SIGNIFICANCE-WEIGHTED FEATURE EXTRACTION FROM HYPER-DIMENSIONAL DATA AND ITS APPLICATIONS

Sadao Fujimura and Senya Kiyasu

Department of Mathematical Engineering and Information Physics
Graduate School of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113, JAPAN

Commission III

KEY WORDS : Algorithms, Feature Extraction, Sensor, Performance, Classification, Accuracy, Significance Weighted, Hyper Dimensional

ABSTRACT

Extracting significant features is essential for processing and transmission of vast volume of hyper-dimensional data. Conventional ways of extracting features are not always satisfactory for this kind of data in terms of optimality and computation time. Here we present a successive feature extraction method designed for significance-weighted supervised classification. After all the data are orthogonalized and reduced by principal component analysis, a set of appropriate features for prescribed purpose is extracted as linear combinations of the reduced components. We applied this method to 411 dimensional hyperspectral data obtained by a ground-based imaging spectrometer. The data were obtained from tree leaves of five categories, soil, stone and concrete. Features were successively extracted, and they were found to yield more than several percents higher accuracy for the classification of prescribed classes than a conventional method. We applied the results of feature extraction for evaluating the performance of current sensors. We used the accuracy of classification as an index of performance for a specific purpose.

1. INTRODUCTION

Recently the dimension of remotely sensed data becomes higher and higher because of higher spectral resolution, increasing number of sensors, and multi-temporal observations. Airborne Visible Infrared Imaging Spectrometer (AVIRIS), for example, has 224 spectral bands in the 0.4-2.5μm region (Vane, 1988). In order to efficiently obtain necessary information from these hyper-dimensional data, or in order to transmit the data through a communication channel, the quantity of data must be reduced. This can be achieved by extracting significant features.

Here we propose a feature extraction method designed for significance-weighted supervised classification and present its application for evaluating the performance of sensors. The basic idea of our feature extraction is as follows: in classification of data we have some kind of objectives or intention. This means that in most of the cases we are interested in classification of a particular set of classes, not all of the terrain objects included in the image. Thus we introduce subjective significance explicitly into feature extraction. The evaluation to be used is the accuracy for the particular classes, though conventional feature extraction methods considering only the average accuracy for all the classes in the image. The purpose of our feature extraction is to extract a set of features which optimally separate one class from another among a particular set of important classes.

One of the conventional methods of feature extraction utilizes an exhaustive search for the best subset of sensor channels using a separability measure between classes (Swain, 1978), and another uses principal component analysis (Ready, 1973). The former requires a lot of computation time to evaluate all the combinations of channels. The latter is not optimal because the features are not selected from the viewpoint of discrimination. Canonical analysis can also be used (Schowengerdt, 1983) and gives better results than principal components. It extracts the features which give the best average separability among classes. However, they are not always suitable for significance-weighted classification.

Our method extracts appropriate features as linear combinations of orthogonalized and reduced components which are obtained by principal component analysis. Each feature is determined successively by considering the distance from the significant classes until the distance satisfies a condition.

We applied the results of feature extraction for evaluating the current sensors' efficiency for a specific purpose. The performance of sensors can be evaluated by comparing the classification accuracy with that by the extracted features.

2. PRINCIPLE OF FEATURE EXTRACTION

2.1 Description of Data

First of all and as usual, we assume that we can get training data for almost all the classes in an image to derive feature with: that is, we can estimate the characteristics of most classes included in the image.

We denote hyper-dimensional data (N dimension) by a vector $\mathbf{y} = (y_1, \cdots, y_N)'$ ($'$: transpose), and suppose that they are classified into one of, say, n classes. Then, $\mathbf{y}$ can be decomposed into class mean $\mathbf{y}_a$ and within-class dispersion $\mathbf{y}_e$: that is, $\mathbf{y}$ is written as

$$\mathbf{y}_{ij} = \mathbf{y}_{a_i} + \mathbf{y}_{e_{ij}} \qquad (1)$$
$$(i = 1 \sim n, \; j = 1 \sim m_i),$$

(see Fig. 1), where $\mathbf{y}_{ij}$ is j-th data of class i. We write the covariance matrix of $\mathbf{y}$, $\mathbf{y}_a$ and $\mathbf{y}_e$ as $\mathbf{C}_{yy}$, $\mathbf{C}_a$ and $\mathbf{C}_e$ respectively. We call $\mathbf{C}_a$ and $\mathbf{C}_e$ between-class and within-class covariance matrix, respectively. Here, we assume that the covariance matrix of each class is identical. This assumption is rather reasonable from the view point of the generality of training data (Fujimura, 1981).

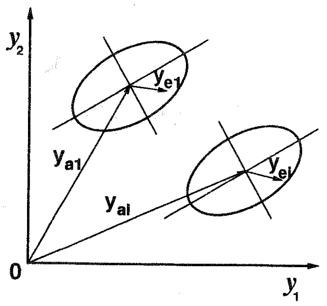


Fig. 1 Description of data

2.2 Feature Extraction

Here, for simplicity we consider two cases where one and two most important classes should be discriminated from all the other classes.

In general, classification accuracy increases as the separability* of classes increases. We use separability to evaluate the performance of features extracted. We extract the features which maximize the separability of a particular pair of classes that we wish to discriminate.

Our method proposed here consists of two steps of processing: pre-processing and feature extraction.

In the pre-processing, hyper-dimensional data $\mathbf{y} = (y_1, \cdots, y_N)'$ are reduced and normalized to m ($m \ll N$) components $\mathbf{z} = (z_1, \cdots, z_m)'$ by a linear transformation $\mathbf{z} = \mathbf{A}'\mathbf{y}$. From the assumption on $\mathbf{C}_e$, the within-class dispersion of each class in the original space has the same ellipsoidal shape shown in Fig. 1. After transformation, they are normalized into an m dimensional sphere. This makes the space uniform: this means that the distance measured in terms of variance does not have directionality in the space.

In the second step, features are successively extracted (Kiyasu,1993, Fujimura,1994) until there remains no class which has distance from the particular classes less than the minimum distance obtained so far. Feature extraction is done by determining sub-space in the feature space: that is, by making a linear combination of $\mathbf{z}$ as $\mathbf{a}'\,\mathbf{z}$, where $\mathbf{a}$ is an m dimensional weight vector which we call here feature vector. Thus, feature extraction is no other than the determination of a feature vector. As the space is uniform now, the direction of an optimal feature vector which discriminates between two classes is obtained just by connecting the centers of these classes. The feature vectors obtained are orthogonalized to make independent.

The procedures for determining successive feature vectors is as follows:

(1) First, we set an optimal feature vector $\mathbf{a}_1$ between the two nearest classes among the prescribed classes.
(2) Next, we evaluate the separability on $\mathbf{a}_1$ for all the combination of the prescribed classes.
(3) If there is any pair of prescribed classes which does not have enough separability, we set an additional feature vector $\mathbf{a}_2$ between them. We ortho-normalize the new vector $\mathbf{a}_2$ with $\mathbf{a}_1$ as shown in Fig. 2, so that this feature is independent of the first one.
(4) Features are successively extracted in the same way until all the distance among the prescribed classes are larger than the minimum distance obtained so far.
(5) Then, we apply the procedures (2)$\sim$(4) to the distance among the prescribed and the other classes.

When only one class is prescribed, the procedure starts from setting a feature vector between the class and its nearest class in the feature space.

A feature $\mathbf{a}_i'\,\mathbf{z}$ is equivalent to $(\mathbf{A}\,\mathbf{a}_i)'\,\mathbf{y}$ expression using original data $\mathbf{y}$, because $\mathbf{z} = \mathbf{A}'\mathbf{y}$, where $(\mathbf{A}\,\mathbf{a}_i)$ means the weighting factor for spectral data.

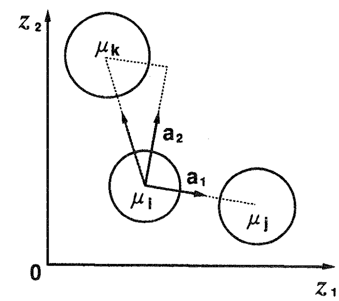


Fig. 2 Feature vectors discriminating between two classes

3. EXPERIMENTAL RESULTS OF FEATURE EXTRACTION

We acquired data for five growth-states of tree leaves (A$\sim$E: from young to fallen), soil, stone and concrete by using an imaging spectrometer which we developed. We obtained 411 dimensional data from the sensor and used for the experiments. For estimating the mean and the variance of each class, 45 training data were used for each class.

*We used the divergence (Kullback, 1959) as a measures of separability. We call it as distance in the rest of this paper.

Averaged relative reflectance is shown in Fig. 3. In the following, the covariance matrices of the classes are assumed to be identical.

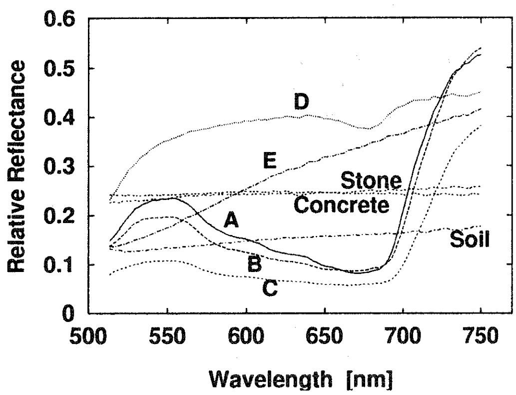


Fig. 3 Spectral reflectance of objects
(A~E : Leaves of plant)

After reducing and normalizing the data to 7 orthogonal components, features were extracted from one to another. At first we selected class A as the most important class to be classified. The first feature was set between class A and the nearest class from A (B in Fig. 3). The next feature was set between A and the next nearest class D. There remains no other classes whose distance from A is less than that between A-B. The two features characterize the weighting factors are shown in terms of wavelength in Fig. 4.

In this case the distance of each class from class A is shown in Table 1: (a) in the original 7 dimensional space, (b) 1 dimension (the first feature), and (c) two dimensional space made by the first two features. From (c), it is seen that the minimum distance is that between A and B which was already obtained in (a). Thus, the two features are sufficient for this case.

To confirm the validity of this method when compared with canonical analysis, the classification accuracy was estimated by test data of 196 samples for each class. Each sample was classified by a maximum likelihood method. Figure 5 shows the classification accuracy for the class A in terms of the number of features.

Table 1 Distance from class A (Relative distance)
(A~E: Leaves, F: Soil, G: Stone, H: Concrete)

(a) Distance in 7 dimension

	B	C	D	E	F	G	H
A	4.4	9.7	18.3	15.5	16.1	16.0	17.1

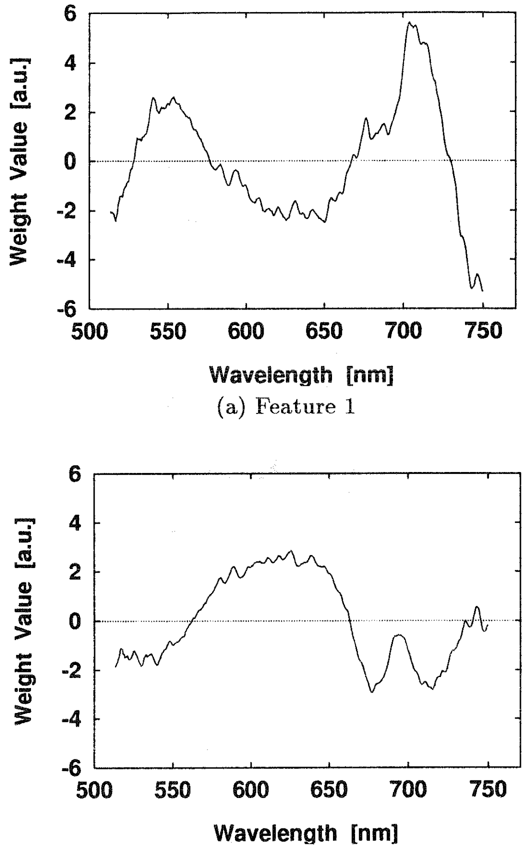
(b) Distance in 1 dimension

	B	C	D	E	F	G	H
A	4.4	7.4	1.6	3.4	2.7	1.9	2.5

(c) Distance in 2 dimension

	B	C	D	E	F	G	H
A	4.4	9.5	18.3	12.5	15.0	15.2	16.3

The accuracy depends on the number of features used, and is higher than that by canonical analysis about 19% (one feature) and 8% (2 features). The confusion matrix is shown in Table 2.



(a) Feature 1

(b) Feature 2

Fig. 4 Weighting factors for the significant class A

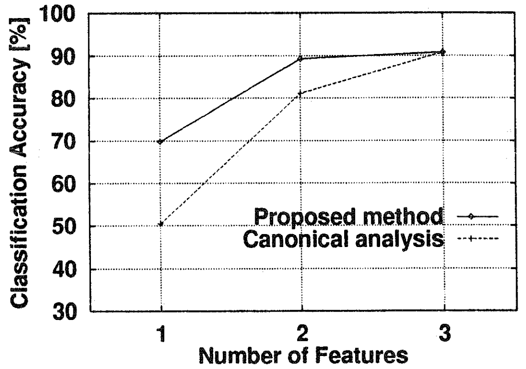


Fig. 5 Classification accuracy of calss A versus number of features

Table 2 Confusion matrix
(A~E: Leaves, F: Soil, G: Stone, H: Concrete)

(a) Classification using first feature

(%)

	A	B	C	D	E	F	G	H
A	69.9	0	0	11.2	3.6	3.1	9.2	3.1
B	0	46.4	25.5	0	19.4	8.7	0	0
C	0	1.5	98.5	0	0	0	0	0
D	27.0	2.6	0	4.6	20.4	27.0	13.3	5.1
E	0	1.0	0	0	38.8	42.9	2.6	14.8
F	0	1.0	0	0	16.8	49.5	7.1	25.5
G	1.0	0	0	6.6	0.5	32.1	34.7	25.0
H	0	0	0	1.0	16.3	52.0	14.3	16.3

(b) Classification using first two features

(%)

	A	B	C	D	E	F	G	H
A	89.3	10.7	0	0	0	0	0	0
B	0	97.5	2.6	0	0	0	0	0
C	0	0	100	0	0	0	0	0
D	0	0	0	90.3	3.1	1.0	4.1	1.5
E	0	0	0	1.0	78.1	17.9	0	3.1
F	0	0	0	0	0	93.4	6.6	0
G	0	0	0	3.1	0	30.6	46.4	19.9
H	0	0	0	23.0	0	17.9	8.2	51.0

Next, we applied our method to the case where two most significant classes A and B are appointed. The distance of each class from class A and B is shown in Table 3 (a). In this case we set the first feature between A and B. The table of distance in this feature space is (b). In the same way above we successively set two features. The distances in 2 dimension are shown in (c).

Table 3 Distance from A and B (Relative distance)
(A~E: Leaves, F: Soil, G: Stone, H: Concrete)

(a) Distance in 7 dimension

	A	B	C	D	E	F	G	H
A	—	4.4	9.7	18.3	15.5	16.1	16.0	17.1
B	4.4	—	7.0	18.5	15.2	16.0	16.1	17.1

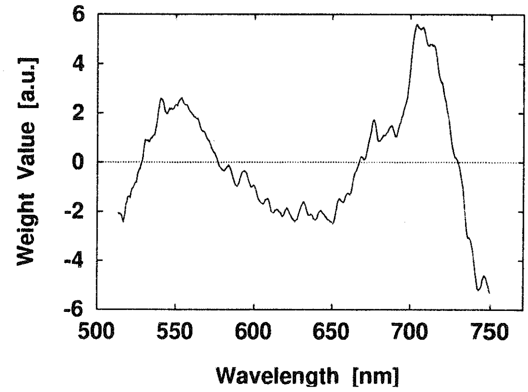
(b) Distance in 1 dimension

	A	B	C	D	E	F	G	H
A	—	4.4	7.4	1.6	3.4	2.7	1.9	2.5
B	4.4	—	3.0	2.8	0.9	1.7	2.5	1.9

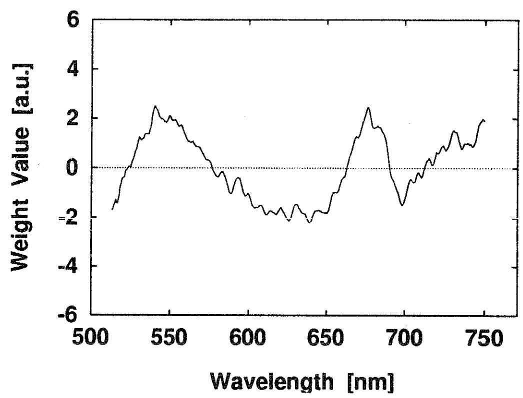
(c) Distance in 2 dimension

	A	B	C	D	E	F	G	H
A	—	4.4	9.2	14.5	15.5	15.0	14.0	14.6
B	4.4	—	6.2	14.7	15.2	14.8	14.1	14.5

It is seen that the two features are sufficient. The features are shown in the form of the weighting factors in Fig. 6. The classification accuracy for A and B is about 16% (one feature) and 6% (2 features) higher than that by canonical analysis (Fig. 7).



(a) Feature 1



(b) Feature 2

Fig. 6 Weighting factors for the significant classes A and B

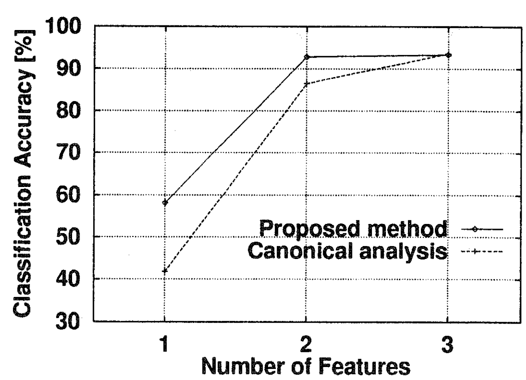


Fig. 7 Classification accuracy of calsses A and B versus number of features

4. APPLICATION FOR SENSOR EVALUATION

We applied the result of feature extraction to the evaluation of performance of sensors. The performance of a sensor is evaluated by classification accuracy of particular classes which are defined to be significant for a specific purpose.

Hyper-dimensional data acquired are separated into two groups, one is used for feature extraction and the other for estimating the classification accuracy. In the first step, significance-weighted features are extracted using the first half of the data. In the next step, the other half of the data are classified by using the extracted features and by using the spectral bands of a sensor under consideration. We use the accuracy of classification as an index of performance for a specific purpose. If the accuracy by a sensor is as high as the accuracy by the extracted features, the sensor can be considered to have sufficient performance.

Our spectrometer for experiments does not cover all the spectral region of current sensors. In order to confirm the validity of this method, we applied it to the subset of bands from the Coastal Zone Color Scanner (CZCS). Though the CZCS has six spectral bands, only three of them (bands 2, 3 and 4 in Fig. 8) are covered by our spectrometer. Figure 9 shows the classification accuracy for classes A and B, by the three bands of CZCS, and by the extracted features. The former was lower than the latter by about 6% when the number of bands is three. We know that the performance of the bands 2, 3 and 4 of CZCS is not sufficient for classifying the classes A and B.

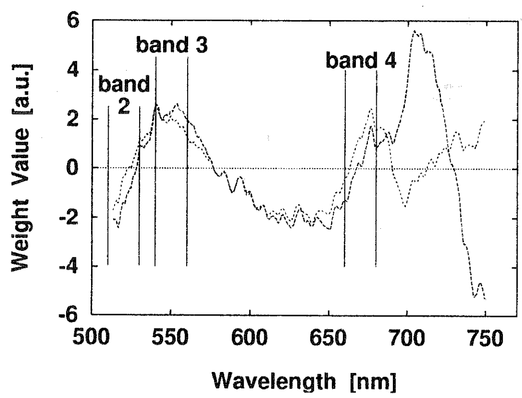


Fig. 8 Three bands used for experiments (bands 2, 3 and 4 of CZCS)

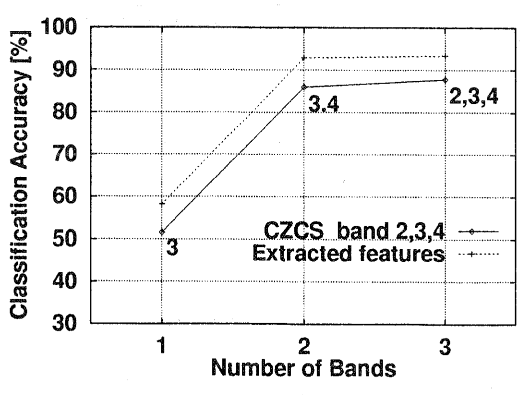


Fig. 9 Classification accuracy by bands 2, 3 and 4 of CZCS and by extracted features

5. CONCLUSIONS

We have proposed a feature extraction method for significance-weighted classification of hyper-dimensional data. The method was tested using 411 dimensional hyperspectral data, in which one or two significant classes were appointed. By successively extracting features, a sufficient number of features to classify the prescribed classes were extracted. It was found that classification accuracy of particular classes increased by more than several percents, compared with classification using the features extracted by canonical analysis. To expand this method to the case of more than two significant classes or of many classes in an image is straight forward.

We have also presented a method for evaluating the performance of current sensors by comparing classification accuracy with extracted features. It would be shown that the spectral bands of current sensors are not always optimal for a specific purpose, and can be improved by designing them appropriately.

Extension of this method to designing spectral bands for a specific purpose and to extracting quantitative information efficiently and accurately from hyper-dimensional data are subjects for a future study.

References from Journals:

Ready,P.J. and Wintz,P.A.,1973. Information Extraction, SNR Improvement, and Data Compression in Multispectral Imagery, IEEE Trans. Comm., COM-21, pp.1123-1131.

References from Books:

Kullback,S.,1959. Information Theory and Statistics, Wiley.

Schowengerdt,R.A.,1983. Techniques for Image Processing and Classification in Remote Sensing, Academic Press, pp.159-167.

Swain,P.H. and Davis,S.M.(ed.), 1978. Remote Sensing: The Quantitative Approach, McGraw-Hill, pp.358-361.

References from Other Literature:

Kiyasu,S. and Fujimura,S., 1993. Successive Feature Extraction from Hyperspectral Data, Proc. Int. Geoscience and Remote Sensing Symp. (IGARSS'93), Tokyo, Japan, pp.469-471.

Fujimura,S. et al., 1981. A Comparison of Automatic Classification Algorithms for Land Use Map by Remotely Sensed Data, Proc. IEEE Southeastcon '81, Huntsville, Alabama, pp.299-303.

Fujimura,S. and Kiyasu,S., 1994. Significance-Weighted Feature Extraction from Hyper-Dimensional Data, Proc. SPIE, vol.2318, pp.63-68.

Vane,G., 1988. First results from the Airborne Visible/ Infrared Imaging Spectrometer (AVIRIS), Proc. SPIE, vol.834, pp.166-174.