EVALUATING THE INITIALIZATION METHODS OF WAVELET NETWORKS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Pai-Hui Hsu *

Dept. of Civil Engineering, National Taiwan University, No.1, Sec. 4, Roosevelt Rd., Taipei City 10617, Taiwan – hsuph@ntu.edu.tw

Commission VII, WG VII/3

KEY WORDS: Hyperspectral, Classification, Wavelet, Neural Networks, Wavelet Networks, Initialization

ABSTRACT:

The idea of using artificial neural network has been proven useful for hyperspectral image classification. However, the high dimensionality of hyperspectral images usually leads to the failure of constructing an effective neural network classifier. To improve the performance of neural network classifier, wavelet-based feature extraction algorithms can be applied to extract useful features for hyperspectral image classification. However, the extracted features with fixed position and dilation parameters of the wavelets provide insufficient characteristics of spectrum. In this study, wavelet networks which integrates the advantages of wavelet-based feature extraction and neural networks classification is proposed for hyperspectral image classification. Wavelet networks is a kind of feed-forward neural networks using wavelets as activation function. Both the position and the dilation parameters of the wavelets are optimized as well as the weights of the network during the training phase. The value of wavelet networks lies in their capabilities of optimizing network weights and extracting essential features simultaneously for hyperspectral images classification. In this study, the influence of the learning rate and momentum term during the network training phase is presented, and several initialization modes of wavelet networks were used to test the performance of wavelet networks.

1. INTRODUCTION

Imaging spectrometer, a remote sensing technology which was developed in 1980's, can obtain images with hundreds of spectral bands simultaneously (Goetz *et al.*, 1985). The images acquired with spectrometers are called as hyperspectral images. These images not only reveal two-dimensional spatial information but also contain rich and fine spectral information. With these characteristics, they can be used to identify surface objects and improve land use/cover classification accuracies. In past three decades, hyperspectral images have been widely used in different fields such as mineral identification, forest vegetation mapping, and disaster investigation based on the spectral analysis (Lillesand and Kiffer, 2000).

Since hyperspectral images contain rich and fine spectral information, an improvement of land use/cover classification accuracy is highly expected from the utilization of such images. However, the statistics-based classification methods which have been successfully applied to multispectral images are not as effective as to hyperspectral images. In fact, problems will arise if too many spectral bands are simultaneously taken on finite training samples. If the training samples are insufficient for the needs, which is a very common case in the hyperspectral images, 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.

A simple but sometimes very effective way of dealing with high-dimensional data is to reduce the dimensionality deliberately (Lee and Landgrebe, 1993; Benediktsson *et al.*, 1995; Hsu, 2007a). In the case of hyperspectral images, this can be done by feature extraction that a small number of salient features are extracted from the hyperspectral images before classification. Since the dimensionality is reduced before classification, the curse of dimensionality can be avoided. Thus most of the traditional classification methods such as the maximum likelihood classifier (MLC) can be directly applied to the extracted features after the feature extraction

In order to avoid the problems caused by the limited training samples, several feature extraction methods based on the wavelet transform (WT) have been proposed for hyperspectral image classification (Hsu, 2003; Hsu, 2007a). In the past decades, 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 translated and scaled versions of the mother wavelet function. When WT is applied to the hyperspectral images, the local energy variations of a spectral signature in different spectral bands at each scale (or frequency) can be detected automatically and provide useful information for hyperspectral image classification. Although the proposed wavelet-based methods perform well for feature

* Corresponding author

extraction and also effectively for classification, however, the relationship between the extracted features and the identified classes are not apparent.
In addition to the dimensionality reduction for the statistics-based classifier, nonparametric classifiers such as the artificial neural networks (ANNs) are also proposed to deal with the problem of high dimensionality and also have been applied to hyperspectral image classification. The use of ANNs is motivated by their power in pattern recognition and classification due to the ultimately fine distribution and non-linearity of the process. However, most of the neural processing algorithms are computationally intensive and involve many iterative calculations, especially for hyperspectral images. A characteristic of neural networks is that the networks need a long training time but are relatively fast data classifiers. For very-high-dimensional data, the training time of a neural network can be very long and the resulting neural network can be very complex. This is a serious drawback, especially when the dimensionality and the sample size of training data are large (Benediktsson *et al*., 1995).

To combine the advantages of ANN's with wavelet-based feature extraction methods, the wavelet networks (WNs) have been proposed with some success in data approximation, identification and classification (Dickhaus and Heinrich, 1996). The value of wavelet networks lies in their capabilities of extracting essential features in time-frequency plane. Furthermore, both the position and the dilation of the wavelets are optimized besides the weights of the network during the training phase. This hierarchical, multiresolution training can result in a more meaningful interpretation of the resulting mapping and adaptation of networks that are more efficient compared to conventional methods. In addition, the wavelet theory provides useful guidelines for the construction and initialization of networks and, consequently, the training times are significantly reduced (Iyengar, 2002).

The performance of the WNs for hyperspectral image classification has been test in Hsu (2007b). The experiment results showed that the WNs exactly is an effective tool for classification of hyperspectral images, and has better results than the traditional feed-forward multi-layer neural networks. The structure of the WNs used in this study is a kind of feed-forward neural network, and the ordinary back-propagation (BP) is used for training WNs. Therefore, the drawbacks of BP may exist in WNs. The first problem is the local minimum of the loss function caused by the gradient descent algorithms (Postalcioglu & Becerikli, 2007). In BP, the learning rate is usually used to control the size of weight changes during the learning phase. Finding a reasonable learning rate of wavelet networks is important to not only curtail processing cost but also classify accurately. A simple gradient decent procedure steps toward the minimum very slowly, and an oscillatory descent occurs with a higher learning rate. In general back-propagation learning process, a simple but powerful improvement algorithm is to add a momentum term (Plaut et al., 1986) to the gradient decent formula. The use of momentum adds inertia to the motion through weight space and smoothes out the oscillations (Bishop, 1995). In this study, the influence of the momentum will be tested.

The second problem is the initialization of the network parameters. Efficient initialization will result to less iterations in the training phase of the network and also avoid local minimums of the loss function in the training phase (Alexandridis and Zapranis, 2013). As for the initialization of wavelet networks, weights are typically started with random numbers. For initial wavelon nodes, there are several initial modes have been proposed to solve best regression problems (Zhang, 1997), but they are not convenient for classification. An easier approach is take prior information into account. For example, differences between two classes are expected for higher frequencies, and the wavelet nodes should be initialized in this region of the time-frequency plane.

In this study, the theory of WNs is firstly introduced for hyperspectral image classification, and then an AVIRIS image was used to test the feasibility and performance of classification using the WNs. The influence of the learning rate and momentum term is presented, and several initialization modes of WNs were used to test the performance of wavelet networks.

2. WAVELET TRANSFORM

2.1 Wavelet Transform

Due to the time-frequency localization properties, wavelet transform (WT) has proven to be appropriate starting point for the classification of the measured signals (Stefan and Sagar, 1999). The wavelet transform decomposes a signal into a series of shifted and scaled versions of the mother wavelet function. 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).

Mathematically, a wavelet is defined as a function $\psi \in \mathbb{L}^2(R)$ that has effectively a limited extent and it has an average value of zero. The family of wavelet bases can be produced by scaling s and translating u from the mother wavelet (Mallat, 1999):

$$\psi_{u,s} = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right) \tag{1}$$

The continue wavelet transform (CWT) of $f \in \mathbb{L}^2(R)$ at time u and scale s can obtained by taking the integral inner product of $f(t)$ with the scaled and translated versions of the basis function ψ:

$$Wf(u,s) = \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*\left(\frac{t-u}{s}\right) \tag{2}$$

where $*$ denotes complex conjugation. In definition, the CWT is a convolution of the input data with a set of functions generated by the mother wavelet. The convolution can be computed by using the Fast Fourier Transform (FFT).

The analysis of a signal using CWT yields a wealth of information. Clearly there will be a lot of redundancy in the CWT The discrete wavelet transform (DWT) is an implementation of the wavelet transform using a discrete set of the scales and translations according to some defined rules to reduce the redundancy. The orthogonal wavelet in terms of multi-resolution analysis (MRA) is commonly used in various applications. The DWT 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 outputs. The decomposition formulas are listed as following:

$$a_{j+1}[p] = \sum_{n=-\infty}^{\infty} h[n-2p]\, a_j[p] = a_j * \bar{h}[2p] \qquad (3)$$

$$d_{j+1}[p] = \sum_{n=-\infty}^{\infty} g[n-2p]\, a_j[p] = a_j * \bar{g}[2p] \qquad (4)$$

where $\bar{h}[n] = h[-n]$ and $\bar{g}[n] = g[-n]$ are 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}.

Either CWT or DWT can be used in WNs. In this study, the CWT is used for the purpose of the hyperspectral image classification.

2.2 Wavelet-Based Feature Extraction

It has been shown that wavelet transform provides good capabilities of time-frequency analysis. Hence, several wavelet-based feature extraction (WFE) methods have been proposed to extract essential features of hyperspectral images for classification (Hsu, 2007a). Such features were performed with wavelet transform described by translating and scaling indices. The values of WT at specific time and scale index can be regarded as meaningful features that can be used to distinguish different classes of land objects in image classification. However, the extracted features were highly dependent on both the values of the translating and scaling parameters that characterize pre-processing of WFE. These values are selected before feature extraction procedure based on a user's prior knowledge that is rarely acquired or unpredictable (Angrisani et al., 2001). To overcome this limitation, adjustable translating and scaling parameters dependent on characteristics of data are expected. In this paper, wavelet networks based on back-propagation networks and wavelet theory is introduced. Wavelet networks can adjust translating and scaling parameters during learning stage and give optimized image classification result. Not only avoid the limited sample problem, but also improve the performance of neural networks. This wavelet networks-based classifier for hyperspectral image is more flexible than neural networks because weights and extracted features both are optimized during training. Figure 1 illustrates the difference between wavelet networks and neural networks with extracted features by the WFE.

3. WAVELET NETWORKS

3.1 Structure of Wavelet Networks

Based on the theory of wavelet transform, the concept of wavelet networks was first proposed by Zhang and Benvenite (Zhang and Benveniste, 1992). Wavelet networks for classification combines the aspects of the wavelet transformation for purposes of feature extraction and selection with the characteristic decision capabilities of neural-network approaches (Dickhaus and Heinrich, 1996). Figure 2 shows the structure of the wavelet networks used in this study, in which consists of one wavelet layer, one hidden layer, and one output layer. A wavelet node (called wavelon) for feature extraction is parameterized by a translation parameter, u_k, and a scale parameter, s_k. The outputs of the wavelons, φ_k, which can be interpreted as the correlation between the signal $x[i]$ and the wavelet $h_k(t)$, server as input to the hidden layer of neural network classifier. The classifier of the right part can be any single-layer or multi-layer perceptrons. During the learning process, the wavelet node parameters are also updated to minimize the error, E.

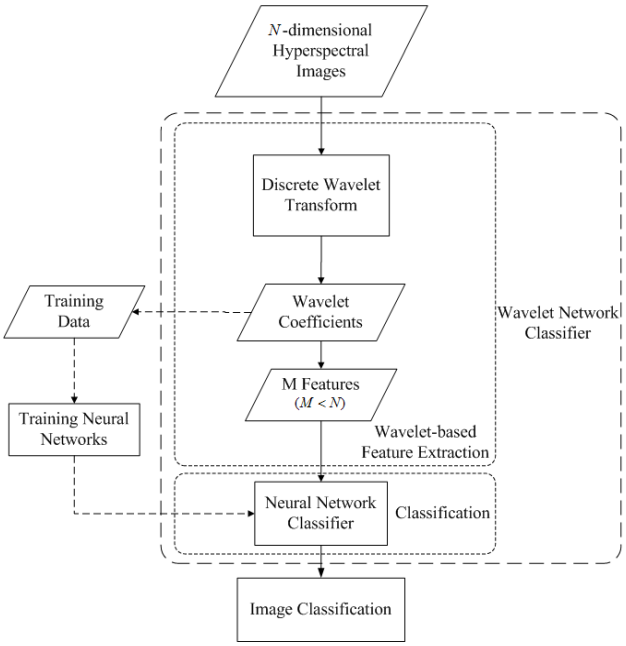


Figure 1. Comparison between neural networks with WFE and wavelet networks

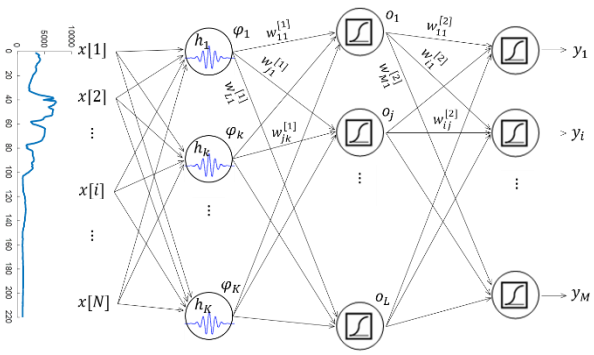


Figure 2. The structure of wavelet networks

3.2 Implementation of Wavelet Networks

A typical wavelet function used in the wavelet networks is the complex Morlet wavelet (Dickhaus and Heinrich, 1996):

$$\mathrm{h}(t) = \exp(j\omega_0 t - 0.5 \cdot t^2) \qquad (5)$$

The wavelet nodes $h_k(t)$ in Figure 2 are scaled and dilated versions of this wavelet mother function:

$$h_k(t) = \frac{1}{\sqrt{s_k}} \exp\left(j\omega_0 \left(\frac{t-u_k}{s_k}\right) - 0.5 \cdot \left(\frac{t-u_k}{s_k}\right)^2\right) \qquad (6)$$

The variable s_k is the scale parameter, and u_k is the dilation parameter of the wavelet function. If the scale s_k is large, the wavelet is a dilated low-frequency function, whereas for small values of s_k, the wavelet is compact, corresponding to a high-frequency function. Formally, the node's output φ_k is the result of the wavelet transform which is defined as the inner product of the node h_k and the signal $x[i]$, which is the input of the wavelet networks (the index $i = 1, \cdots, N$ denotes the signal number)

$$\varphi_k = \langle h_k, x \rangle = \left| \int_t h\left(\frac{t-u_k}{s_k}\right) x(t) dt \right| \tag{7}$$

For the Morlet wavelet, φ_k can be calculated for each wavelet node

$$\varphi_k = \sqrt{o_{cos_k}^2 + o_{sin_k}^2} \tag{8}$$

where

$$o_{cos_k} = \sum_{i=1}^{N} x[n] \cos\left(\omega_0 \frac{n-u_k}{s_k}\right) \cdot \exp\left(-0.5 \cdot \left(\frac{n-u_k}{s_k}\right)^2\right) \tag{9}$$

$$o_{sin_k} = \sum_{i=1}^{N} x[n] \sin\left(\omega_0 \frac{n-u_k}{s_k}\right) \cdot \exp\left(-0.5 \cdot \left(\frac{n-u_k}{s_k}\right)^2\right) \tag{10}$$

The neuron's output y_i is calculated by the weighted sum of the outputs of the pervious layer, o_j, the neuron's threshold, b_j, and its activation function f , that the sigmoidal function is used in this paper:

$$\begin{aligned} y_i = f\left(u_i^{[2]}\right) &= \frac{1}{1+\exp\left(u_i^{[2]}\right)} \\ &= \frac{1}{1+\exp\left(\sum_{j=1}^{J} w_{ij}^{[2]} \cdot o_j + b_i^{[2]}\right)} \end{aligned} \tag{11}$$

3.3 Training of Wavelet Networks

During the training phase, the ANN weights w are adjusted to minimize the total least-square error E_T between the net's desired output vector d_i and its actual output y_i for all input vectors $x^{(p)}$.

$$E_T = \sum_{p=1}^{P} E^{(p)} = \frac{1}{2} \sum_{p=1}^{P} \sum_{i=1}^{M} \left(d_i^{(p)} - y_i^{(p)}\right)^2 \to min \tag{12}$$

The minimization problem can be solved by an iterative gradient technique. The partial derivative of the weights, w, are calculated according to the generalized delta rule:

$$w_{t+1} = w_t - \eta \frac{\partial E}{\partial w_t} + \alpha(w_t - w_{t-1}) \tag{13}$$

$$w = \left\{u_k, s_k, w_{jk}^{[1]}, b_j^{[1]}, w_{ij}^{[2]}, b_i^{[2]}\right\} \tag{14}$$

$$\frac{\partial E_T}{\partial w} = \sum_{p=1}^{P} \frac{\partial E^{(p)}}{\partial w} \tag{15}$$

The partial derivatives of the weights to the neurons in the output and hidden layer are calculated as follows:

$$\frac{\partial E_T}{\partial w_{ij}^{[2]}} = -\sum_{p=1}^{P} \delta_{ij}^{(p)} o_j^{(p)} \tag{16}$$

$$\frac{\partial E_T}{\partial w_{jk}^{[1]}} = -\sum_{p=1}^{P} \delta_{jk}^{(p)} \varphi_k^{(p)} \tag{17}$$

These two equations hold for neurons with a sigmoidal activation function.

In the wavelet network, not only the weights are adjusted, but also the parameters of the wavelet nodes. The partial derivatives for a wavelet node's scale parameter, s_k, and it shift parameter, u_k, depend on the wavelet basis chosen and are determined using the backpropagated error E_T:

Thus, in each iteration of the training cycle, the weights and the wavelet parameters are varied to reduce the error, *E*. This procedure is repeated until the net has settled down to a minimum

$$\frac{\partial E^{(p)}}{\partial u_k} = -\sum_{j=1}^{L} \delta_{jk} w_{jk}^{[1]} \frac{\partial \varphi_k}{\partial u_k} \tag{18}$$

$$\frac{\partial E^{(p)}}{\partial s_k} = -\sum_{j=1}^{L} \delta_{jk} w_{jk}^{[1]} \frac{\partial \varphi_k}{\partial s_k} \tag{19}$$

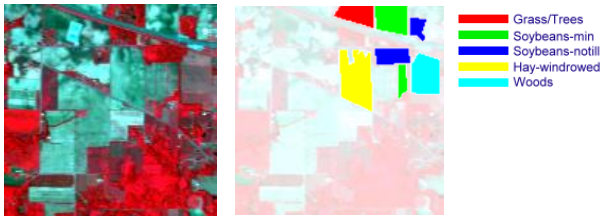
$$\begin{aligned} \frac{\partial \varphi_k}{\partial u_k} &= \frac{1}{\varphi_k} \sum_{i=1}^{N} x[n] \exp\left(-0.5 \cdot \left(\frac{n-u_k}{s_k}\right)^2\right) \frac{1}{s_k} \\ &\cdot \left(\begin{array}{l} o_{cos_k}\left[\omega_0 \sin\left(\omega_0 \frac{t-u_k}{s_k}\right) + \frac{t-u_k}{s_k} \cos\left(\omega_0 \frac{n-u_k}{s_k}\right)\right] \\ +o_{sin_k}\left[-\omega_0 \cos\left(\omega_0 \frac{n-u_k}{s_k}\right) + \frac{n-u_k}{s_k} \sin\left(\frac{t-u_k}{s_k}\right)\right] \end{array} \right) \end{aligned} \tag{20}$$

$$\frac{\partial \varphi_k}{\partial s_k} = \frac{n-u_k}{s_k} \frac{\partial \varphi_k}{\partial u_k} \tag{21}$$

4. EXPERIMENTS

In this study, a set of hyperspectral data was used to test the classification performance using the two methods mentioned above. The diagram of the experiment is illustrated in Figure 1. The study image is a small segment of AVIRIS image. The image is located in NW Indiana Indian Pine test site and was taken in 1991. The image size is 145 pixels×145 pixels. The original data set has 224 spectral bands from 400 nm to 2450 nm with 10 nm spectral resolution. The number of bands is 220 after removing 4 noisy bands. The radiance spectra are directly used to test the feature extraction method without performing any kind of atmospheric correction. The ground truth data includes five different classes which are "grass/trees", "soybeans-min", "soybeans-notill", "hay-windrowed" and "woods". Figure 4 shows the information related to test image. The number of training samples of each known class was 50 from ground truth data and additional labeled test samples were used to assess the accuracy of the classification.

In order to develop a wavelet networks image classifier suitable for hyperspectral image, three experiments were designed to look for a proper wavelet network classifier. The performance of wavelet networks was evaluated generally by the criterion MSE (mean square error).



(a) Test image (b) Ground truth data

Figure 3. Test Image Data

4.1 Experiment 1: Learning Rate and Momentum Term

An experiment was set to test three back-propagation algorithms including general gradient decent algorithm, gradient decent with momentum term, and quickprop method. Momentum term and quickprop method are improved algorithms to accelerate the learning process. The performance (MSE) was used to evaluate the efficiency of learning in Figure 4. An oscillatory curve was obtained by the simple gradient descent method. It indicates that a large initial learning rate was chosen in this case. Compare to the result of added momentum term, the oscillations was smoothed out because of adaptive learning strategy. In addition, the quickprop method lead the most effective learning process in this experiment because of the quick and smoother convergence in Figure 4.

Table 1 shows the classification results acquired by trained networks whose MSE = 0.5. As discussed above, quickprop required the least iteration computation. Further, good generalization ability was provided because the minimum MSE was corresponding to the best classification result in Table 1. These results suggest that quickprop is a practical improved gradient algorithm for wavelet networks.

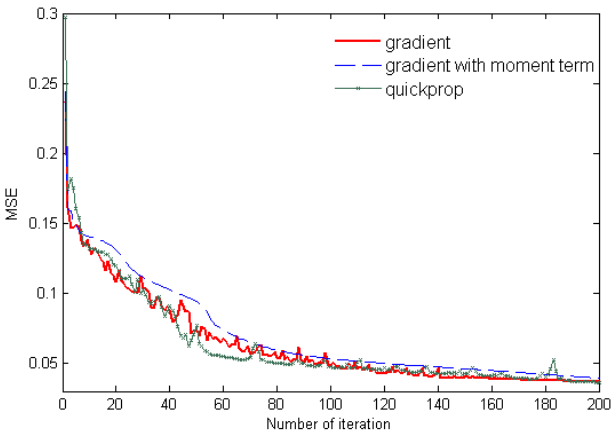


Figure 4. The performance (MSE) of three back-propagation algorithms

	overall accuracy	the number of iterative learning
gradient	87.1%	92
gradient with moment term	88.8%	119
quickprop	89.8%	80

Table 1. Classification result and the number of learning at MSE = 0.5

4.2 Experiment 2: Different Initialization Modes

Several initialization modes were used to test the performance of wavelet networks. As listed in Table 2, better classification results were given by mode1 and mode 5. Initial values set by these two modes were well distributed in all band-scale space. The results indicated that useful features can be found in both high frequency field (detail space) and low frequency field (approximation space). Further, using the wavelet networks with the initial scaling and translating values chosen randomly is a feasible and easy procedure.

Mode	Initialized parameters		Overall accuracy
	translating	scaling	
Mode 1	equal interval	each resolution space	88.1%
Mode 2	equal interval	random	86.3%
Mode 3	equal interval	finest resolution	85.3%
Mode 4	random	each resolution space	86.6%
Mode 5	random	random	87.7%
Mode 6	random	finest resolution	81.7%
Mode 7	non-linear	non-linear	84.6%

Table 2. Initialization mode

Figure 5 represented the locations of the wavelons for all modes on the time-frequency plane after training. It can be found that several common regions were selected as features by different modes. It can be inferred that wavelet networks can extract useful features during training. Moreover, these features were distributed mostly in high frequency domain. Features extracted from wavelet networks initialized by different modes were converged to the same spot of time-frequency plane, especially higher frequency or detail space. In multi-resolution analysis, useful features can be found in detail space to distinguish objects.

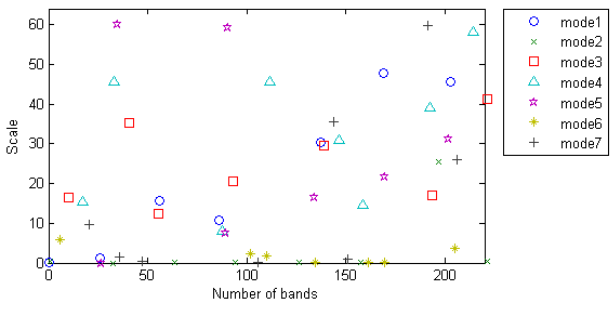


Figure 5. Locations of learned wavelons on time-frequency plane

Then by projecting scaling and translating indices of learned wavelons to time-frequency plane of the tested data, one can examine the efficiency of wavelons, as shown in Figure 6. It can

be seen approximately that features extracted by wavelet networks corresponded to the modulus of coefficients characterizing the five classes.

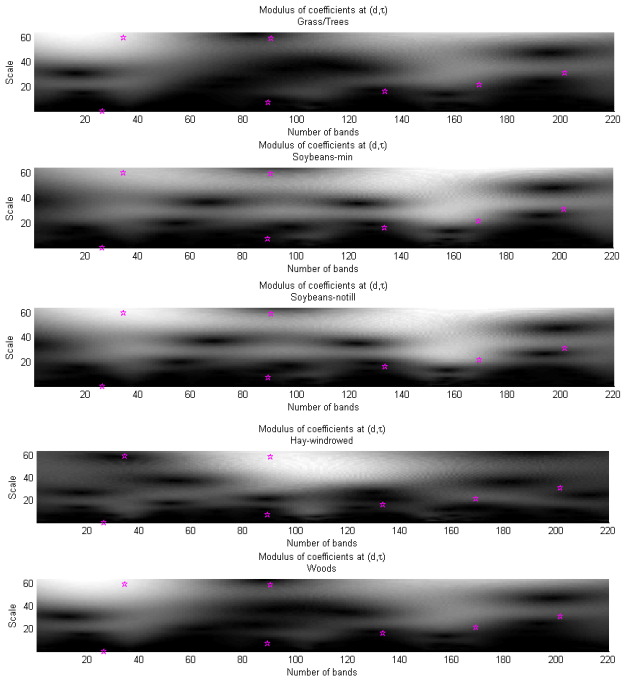


Figure 6. Modulus of coefficients and learned wavelons of mode 5

4.3 Experiment 3: Varying Number of Wavelons

Statistically, the effective number of features can be approximately selected upon the number of classes by $N_{featres} = N_{classes} + 1$. Unnecessary features increase the computation burden. Moreover, the problem of insufficient samples discussed before arises. The issue about how many efficient wavelons should have in wavelet networks is crucial. The experiment results in Figure 7 showed that growing features is not beneficial to accuracy but increase the training time. The appropriate number of wavelons was approximately acquired by training with 5 wavelons. No significant accuracy improvement was obtained by adding wavelons.

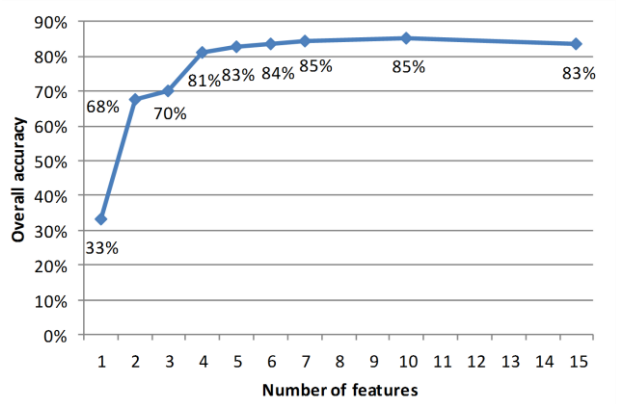


Figure 7. Classification results with increasing the numbers of wavelons (features)

5. CONCLUSION

Supervised image classification requires a certain amount of training samples to obtain reliable statistical parameters. However, we often have limited ground truth data in practical applications. Neural networks is an alternative approach of image classification without necessary statistical computation. However, neural networks increase training cost and complexity when applying to high dimensional data. Wavelet networks can successfully improve the performance of traditional neural networks by combining wavelet decomposition theory and learning ability of neural networks. Moreover, it yields a good generalization ability .Some parameters of wavelet networks are discussed in this paper to design a proper hyperspectral image classifer. By optimizing the parameters of wavelet networks, an effective tool to classify can be expected.

Since the mother wavelet function used in this study was complex Morlet wavelet only, it is expected that future research could examine further the correlation between different mother wavelet functions and performance of wavelet networks. Instead of back-propagation wavelet networks, an additional interesting research might be to consider other wavelet networks, such as radical wavelet neural networks and recurrent wavelet networks. Moreover, a more extensive study about the applicability of other artificial intelligence techniques, such as support vector machine, fuzzy logic and genetic algorithm is another interesting topic.

REFERENCES

Alexandridis, A. K., and Zapranis, A. D., 2013. Wavelet neural networks: A practical guide, *Neural Networks*, 42, pp. 1-27.

Angrisani, L., Daponte, P., and D'Apuzzo, M, 2001. Wavelet network-based detection and classification of transients. *IEEE Transactions on Instrumentation and Measurement*, 50(5), pp. 1425-1435.

Bellman, R. E., 1961. *Adaptive control processing* - a guided tour, Princeton University Press.

Benediktsson, J. A., Sveinsson, J. R., and Amason, K., 1995. Classification and feature extraction of AVIRIS data. *IEEE Transactions on Geoscience and Remote Sensing*, 1194-1205.

Bishop, C. M., 1995. Neural Networks for Pattern Recognition Oxford University Press, New York.

Dickhaus, H., and Heinrich, H., 1996. Classifying biosignals with wavelet networks [a method for noninvasive diagnosis]. *Engineering in Medicine and Biology Magazine*, IEEE, 15(5), pp. 103-111.

Goetz, A. F. H., G. Vane, J. E. Solomon, and B. N. Rock, 1985. Imaging Spectrometry for Earth Remote Sensing. *Science*, 228, pp. 1147-1153.

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

Hsu, P.-H., 2007a. Feature extraction of hyperspectral images using wavelet and matching pursuit. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(2), pp. 78-92.

Hsu, P.-H. and Yang, H.-H., 2007b. Hyperspectral image classification using wavelet networks. 2007 IEEE International Geoscience and Remote Sensing Symposium, Barcelona, Spain, pp. 1767-1770.

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

Iyengar, S. S., Cho, E. C., and Phoha, V. V., 2002. *Foundations of Wavelet Networks and Applications*, Chapman&Hall/CRC, Boca Raton.

Lillesand, T. M., Kiefer, R. W., and Chipman, J. W., 2004. *Remote Sensing and Image Interpretation*. John Wiley & Sons., New York.

Mallat, S., 1999. *A Wavelet Tour on Signal Processing*, Academic Press, New York.

Lee, C., and Landgrebe, D. A., 1993. Feature extraction based on decision boundaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4), pp. 388-400.

Plaut, D. C., Nowlan, S. J., and Hinton, G. E., 1986. Experiments on learning by back propagation. Technical Report CMU-CS-86-126, Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA.

Postalcioglu, S. and Becerikli, Y., 2007. Wavelet networks for nonlinear system modelling. Neural Computing and Applications, 16(4), pp 433-441.

Stefan, P., and Sagar, V. K., 1999. Feature Extraction From Wavelet Coefficients for Pattern Recognition Tasks. *IEEE Computer Society*, pp. 83-88.

Zhang, Q., and Benveniste, A., 1992. Wavelet networks. *IEEE Transactions on Neural Networks*, 3(6), pp.889-898.

Zhang, Q., 1997. Using wavelet network in nonparametric estimation. *IEEE Transactions on Neural Networks*, 8(2), pp. 227-236.