make computation feasible (Young and Fu, 1986). In the experimental test, the Sequential Forward algorithm was applied based on the calculation of Bhattacharyya distance.

Feature extraction

The purpose of feature extraction is to reduce the high dimensional data to relatively few features without much loss in overall classification accuracy. The type of feature extraction can be linear or nonlinear. The former has been widely used in multispectral images, and the latter is more complex for practical application. The most commonly used method of feature extraction is Principal Components Transformation (PCT). In PCT, an orthogonal subspace projection is performed on the hyperspectral images and produces a new sequence of uncorrelated images. Usually the first few components con-

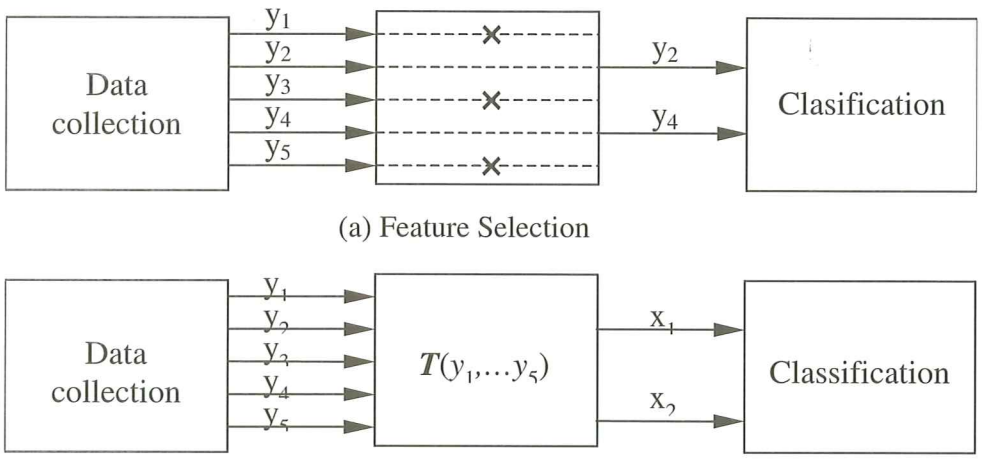


Figure 1. The diagram of dimension reduction.

tain the most variances, and the later components tending to show little variance could be ignored. Therefore, the essential dimensionality of the classification space will be reduced and thus the classification speed will be improved. Although this method can effectively provide good classification accuracy, it is sensitive to noise and has to be performed with the whole data set (Schowengerdt 1997).

In contrast to the PCT concept of taking the global covariance matrix of the full data set into account, Discriminant Analysis Feature Extraction (DAFE), or called Canonical Analysis (Richard, 1986), generates a transformed set of feature axes, in which class separation is optimized. This approach uses the ratio of a between-class covariance matrix to within-class covariance matrix as a criterion function. Thus a transformation matrix is determined to maximize the ratio, that is, the separability of classes will be maximized after the transformation. Although the discriminant analysis performs well for most cases, there are several drawbacks for this method. First, the approach delivers features only up to the number of classes minus one. Second, when the mean values are similar or the same, the extracted feature vectors are not reliable. Furthermore, if a class has a mean vector very different from the other classes, the between-class covariance matrix will be biased toward this class and will result in ineffective features (Tadjudin and Landgrebe, 1998).

Lee and Landgrebe (1993) showed that useful features could be separated from redundant features the by decision boundaries. The approach is called Decision Boundary Feature Extraction (DBFE). It was shown that all the features needed for classification are normal to the effective decision boundary. A decision boundary feature matrix (DBFM) was defined to predict the intrinsic discriminant dimension and to extract discriminative information from the decision boundary. In order to determine the effective decision boundary, the majority of training samples are first selected. The number of training samples required could be large for high dimensional data. For hyperspectral images, the number of training samples is usually not enough to prevent singularity or yield a good covariance estimate. In addition, DBFE for more than two classes is sub-optimal (Tadjudin and Landgrebe, 1998).

From the perspective of signal processing, the result of a Fourier Transform (FT) is useful because of that a signal's frequency content is of great importance. The frequency power spectrum of FT can localize information about global patterns of the spectral curve. Especially the low-frequency content gives the signal its identity and is the most important part. The Fourier Feature Extraction (FFE) method was proposed by Hsu and Tseng (1999) to obtain spectral features. In FFE, the discrete Fourier Transform (DFT) is implemented on the spectral data for each pixel and a series of Fourier power spectrum is produced. Like PCT method, the first few components formed by DFT may contain the most important information. Thus they can be treated as the most important features for classification.

III. WAVELET BASED FEATURE EXTRACTION

In the past two decades, wavelet transform has been developed as a powerful analysis tool of signal processing, and also been successfully applied in applications such as image processing, data compression and pattern recognition (Mallat, 1999). From the perspective of signal processing, the hyperspectral curve of each pixel can be thought as a one-dimensional signal. By using the wavelet transform, the hyperspectral signal is transformed from the spectral space to the time-scale space. Furthermore, the wavelet transform decomposes a signal into a series of shifted and scaled versions of the mother wavelet function. Thus, the local energy variation of a hyperspectral signal in different spectral band at each scale (or frequency) can be detected automatically and provide some useful information for hyperspectral data analysis. In this study, some feature extraction methods based on wavelet transform were used to extract important features from the hyperspectral images.

Linear wavelet feature extraction

The orthogonal wavelet transform in terms of multi-resolution analysis (MRA) can decompose a signal into the low-frequency components that represent the optimal approximation, and the high-frequency components that represent the detailed information of the original signal (Mallat, 1989). The decomposition coefficients in a wavelet orthogonal basis can be computed with a fast algorithm that cascades discrete convolutions with conjugate mirror filters h and g, and subsamples the output. The decomposition formulas are described as following (Mallat, 1999):

$$a_{j+1}[p] = \sum_{n=-\infty}^{\infty} h[n-2p]a_j[n] \quad (1)$$

$$d_{j+1}[p] = \sum_{n=-\infty}^{\infty} g[n-2p]a_j[n] \quad (2)$$

where a_j is the approximation coefficients at scale 2^j, a_{j+1} and d_{j+1} are respectively the approximation and detail components at scale 2^{j+1}. In practice the original signal s is always expressed as coefficient a_L. A multilevel orthogonal wavelet decomposition of a_L is composed of wavelet coefficients of signal s at scales $2^L < 2^j \leq 2^J$ plus the remaining approximation at the largest scale 2^J:

$$[\{d_j\}_{L<j\leq J}, a_J] \quad (3)$$

Assume that the length of a_j is N, one may notice that the downsampling procedure in the wavelet decomposition which reduces the length of a_{j+1} to $N/2$ achieves the dimension reduction of a_j. Specifically, we decomposed the hyperspectral signature using wavelet decomposition and then selected the fewest wavelet coefficients required to perform the dimensionality reduction. In this paper, both the approximation a_{j+1} and detail d_{j+1} extracted by using the multilevel

wavelet decompositions are used as features for classification. Figure 2 shows the shape of a_{j+1} and d_{j+1} of a hyperspectral curve. This method is referred to Linear Wavelet Feature Extraction (Linear WFE) because of the property of linear data transformation.

Usually the detail components $\{d_j\}_{L<j\leq J}$ are treated as noises or unimportant information. However, our experiments in this study showed that the detail components extracted by the band-pass filters are important for classification (see experiment I). Therefore, the approximation and detail components should be combined to perform the feature extraction of hyperspectral images.

Nonlinear wavelet transform feature extraction

Let us sort the wavelet coefficients $[\{d_j\}_{L<j\leq J}, a_J]$ in decreasing order and take the first a few biggest coefficients as the important features for classification. Suppose that the band number of a hyperspectral signature is n, m features are selected from the first a few coefficients after wavelet transform where $m \leq n$. One may notice that the m largest coefficients may include the detail components of wavelet decomposition if there are singular points or sharp variation in the hyperspectral curve. This method which combines the approximation and detail information of wavelet transform is referred to as Non-Linear Wavelet Feature Extraction (Non-linear WFE).

The best-basis algorithm

The main purpose of best-basis algorithm is to search a best basis which can be used to reconstruct the best approximation of the original signal (Coifman and Wickerhauser ,1992). This method first expands a given signal into a library of orthonormal bases. The library can be constructed by wavelet packets or local trigonometric bases which have a binary tree structure. The nodes of the binary tree represent the subspaces of a signal with different time-frequency localization characteristics. Then a complete basis called a best basis which minimizes a certain cost functional C (e.g. entropy or norm) is searched among this binary tree from leaves to root. Entropy used in the best-basis algorithm is an index that measures the flatness of the energy distribution of a signal. Minimizing entropy will lead to an efficient representation for the signal. Therefore, the best-basis algorithm is good for signal compression but may not be good for classification problems (Saito and Coifman, 1994).

Local discriminant bases

The Local Discriminant Bases (LDB) was proposed by Saito and Coifman (Saito and Coifman, 1994) to search for a best basis for classification. In this method, the discriminating function D between the nodes of the tree is calculated from a known training data set. Function D can be a certain distance function between different classes. Then a complete orthonormal basis, called LDB, that can distinguish signal features among different classes is selected form the library tree. To make this algorithm fast, the discriminant functional D needs to be additive. In this paper, two kinds of distance are used as discriminant measurements. They are respectively cross entropy and the l^2 norm of the wavelet coefficients (Saito and Coifman, 1994). One may expect that the LDB which combines the characteristics of both the statistic-based and the frequency-based methods can obtain a better result than other methods.

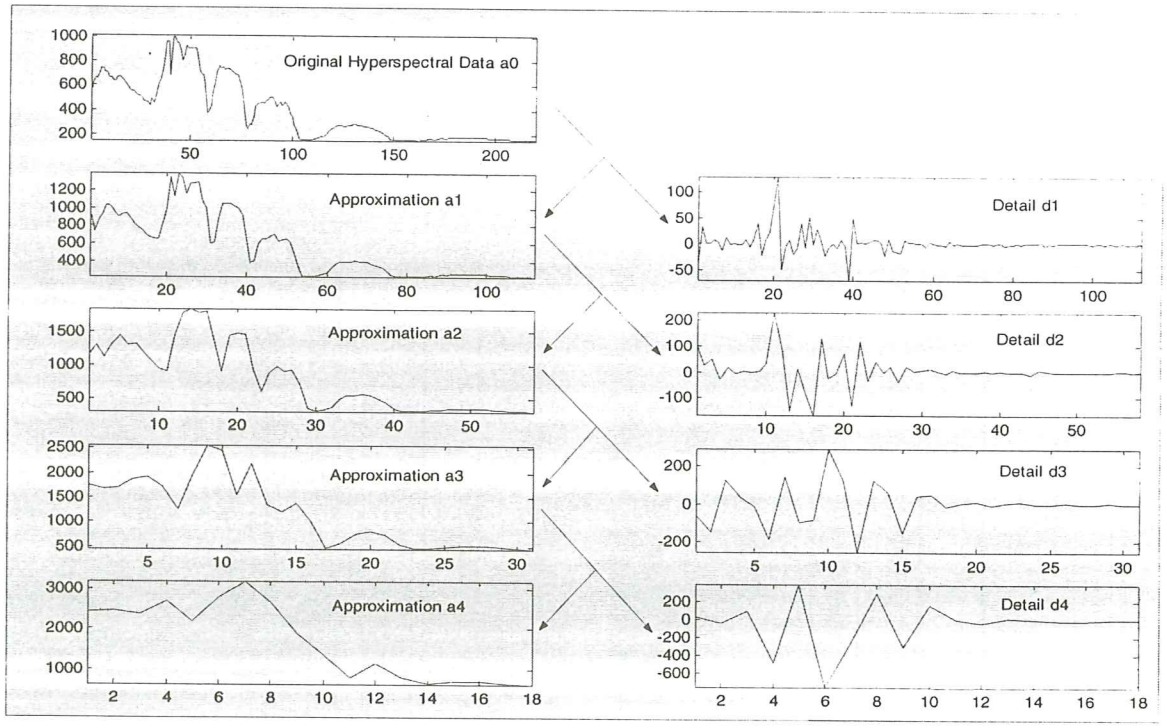


Figure 2. The results of linear wavelet feature extraction of hyperspectral data

IV. EXPERIMENTIAL RESULTS AND ANALYSIS

Experiment I

The test data sets (Figure 3) of experiment I is an AVIRIS image taken in 1992 and acquired from the website of School of Electrical and Computer Engineering at Purdue University (http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/).The image is an agriculture field on northwest Indiana. The original data set has 224 spectral bands from 0.4 to 2.45 μm with 10 nm spectral resolution. The number of bands is 220 after removing 4 noisy bands. The image size of the test field is 68´85. The ground truth data shown in Figure 3(b) include four classes Corn, Grass, Soybean-1 and Soybean-2. The number of pixels of each known class is listed in Figure 3(c).

The main purpose of experiment I is to compare the feature extraction methods mentioned in section 2 with the linear wavelet feature extraction methods in terms of classification accuracy. The feature extraction methods were firstly applied to the test data to reduce the dimensionality then the extracted features were used as the inputs in classification. The approach used for classification was the Maximum Likelihood (ML) Classifier. The accuracy of each feature extraction method was shown in Figure 4. Because the number of features delivered by the DAFE method is only up to the number of classes minus one, the classification accuracy is only estimated in two and three features. Furthermore, in order to see the effect of limited training samples, four different sizes (all, 300, 175, and 105) were used to estimate the statistics of the training areas for classification.

We summarize the results in the following. Firstly, as the number of features increase, the accuracy of classification in the beginning stage increases. Secondly, when the number of training samples reduces, the classification accuracy reduces. These two results conform to the Hughes phenomenon. Thirdly, when the number of bands is reduced to the number smaller than 30, all the methods except for the DBFE and WFE_D (the detail of linear WFE) have similar classification accuracies. Comparing other methods, the approximation of linear WFE (denoted by WFE_A) has the advantage of fast computation time. Fourthly, one may find that the result of DBFE is not as good as expected, the reason mentioned above is that the efficiency of DBFE strongly depends on the number and the distribution of the training samples in the feature extraction stage. Finally, when the number of extracted feature is 5, the result of FFE is the best. When the number of extracted features is 3, the classification accuracy using DAFE is the best.

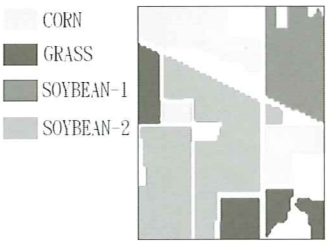
One interesting thing in this experiment is the result of the linear WFE. The detail components of wavelet decomposition are often interpreted as noises or unimportant information of a signal. However, when the number of features reduces to 11, the classification accuracies of WFE_D are better than WFE_A and other methods. It means that the detail components extracted by the band-pass filters are most important for classification.

Experiment II

The major purpose of experiment II is to compare the nonlinear WFE, best basis algorithm and LDB with the linear WFE and PCT. The data set is the same as the one used in experiment I. The nonlinear WFE and best basis algorithm do not need training samples. On the opposite, the LDB method needs the training samples before classifications in order to determine what the best tree is. The size of training samples for classification is 105. Figure 5(a) and 5(b) respectively show the best tree of LDB using cross entropy and l^2 norm. One may find that the number of decomposition level using l^2 norm is smaller than using cross entropy. The smaller in the level numbers of the best tree is, the shorter is the computation time required when performing feature extraction on the unknown data set.

Figure 6 illustrates the results of the classification by using the wavelet-based feature extraction methods. Firstly, as the number of features increases, the classification accuracy first increases and then decreases. This phenomenon has a great resemblance to the result of experiment I. Secondly, the classification results using Nonlinear WFE, best basis and LDB are better than the results using linear WFE and PCT. Thirdly, when the number of features reduces to 31 and 5, the LDB using l^2 norm is better than other methods. When the number of features reduces to 18 and 11, the best basis algorithm provides the best results. When the number of features re-

(a) Test Field



(b) Ground Truth Map

Class Name	Pixel Number
corn	998
grass	702
soybean-1	727
soybean-2	1905

(c) Pixel Number of each classes

Figure 3. The test data delivered by AVIRIS on NW Indiana in 1992

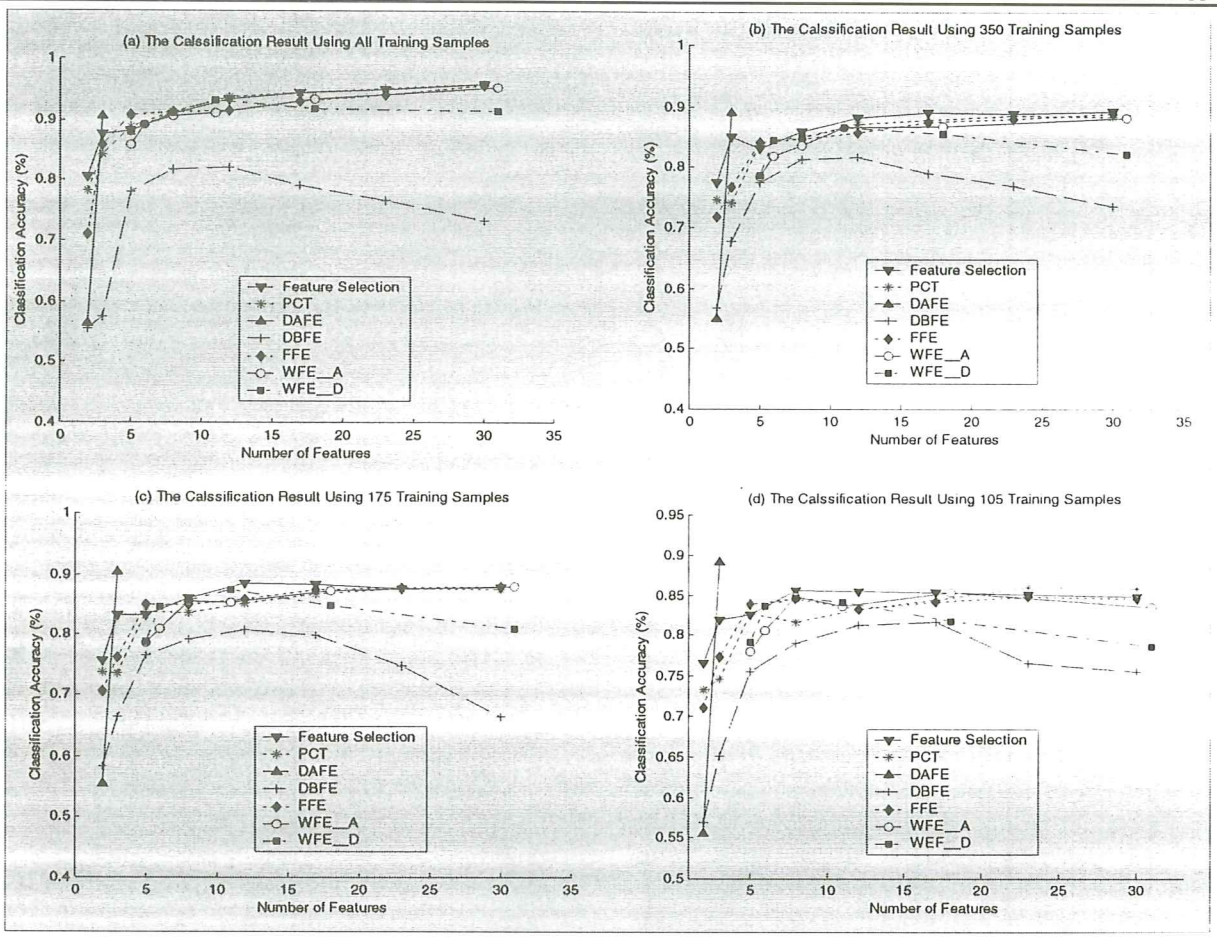
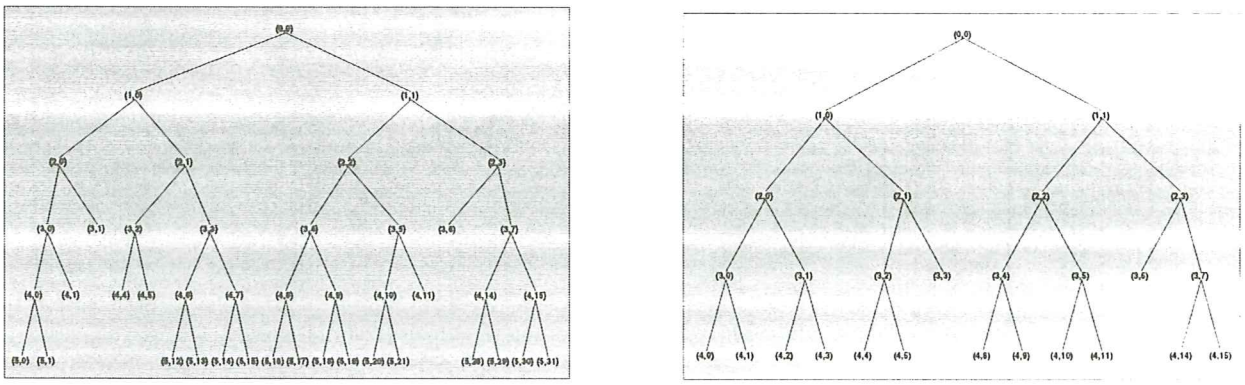


Figure 4. Classification accuracy using different feature extraction methods.

duces to 8 and 6, the non-linear WFE is the best. All these results show that using the combination of approximation and detail components of wavelet decomposition will increase the ability of searching for useful features and obtain better classification accuracy than the linear WFE. In general, the result of best basis algorithm is more stable than that of the other wavelet based methods.

Experiment III

The test data sets (see Figure 7) of experiment III is a hyperspectral image from Hyperion acquired on March 27, 2001 in Argentina. The size of the image is 85´100. A simplified atmospheric correction approach was performed on the data set to retrieve the surface reflectance for each pixel. The number of bands is 193 from 0.4 to 2.4 μm with 10 nm spectral resolution after the preprocessing. The training areas shown



(a) (b)

Figure 5. (a) The best tree of LDB using cross entropy (b) The best tree of LDB using l^2 norm.

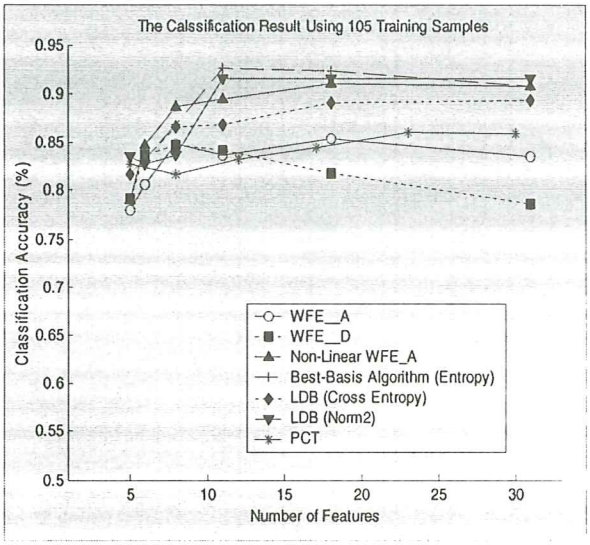


Figure 6. The results of classification accuracy using wavelet-based feature extraction methods.

in Figure 7(b) are selected using a hybrid method of unsupervised and supervised training. The ISODATA algorithm was first applied to the data, then a labeled cluster map of the training area is produced by an analyst using the n-dimensional visualizer tool of the ENVI software. Four classes are labeled and the pixel number for each class is listed in figure 7(c.)

Figure 8 shows the results of classification accuracies using most of the feature extraction methods discussed in this paper. Like in experiment I, two different sizes of training samples were used to see the effect of limited training samples. Figure 8(a) and 8(b) are respectively the results by using 100 and 50 training samples in classification. The results of DAFE and DBFE were not affected by the number of training samples, but their results are not as good as the PCT and the wavelet-based methods of feature extraction. In general, the result of the PCT is the best and of the DBFE is poor. These two figures do not show prominent difference among wavelet based feature extractions. However, the nonlinear WFE has a slightly better result than the other wavelet-based methods.

V. CONCLUSION

The goal of employing feature extraction is to reduce the number of features substantially without sacrificing significant information. Thus, the accuracy of classification could be preserved and the speed of computation could be increased. This paper compared some existing methods of feature extraction. In addition, the wavelet-based feature extraction methods were developed by using the wavelet decomposition algorithm and compared with other existing methods for their effect on classification results. The results showed that all of the wavelet-based feature extraction methods can reduce the dimensionality of hyperspectral data and preserve the accuracy of classification. In the existing methods, the PCT is based upon the global covariance matrix of the full set of image data and not suitable for the case of multiple classes. And the result of DBFE and DAFE depends on the number and distribution of training samples. For the wavelet-based methods, the new feature spaces formed by the wavelet transform are more meaningful and more stable than other methods. This paper also showed that the detail components of wavelet decomposition contain useful and important information for classification. Thus, the combination of approximation and detail information is needed to obtain more suitable and stable spectral features for classification.

ACKNOWLEDGMENTS

This research project was sponsored by the National Science Council of Republic of China under the grants of NSC89-2211-E006-118.

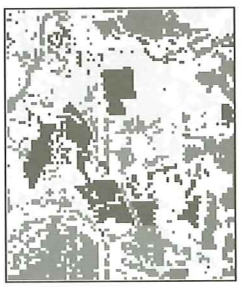
REFERENCES

[1] Coifman, R. R. and M. V. Wickerhauser, 1992, Entropy-based algorithms for best basis selection, *IEEE Trans. Inform. Theory*, 38(2):713-719.

[2] Hughes, G. F., 1968, On the mean accuracy of statistical pattern recognizers, *IEEE Trans. Inform. Theory*, IT-14, pp. 55-63,

[3] Hsien, P. F. and D. Landgrebe, 1998, *Classification of High Dimensional Data.* Ph. D. Dissertation, School of Electrical and Computer Engineering, Purdue University, West Lafayette, In-

(a) Test Field



(b) Training areas

Class Name	Pixel Number
class 1	1090
class 2	940
class 3	1698
class 4	1316

(c) Pixel Number of each classes

Figure 7. The test data is Hyperion image acquired on March 23, 2001, Argentina.

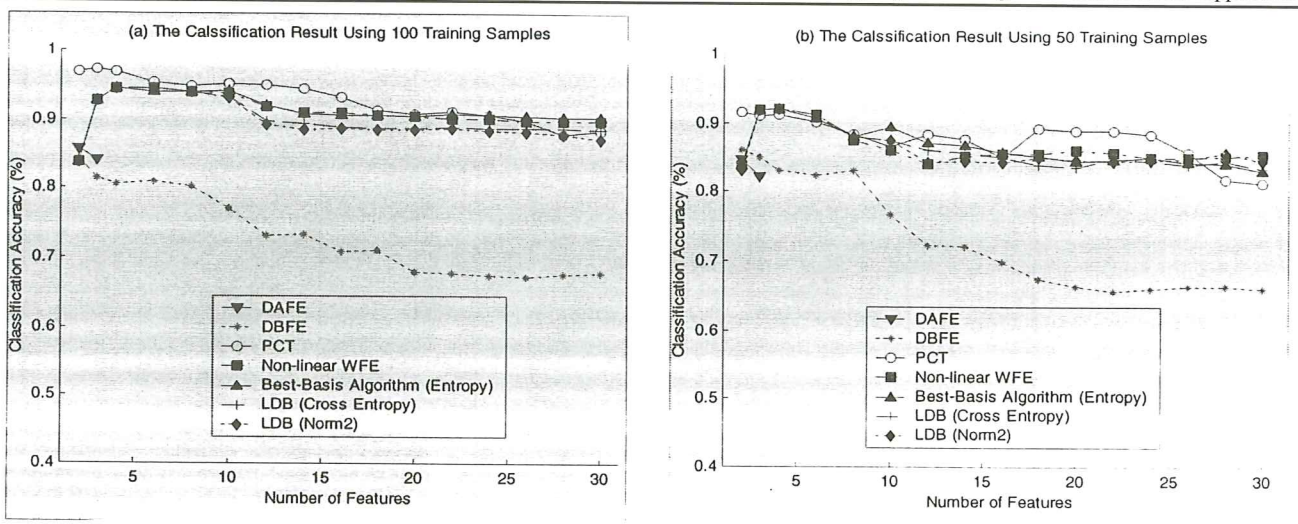


Figure 8. The results of classification accuracy after using wavelet-based feature extraction methods

diana.

[4] Hsu, P. H. and Y. H. Tseng, 1999, Feature extraction for hyperspectral image, *Proc. 20th Asia Conference on Remote Sensing*, vol. 1, pp. 405-410, Hong Kong, Nov. 1999.

[5] Jimenez, L. O. and D. Landgrebe, 1995, *High Dimensional Feature Reduction Via Projection Pursuit*, PhD Thesis, School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana.

[6] Lee, C. and D. Landgrebe, 1993, *Feature Extraction and Classification Algorithms for High Dimensional Data*, PhD Thesis, School of Electrical Engineering, Purdue University, West Lafayette, Indiana, January.

[7] Mallat, S. G., 1989, A theory for multiresolution signal decomposition: The Wavelet Representation, *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 11(7): 674-693.

[8] Mallat, S. G., 1999, *A Wavelet Tour of Signal Processing*, Academic Press, 2nd edition, San Diego.

[9] Richards, J. A.,1993, *Remote Sensing Digital Image Analysis?An Introduction*, Springer-Verlag Berlin Heidelberg, 2nd Edition.

[10] Saito, H. and R. R. Coifman, 1994, Local discriminant basis, *Proc. SPIE*, vol. 2303, pp. 2-14, Jul. 1994.

[11] Schowengerdt, R. A.,1997, *Remote Sensing: Models and Methods for Image Processing*, Academic Press.

[12] Swain , P. H. and Shirley M. D, 1978, *Remote Sensign: The Quantitative Approach*, McGRAW W-HILL.

[13] Tadjudin, S. and D. Landgrebe, 1998, *Classification of High Dimensional Data with Limited Training Samples*, PhD Thesis, School of Electrical and Computer Engineering, West Lafayette, Indiana.

[14] Young, T. Y. and K. S. Fu, 1986, *Handbook of Pattern Recognition and Image Processing*, College of Engineering, University of Miami, Coral Gables, Florida.