FEATURE EXTRACTION OF HYPERSPECTRAL IMAGES USING MATCHING PURSUIT

P.H. Hsu

Associate Research Fellow, National Science and Technology Center for Disaster Reduction,
106, No. 200, Sec. 3, Hsinhai Rd., Taipei, Taiwan, Republic of China - paihui@naphm.ntu.edu.tw

TS Ths 5

KEY WORDS: Hyper-Spectral Sensing, Wavelet Transform, Matching Pursuit, Feature Extraction, Classification

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, the classification methods that have been successfully applied to multispectral data in the past are not as effective as to hyperspectral data. The major cause is that the size of training data set does not correspond to the increase of dimensionality of hyperspectral data. Actually, the problem of the "curse of dimensionality" emerges when a statistic-based classification method is applied to the hyperspectral data. A simpler, but sometimes very effective way of dealing with hyperspectral data is to reduce the number of dimensionality. This can be done by feature extraction that a small number of salient features are extracted from the hyperspectral data when confronted with a limited set of training samples. In this paper, we tested some proposed feature extraction methods based on the wavelet transform to reduce the high dimensionality without losing much discriminating power in the new feature space. In addition, a new feature extraction method based on the matching pursuit with wavelet packet is used to extract useful features for classification. An AVIRIS data set was tested to illustrate the classification performance of the new method and be compared with the existing wavelet-based methods of feature extraction.

1. INTRODUCTION

Since the mid 1980s, the new technology of imaging spectrometer with two-dimensional area arrays of detector elements was developed to collect spectral data with a large number of bands simultaneously (Goetz *et al.*, 1985). The value of this technique lies in the ability to construct an effectively continuous reflectance spectrum for each pixel of the sense. Because of the large number of spectral bands, the images acquired with imaging spectrometers are also referred to as hyperspectral images which are distinguished from the multispectral images with only three to ten bands. The rich and detailed spectral information provided by hyperspectral images can be used to identify and quantify a large range of surface materials which cannot be identified by multispectral images. By means of the solar reflected spectrum measured by imaging spectrometers, a wide range of scientific researches and applications have being proposed based on the spectral analysis (Lillesand and Kiffer, 2000).

1.1 Curse of Dimensionality

Seemingly the high dimensionality of hyperspectral data should increase the abilities and effectiveness in classifying land use/cover types. However, the classification methods that have been successfully applied to multispectral data in the past are not as effective as to hyperspectral data. The major cause is that the size of training data set does not adapt to the increasing dimensionality of hyperspectral data. If the training samples are insufficient for the needs, which is common for the hyperspectral case, the estimation of statistical parameters becomes inaccurate and unreliable. As the dimensionality increases with the number of bands, the number of training samples needed for training a specific classifier should be increased exponentially as well. The rapid increase in training samples size for density estimation has been termed the "curse of dimensionality" by Bellman (1961), which leads to the "peaking phenomenon" or "Hughes phenomenon" in classifier design (Hughes, 1968). The consequence is that the classification accuracy first grows and then declines as the number of spectral bands increases while training samples are kept the same. For a given classifier, the "curse of dimensionality" can only be avoided by providing a sufficiently large sample size. The more complex the classifier, the larger should the ratio of sample size to dimensionality be to avoid the curse of dimensionality. However, in practice, the number of training samples is limited in most of the hyperspectral applications. Furthermore, the high dimensionality of hyperspectral data makes it necessary to seek new analytic methods to avoid a vast increase in the computational time. A simpler, but sometimes very effective way of dealing with high-dimensional data is to reduce the number of dimensions (Lee and Landgrebe, 1993; Benediktsson *et al.*, 1995; Landgrebe, 2001). This can be done by feature selection or extraction that a small number of salient features are extracted from the hyperspectral data when confronted with a limited set of training samples.

1.2 Spectral Feature Extraction

Feature extraction is generally considered a data mapping procedure which determines an appropriate subspace of dimensionality M from the original feature space of dimensionality N ($M \leq N$) (Fukunaga, 1990; Lee and Landgrebe, 1993; Jain *et al.*, 2000). The way of feature extraction can be a linear or nonlinear data transformation. Regardless of how the data transformation is implemented, the feature extraction algorithm must be designed to preserve the information of interest for a special problem such as compression, denoising, or classification. For example, in

hyperspectral image classification, effective features are those which are most capable of preserving class separability.

The most commonly used method of feature extraction is Principal Components Transformation (PCT) (Fukunaga, 1990; Jain *et al.*, 2000; Landgrebe, 2001). PCT is an orthogonal transformation to produce a new sequence of uncorrelated images called principal components. Only the first *M* components are used as the features for the image representation or classification. The transformation matrix of PCT consists of a Karhunen-Loéve basis whose vectors are ordered by the decreasing sequence of the eigenvalues of covariance matrix of the total hyperspectral data set. This would result in the best fit of the approximation which has the minimum mean-square error (Mallat, 1999). However, it is sensitive to noise and has to be performed with the whole data set. In contrast to the PCT which takes the global covariance matrix into account, Linear Discriminant Analysis, or called Canonical Analysis (Richards, 1993), generates a transformed set of feature axes, in which class separation is optimized (Lee and Landgrebe, 1993; Jimenez and Landgrebe, 1995). This approach called Discriminant Analysis Feature Extraction (DAFE) uses the ratio of between-class covariance matrices to within-class covariance matrices as a criterion function. A transformation matrix is then determined to maximize the ratio, that is, the separability of classes will be maximized after the transformation. Although the discriminant analysis is an effective and practical algorithm for deriving effective features in many circumstances, 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 of different classes 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). Finally, in order to estimate the between-class and within-class scatter matrices reliably, the number of training samples should be large enough. However, this is often not a common circumstance for hyperspectral images. Lee and Landgrebe (1993) showed that useful features could be separated from redundant features by decision boundaries. The algorithm is called Decision Boundary Feature Extraction (DBFE) because it takes full advantages of the characteristics of a classifier by selecting features directly from its decision boundary. Since the method depends on how well the training samples approximate the decision boundaries, the number of training samples required could be much more for high dimensional data because it computes the class statistical parameters at full dimensionality. For hyperspectral images, the number of training samples is usually not enough to prevent singularity or to yield a good covariance estimate. In addition, DBFE for more than two classes is sub-optimal (Tadjudin and Landgrebe, 1998). The DBFE method is also computationally more intensive than the other methods.

2. WAVELET-BASED FEATURE EXTRACTION

In the past two decades, wavelet transform (WT) has been developed as a powerful analysis tool for signal processing, and also has been successfully applied in applications such as image processing, data compression and pattern recognition (Mallat, 1999). Due to the time-frequency localization properties, discrete wavelet and wavelet packet transforms have proven to be appropriate starting point for the classification of the measured signals (Pittner and Kamarthi, 1999). The WT decomposes a signal into a series of shifted and scaled versions of the mother wavelet function. 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 image classification. Several feature extraction methods based on the WT have been proposed for hyperspectral images (Hsu and Tseng, 2000; Hsu, 2003). The general process of the wavelet-based feature extraction methods is illustrated in Figure 1. Firstly, wavelet or wavelet packet transforms are implemented on the hyperspectral images and a sequence of wavelet coefficients is produced. Then, a simple feature selection procedure associated with a criterion is used to select the effective features for classification. The criterion of feature selection can be designed for signal representation or classification. In the stage of feature selection shown in Figure 1, some training data may be needed as samples to find the effective features for classification. Unlike the existing feature extraction methods such as DAFE and DBFE which need to estimate the statistic parameters at full dimensionality, the wavelet-based feature extraction optimizes the criterion in a lower dimensional space. Thus the problem of limited training sample size can be avoided.

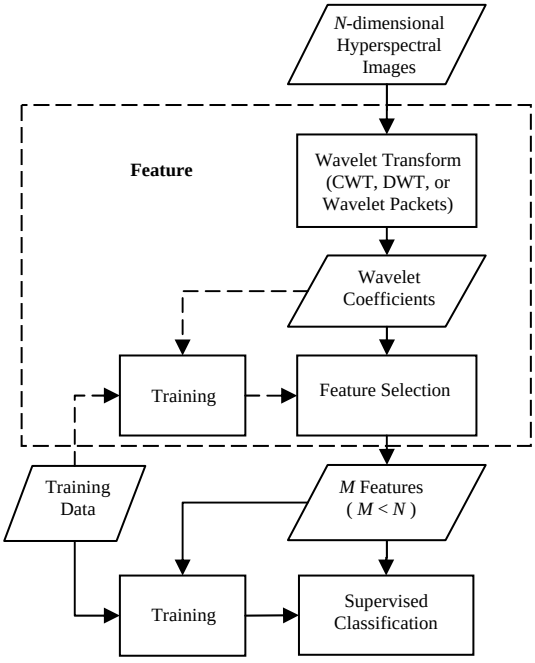

Figure 1. The general flow chart of wavelet-based feature extraction.

2.1 Orthogonal Wavelet Decomposition

The orthogonal wavelet transform in terms of multi-resolution analysis (MRA) can decompose a signal x into the low-frequency components that represent the optimal approximation, and the high-frequency components that represent the detailed information (Mallat, 1989. The inner coefficients of x in a wavelet orthogonal basis can be computed with a fast algorithm that cascades discrete convolutions with Conjugate Mirror Filters (CMF) h and g, and sub-samples the output. The decomposition formulas are described as following):

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

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

where $\bar{h}[n] = h[-n]$ and $\bar{g}[n] = g[-n]$. a_j is the approximation coefficients at scale 2^j, and a_{j+1} and d_{j+1} are respectively the approximation and detail components at scale 2^{j+1}. There are some necessary and sufficient conditions associated with the conjugate mirror filters h and g, so that the perfect reconstruction of signal x can be achieved without losing information. Figure 2 shows the diagram of a fast wavelet decomposition calculated with a cascade of filtering with $\bar{h}$ and $\bar{g}$ followed by a factor 2 sub-sampling. Assume that the length of a_j is N, one may notice that the sub-sampling procedure in the wavelet decomposition shown in Figure 2 which reduces the length of a_{j+1} to $N/2$ achieves the dimensionality reduction of a_j. In practice, the original signal x in Figure 2 is always expressed as a sequence of coefficients a_L. A multilevel orthogonal wavelet decomposition of a_L is composed of wavelet coefficients of signal x at scales $2^L < 2^j \le 2^J$ plus the remaining approximation at the largest scale 2^J:

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

It is calculated from a_L by iterating formula (1) and (2).

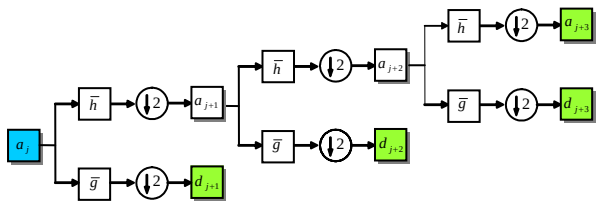

Figure 2. Fast orthogonal wavelet decomposition

2.2 Linear Wavelet Feature Extraction

The sub-sampling shown in Figure 2 motives us to reduce the dimensionality of hyperspectral data by wavelet decomposition. Firstly, the wavelet decompositions of (1) and (2) were implemented on the hyperspectral data, and then only the $M = 2^{-l}$ first scaling and wavelet coefficients at scales $2^j > 2^l$ are selected as features. One may prove that the selected features $[\{d_j\}_{l<j\le J}, a_J]$ are useful for data representation. Because the linear wavelet transformation of x from large scale wavelet coefficients are equivalent to the finite element approximation over uniform grids, we call this method Linear Wavelet Feature Extraction (Linear WFE).

In this method, the large amplitude wavelet coefficients at small scales would not be selected as features. However, the wavelet coefficients with large amplitudes are generated by the singularities of the spectral curve which may involve important information for representation or classification. Hsu (2003) suggested that the approximation and detail components at each scale of linear WFE should be combined together to extract better features of hyperspectral images for classification. This can be done by non-linear wavelet feature extraction.

2.3 Non-Linear Wavelet Feature Extraction

Linear WFE method which selects the M wavelet coefficients independently of original spectrum x at larger scales can be improved by choosing the M wavelet coefficients depending on the x. This can be done by sorting the coefficients $[\{d_j\}_{L<j\le J}, a_J]$ calculated by the multilevel orthogonal wavelet decomposition in decreasing order. Then the M largest amplitude wavelet coefficients are selected as the important features of x for classification. The non-linear approximation calculated from the M largest amplitude wavelet coefficients including the approximation and detail information can be interpreted as an adaptive grid approximation, where the approximation scale is refined in the neighborhood of singularities (Mallat, 1999). Thus this feature extraction method based on the non-linear approximation is called Non-Linear Wavelet Feature Extraction (Non-linear WFE).

2.4 Best Basis Feature Extraction

2.4.1 Wavelet Packets: Wavelet packets were introduced by Coifman *et al.* (1992) by generalizing the link between multiresolution approximations and wavelets. In the orthogonal wavelet decomposition algorithm described in Section 2.1, only the approximation coefficients are split iteratively into a vector of approximation coefficients and a vector of detail coefficients at a coarser scale. In the wavelet packet situation, each detail coefficients vector is also decomposed into two parts using the same approach as in approximation vector splitting. This recursive splitting shown in Figure 3 defines a complete binary tree of wavelet packet spaces where each parent node is divided in two orthogonal subspaces. The nodes of the binary tree represent the subspaces of a signal with different time-frequency localization characteristics. Any node in the binary tree can be labelled by (j, p), where 2^j is the scale and p is the number of nodes that are on its left at the same scale. Suppose that we have already constructed a wavelet packet space $\mathbf{W}_j^p$ and its orthogonal basis $B_j^p = \{\psi_j^p(t - 2^j n)\}_{n\in\mathbf{Z}}$ at node (j, p). The two successor wavelet packet orthogonal bases at the children nodes are defined by the splitting relations (Coifman *et al.*; 1992; Mallat, 1999):

$$\psi_{j+1}^{2p} = \sum_{n=-\infty}^{+\infty} h[n]\psi_j^p(t - 2^j n) \qquad (4)$$

$$\psi_{j+1}^{2p+1} = \sum_{n=-\infty}^{+\infty} g[n]\psi_j^p(t - 2^j n) \qquad (5)$$

One may prove that $B_{j+1}^{2p} = \{\psi_{j+1}^{2p}(t - 2^{j+1} n)\}_{n\in\mathbf{Z}}$ and $B_{j+1}^{2p+1} = \{\psi_{j+1}^{2p+1}(t - 2^{j+1} n)\}_{n\in\mathbf{Z}}$ are orthonormal bases of two orthogonal spaces $\mathbf{W}_{j+1}^{2p}$ and $\mathbf{W}_{j+1}^{2p+1}$ such that

$$\mathbf{W}_{j+1}^{2p} \oplus \mathbf{W}_{j+1}^{2p+1} = \mathbf{W}_j^p \qquad (6)$$

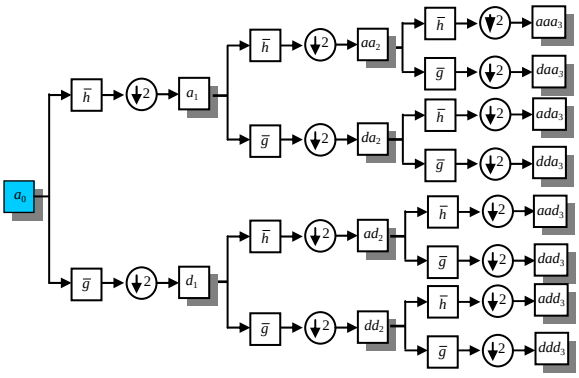

Figure 3. Fast wavelet packet decomposition

2.4.2 Best Basis Feature Extraction: For a given orthogonal wavelet function, one may generate a large family of orthogonal bases that include different types of time-frequency atoms. The basis family is always interpreted as a dictionary D that is a union of orthonormal bases in a signal space of finite dimension N:

$$D = \bigcup_{\lambda \in \Lambda} B^{\lambda} \tag{7}$$

Wavelet packet is an example of dictionary where the bases share some common vectors. Each orthonormal basis in the dictionary is a family of N wavelet functions: $B^{\lambda} = \{\psi_m^{\lambda}\}_{1 \le m \le N}$ and offers a particular way of coding signals, preserving global energy, and reconstructing exact features. For discrete signals of size N, the number of wavelet packet bases is more than $2^{N/2}$ (Mallat, 1999). In order to optimize the non-linear feature extraction of a given hyperspectral signal x, one may adaptively choose the "best" basis in the dictionary D depending on the spectral structures. Then the features are selected from the M largest wavelet coefficients calculated by this best basis. This can be done by the "fast best basis algorithm" proposed by Coifman and Wickerhauser (1992). This algorithm first expands a given signal x into a family of orthonormal bases such as the wavelet packets. Then a complete basis called a best basis which minimizes a certain cost functional $C(x, B^{\lambda})$ is searched among the binary tree with a bottom-up progression. The best basis A_j^p at each subspace $\mathbf{W}_j^p$ is determined by minimizing the cost function C:

$$A_j^p = \begin{cases} B_j^p & \text{if } C(x, B_j^p) \le C(x, A_{j+1}^{2p}) + C(x, A_{j+1}^{2p+1}) \\ A_{j+1}^{2p} \cup A_{j+1}^{2p+1} & \text{if } C(x, B_j^p) > C(x, A_{j+1}^{2p}) + C(x, A_{j+1}^{2p+1}) \end{cases} \tag{8}$$

The cost function C should be defined by the Schur concave sum and with the additive property for efficient computation. The cost function used in this study is entropy of the energy distribution of the hyperspectral curve x for each pixel:

$$C(x, B) = -\sum_{m=1}^{N} \frac{\left|\langle x, \psi_m \rangle\right|^2}{\|x\|^2} \log_e \left(\frac{\left|\langle x, \psi_m \rangle\right|^2}{\|x\|^2} \right) \tag{9}$$

Because of the advantage of the tree structure of wavelet packets, the fast dynamic programming algorithm finds the best basis with $O(N \log_2 N)$ operations (Mallat, 1999).

2.5 Local Discriminant Basis Feature Extraction

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. The Local Discriminant Bases (LDB) method was proposed by 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. The discriminating 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 study, J-divergence is used as the discriminant function. Once the discriminant function D is specified, the goodness of each node in the wavelet packet tree can be compared with the two children nodes for a classification problem. According the discriminant measure, we can determine whether we should keep the children nodes or not. This manner is the same as the best basis search algorithm. Because the discriminant measures are calculated within the subspace of wavelet packets, we don't need too much training samples to estimate the discriminant measures.

3. MATCHING PURSUIT FEATURE EXTRACTION

Both the best basis algorithm and LDB method are based on the wavelet packets which divide the frequency axis into intervals of varying sizes. Thus a best wavelet packet basis can be interpreted as a "best" frequency segmentation. If the signal includes different types of high energy structures at different times but in the same frequency interval, such as the case of spectral mixture of hyperspectral data, the wavelet packet basis could not well adapt to the signal. Furthermore, the set of orthogonal bases in the wavelet packet is much smaller then the set of non-orthogonal bases which can be used to improve the approximation of complex signals. The pursuit algorithms generalize the adaptive approximation by selecting the vectors from redundant dictionaries of time-frequency atoms, with no orthogonal constraints.

The Matching Pursuit (MP) introduced by Mallat and Zhung (1993) uses a greedy strategy to find the best basis for signal approximation. Vectors are selected from the dictionary one by one in order to best match the signal structures. It is closely related to projection pursuit algorithm developed by Friedman and Stuetzle (1981) for statistical parameter estimation. In this study, we attempt to use the matching pursuit algorithm to extract the features for hyperspectral image classification. Let $D = \{g_\gamma\}_{\gamma \in \Gamma}$ be a redundant dictionary with $P > N$ vectors, where $\|g_\gamma\| = 1$. A matching pursuit begins by projecting x on a vector $g_{\gamma_0} \in D$ and computing the residue Rx:

$$x = \langle x, g_{\gamma_0} \rangle g_{\gamma_0} + Rx \tag{10}$$

In order to minimize the residue Rx, g_{γ_0} is chosen to maximize $\left|\left\langle x, g_{\gamma_0}\right\rangle\right|$ such that

$$g_{\gamma_0} = \arg\max_{g_\gamma \in \Gamma} \left|\left\langle x, g_\gamma \right\rangle\right| \tag{11}$$

In each of the consecutive steps, the vector $g_{\gamma_m} \in D$ is matched to the residual $R^m x$, which is the m^{th} order residue left after subtracting results of previous iterations:

$$R^m x = \left\langle R^m x, g_{\gamma_m} \right\rangle g_{\gamma_m} + R^{m+1} x \tag{12}$$

Summing (12) from m between 0 and M-1 yields

$$x = \sum_{m=0}^{M-1} \left\langle R^m x, g_{\gamma_m} \right\rangle g_{\gamma_m} + R^M x \tag{13}$$

The orthogonality of $R^{m+1}x$ and g_{γ_m} in each iteration implies energy conservation.

$$\|x\|^2 = \sum_{m=0}^{M-1} \left|\left\langle R^m x, g_{\gamma_m} \right\rangle\right|^2 + \left\|R^M x\right\|^2 \tag{14}$$

One may prove that the residue $\left\|R^m x\right\|$ will converge exponentially to 0 when m tends to infinity (Mallat, 1999). A matching pursuit can be implemented with a fast algorithm that $\left\langle R^{m+1} x, g_\gamma \right\rangle$ is calculated from $\left\langle R^m x, g_\gamma \right\rangle$:

$$\left\langle R^{m+1} x, g_\gamma \right\rangle = \left\langle R^m x, g_\gamma \right\rangle - \left\langle R^m x, g_{\gamma_m} \right\rangle \left\langle g_{\gamma_m}, g_\gamma \right\rangle \tag{15}$$

Finally, the M vectors $\{g_{\gamma_m}\}_{0<m\le M}$ chosen to minimize the residues at each iteration are directly used as the features for hyperspectral image classification.

4. EXPERIMENTS

The main purpose of this experiment is to compare the wavelet-based feature extraction methods described in this paper in terms of classification accuracy. Figure 4 shows the AVIRIS dataset tested in this experiment which is available in the AVIRIS website of NASA JPL (http://popo.jpl.nasa.gov/html/aviris. freedata.html). This image was acquired in 1996 and covered the Jasper Ridge Biological Preserve of Stanford University. Figure 5 shows the vegetation map corresponding to the test field. The image size of the test field is 180×200. The radiance spectra have been corrected to surface reflectance. From the original 224 spectral channels, 98 spectral bands corresponding to the visible and near-infrared regions are used in this test, discarding the atmospheric absorption bands and short-wave infrared region. There are 10 known classes of different vegetation type and one class of water in this test field. The mean spectra of each class are shown in Figure 6. The results of various feature extraction method are fed to the Maximum Likelihood classifier (MLC) to test the classification effectiveness of the extracted features.

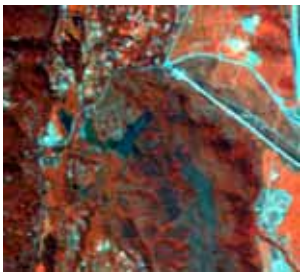

Figure 4. An AVIRIS data set of Jasper Ridge Biological Preserve

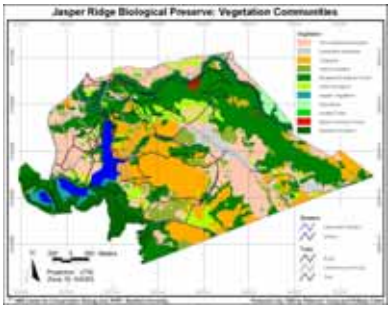

Figure 5. Jasper Ridge Vegetation Map (© JRPB, Copyright 1996, Stanford University)

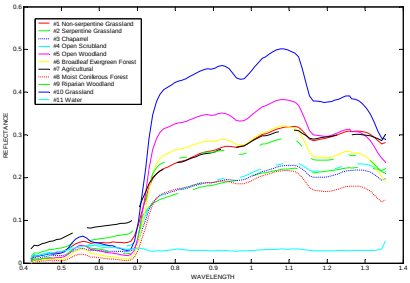

Figure 6. The mean spectra of 11 different classes

Figure7 illustrates the classification results of the extracted features using the wavelet-based and matching pursuit feature extraction methods. In this experiment, the classification accuracies are calculated for various numbers of features extracted by different wavelet-based methods. We summarized the results of this experiment in the following. Firstly, as the number of features increase, the accuracies of classification increase in the beginning and then decrease. The results conform to the Hughes phenomenon. Secondly, the results of nonlinear wavelet-based methods including the matching pursuit, best basis algorithm and LDB methods have the similar results which are better than the results of linear WFE and PCT methods. The basic concept of linear WFE is similar to the PCT methods. They are based on the same criterion that the best approximation with the minimum error is used as a set of important features. The experiment results show that the features extracted by these two methods almost have the identical effectiveness. Thirdly, when the number of features is

smaller than 10, the LDB methods which take into account the discriminant information from the training data have better results. Furthermore, the matching pursuit method has the best accuracies when the number of features is larger than 20. One may notice that the best classification accuracy of some nonlinear wavelet-based methods is occurred when the number of feature is 10. This corresponded with the conclusion of ideal features that the L-1 features are the smallest set needed to classify L classes where L = 11 in this experiment (FuKunaga, 1990).

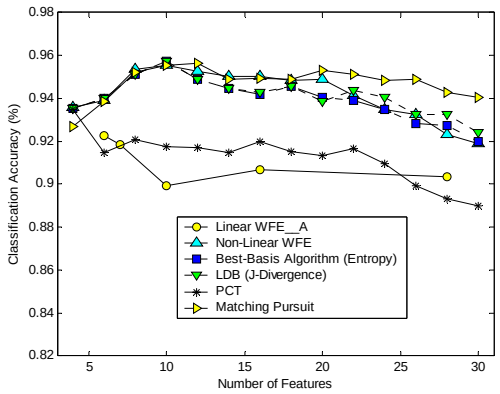

Figure 7. Classification results using wavelet-based feature extraction methods

5. CONCLUSIONS

In this study, several feature extraction methods based on WT and matching pursuit algorithm are used to reduce the dimensionality of hyperspectral data. The experiment results show that the WT is exactly an effective tool for feature extraction. Although some wavelet-based methods such as the nonlinear WFE, best basis algorithm and matching pursuit are based on the best approximation for data representation, they are still effective for classification. Especially, the nonlinear wavelet-based methods are more effective for classification than linear methods. In some circumstances, the matching pursuit basis has better results than the best wavelet packet basis. In the LDB methods, the resulted features are selected within the subspace of wavelet packets, thus the problem of limited training sample size is avoided. In the future, the matching pursuit methods based on the discriminant information between different classes derived from the training data set will be studied for feature extraction. Furthermore, because the results of wavelet-based feature extraction methods are strongly depend on the choice of wavelet basis, the classification accuracies of wavelet-based features using different wavelets function will be tested in the future.

6. REFERENCES

Bellman, R., 1961. *Adaptive Control Processes: A Guided Tour*, Princeton University Press.

Benediktsson, J. A., J. R. Sveinsson, and K. Arnason, 1995. Classification and feature extraction of AVIRIS data. *IEEE Trans. Geoscience and Remote Sensing*, 33(5), pp. 1194-1205.

Coifman, R. R., Y. Meyer, and M. V. Wickerhauser, 1992. Wavelet analysis and signal processing. In: *Wavelets and Their Applications*, Jones and Barlett, Boston.

Fukunaga, K., 1990. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego.

Firedman, J. D. and W. Stuetzle, 1981. Projection pursuit regression. *J. of Amer. Stat. Assoc.*, 76, pp. 817-823.

Goetz, A. F. H., G. Vane, J. E. Solomon, and B. N. Rock, 1985. Imaging Spectrometry for earth remote sensing. *Science*, 228(4704), pp. 1147-1153.

Hughes, G. F., 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Information Theory*, IT-14(1), pp. 55-63.

Hsu, P. H., Y. H. Tseng and P. Gong, 2002. Dimension reduction of hyperspectral images for classification applications. *Geographic Information Sciences*, 8(1), pp. 1-8.

Hsu, P.H., 2003. *Spectral Feature Extraction of Hyperspectral Images Using Wavelet Transform*. Ph.D. Thesis, Department of Surveying Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.

Jain, A. K., R. P. W. Duin, and J. Mao, 2000. Statistical pattern recognition: A Review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(1), pp. 4-37.

Jimenez, L. O. and D. A. Landgrebe, 1995. *High Dimensional Feature Reduction Via Projection Pursuit*. Ph.D. thesis, School of Electrical & Computer Engineering, Purdue University, West Lafayette.

Landgrebe, D. A., 2001. Analysis of multispectral and hyperspectral image data. In: *Introduction to Modern Photogrammetry*, John Wiley & Sons, Inc.

Lee, C. and D. A. Landgrebe, 1993. Analyzing high-dimensional multisepctral data. *IEEE Trans. Geoscience and Remote Sensing*, 31(4), pp. 792-800.

Lillesand, T. M. and R. W. Kiefer, 2000. *Remote Sensing and Image Interpretation*. John Wiley & Sons, New York.

Mallat, S., 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11(7), pp. 674-693.

Mallat, S., Z. Zhang, 1993. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Processing*, 41(12), pp. 3397-3415.

Mallat, S., 1999. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego.

Pittner, S. and S. V. Kamarthi, 1999. Feature extraction from wavelet coefficients for pattern recognition tasks. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(1), pp. 83-88.

Saito, N. and R. R. Coifman, 1994. Local discriminant bases. In: *Proceeding of SPIE*.

Tadjudin, S. and D. A. Landgrebe, 1998. *Classification of high dimensional data with limited training samples*. Ph.D. thesis, School of Electrical & Computer Engineering, Purdue University, West Lafayette.