An Application of Canonical Correlation Analysis Technique to Land Cover Classification of LANDSAT Images

Jong-Hun Lee[a], Min-Ho Park, and Yong-Il Kim

This research is an attempt to obtain more accurate land cover information from LANDSAT images. Canonical correlation analysis, which has not been widely used in the image classification community, was applied to the classification of a LANDSAT image. It was found that it is easy to select training areas on the classification using canonical correlation analysis in comparison with the maximum likelihood classifier of ERDAS® software. In other words, the selected positions of training areas hardly affect the classification results using canonical correlation analysis. When the same training areas are used, the mapping accuracy of the canonical correlation classification results compared with the ground truth data is not lower than that of the maximum likelihood classifier. The kappa analysis for the canonical correlation classifier and the maximum likelihood classifier showed that the two methods are alike in classification accuracy. However, the canonical correlation classifier has better points than the maximum likelihood classifier in classification characteristics. Therefore, the classification using canonical correlation analysis applied in this research is effective for the extraction of land cover information from LANDSAT images and will be able to be put to practical use.

Manuscript received March 15; revised November 2, 1999.

[a]Electronic mail: jong@etri.re.kr

I. INTRODUCTION

Recently, new methods of classifying multispectral imagery have been developed and various implementations to improve the classification accuracy of satellite imagery have been performed [16], [24]. These methods for image classification are based on multivariate statistical analysis. This paper is the study of image classification, applying canonical correlation analysis (CCA) of multivariate statistical analyses used, to the principle of image classification. Up until now, a few researches using CCA have been published in remote sensing journals [11], [12], [19], [20], [23], [25]. Besides, the canonical correlation analysis has been applied to many fields including quantitative geography for spatial analysis; hydrology for snowmelt runoff forecast; sociological literature for occupational mobility; psychometrics; biometrics and so on [2], [10], [21]. The objective of this research is to obtain more accurate land cover information from LANDSAT images. In this experiment, the efficiency of the classification algorithm applying canonical correlation analysis was tested and evaluated.

This report is composed of the following three parts;

① Theoretical review of the canonical correlation analysis: The statistical theory of the canonical correlation analysis, which is based on the multivariate statistical analysis, is introduced for programming of a 'Canonical Correlation Classifier (CCC)' for image classification.

② Construction of an image classification algorithm using canonical correlation analysis: Various algorithms using the canonical correlation analysis are applied to image analysis, and an optimum algorithm among them is proposed.

③ Applications of the CCC to the LANDSAT image classification: The CCC is applied to a LANDSAT image of the research area, and the results are compared and discussed. In this study, LANDSAT-5 TM data was used. The maximum likelihood classifier (MLC) [26] with the same data was employed to be compared with the CCC. ERDAS® Image Processing System generated the land cover classification results using MLC [9]. The characteristics and advantages of CCC for image classification were examined. In addition, the significance of eigenvalue is tested, and the unclassified pixels are discriminated. MATLAB®, which is a program for matrix calculations, was used for coding the algorithm of the canonical correlation classifier. The image file of ground truth data for accuracy assessment was obtained from the .gis file generated by ERDAS® software through referencing a 1:10,000 topographic map and a 1 : 20,000 aerial photo.

II. THEORETICAL REVIEW OF THE CANONICAL CORRELATION ANALYSIS

1. Conceptual Overview

Canonical correlation analysis developed by H. Hotelling [22] in 1935 is a statistical method to identify and quantify the associations between two sets of variables. Canonical correlation analysis focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. The idea is, first, to determine the pair of linear combinations having the largest correlation. Next, we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair. The process continues. The pairs of linear combinations are called the canonical variables, and their correlations are called canonical correlations. The canonical correlations measure the strength of association between the two sets of variables. The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between the two sets of variables into a few pairs of canonical variables. In this study, CCA relates each pixel of a LANDSAT image to a land cover type.

2. The Mathematical Model

We are interested in measures of association between two groups of variables. The first group of p variables is represented by the $(p\times n)$ random vector $\mathbf{X}^{(1)}$. The second group of q variables is represented by the $(q\times n)$ random vector $\mathbf{X}^{(2)}$. We assume, in the theoretical development, that $\mathbf{X}^{(1)}$ represents the smaller sets, so that $p \le q$. For the random vectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, let

$$
\begin{aligned}
&E(\mathbf{X}^{(1)}) = \mu^{(1)}; \quad \mathrm{Cov}(\mathbf{X}^{(1)}) = \textstyle\sum_{11} \\
&E(\mathbf{X}^{(2)}) = \mu^{(2)}; \quad \mathrm{Cov}(\mathbf{X}^{(2)}) = \textstyle\sum_{22} \\
&\mathrm{Cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \textstyle\sum_{11} = \sum_{21}^{/}
\end{aligned} \tag{1}
$$

It will be convenient to consider $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ jointly. The random vector

$$
\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \hline \mathbf{X}^{(2)} \end{bmatrix} = \left[\begin{array}{cccc} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1n}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1}^{(1)} & x_{p2}^{(1)} & \cdots & x_{pn}^{(1)} \\ \hline x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1n}^{(2)} \\ x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{q1}^{(2)} & x_{q2}^{(2)} & \cdots & x_{qn}^{(2)} \end{array}\right] = [\mathrm{x}_1, \mathrm{x}_2, .., \mathrm{x}_n] \tag{2}
$$

where $\mathrm{x}_j = \begin{bmatrix} \mathrm{x}_j^{(1)} \\ \hline \mathrm{x}_j^{(2)} \end{bmatrix}$ has mean vector

$$
\underset{((p+q)\times 1)}{\mu} = E(\mathbf{X}) = \begin{bmatrix} E(\mathbf{X}^{(1)}) \\ \hline E(\mathbf{X}^{(2)}) \end{bmatrix} = \begin{bmatrix} \mu^{(1)} \\ \hline \mu^{(2)} \end{bmatrix}
$$

$$
\text{where} \quad \mu^{(1)} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{x}_j^{(1)}
$$

$$
\mu^{(2)} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{x}_j^{(2)} \tag{3}
$$

and covariance matrix, as following.

$$
\begin{aligned}
\underset{(p+q)\times(p+q)}{\sum} &= E(\mathbf{X}-\mu)(\mathbf{X}-\mu)^{/} \\
&= \left[\begin{array}{c|c} E(\mathbf{X}^{(1)}-\mu^{(1)})(\mathbf{X}^{(1)}-\mu^{(1)})^{/} & E(\mathbf{X}^{(1)}-\mu^{(1)})(\mathbf{X}^{(2)}-\mu^{(2)})^{/} \\ \hline E(\mathbf{X}^{(2)}-\mu^{(2)})(\mathbf{X}^{(1)}-\mu^{(1)})^{/} & E(\mathbf{X}^{(2)}-\mu^{(2)})(\mathbf{X}^{(2)}-\mu^{(2)})^{/} \end{array}\right] \\
&= \left[\begin{array}{c|c} \underset{(p\times p)}{\sum_{11}} & \underset{(p\times q)}{\sum_{12}} \\ \hline \underset{(q\times p)}{\sum_{21}} & \underset{(q\times q)}{\sum_{22}} \end{array}\right]
\end{aligned} \tag{4}
$$

The covariances between pairs of variables from different sets-one variable from $\mathbf{X}^{(1)}$, one variable from $\mathbf{X}^{(2)}$ are contained in $\sum_{12}$ or, equivalently, in $\sum_{21}$. That is, the pq elements of $\sum_{12}$ measure the association between the two sets. When p and q are relatively large, interpreting the elements of $\sum_{12}$ collectively is ordinarily hopeless. Moreover, it is often linear combinations of variables that are interesting and useful for predictive or comparative purposes. The main task of canonical correlation analysis is to summarize the associations between the $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ sets in terms of a few carefully chosen covariances (correlation) rather than the pq covariances in $\sum_{12}$.

Linear combinations provide simple summary measures of a set of variables. Set

$$\begin{aligned} \boldsymbol{U} &= \mathbf{a}'\mathrm{X}^{(1)} \\ \boldsymbol{V} &= \mathbf{b}'\mathrm{X}^{(2)} \end{aligned} \tag{5}$$

for some pair of coefficient vectors **a** and **b.** Using (5)

$$\begin{aligned} \mathrm{Var}(\boldsymbol{U}) &= \mathbf{a}'\mathrm{Cov}(\mathbf{X}^{(1)})\mathbf{a} = \mathbf{a}'\sum_{11}\mathbf{a} \\ \mathrm{Var}(\boldsymbol{V}) &= \mathbf{b}'\mathrm{Cov}(\mathbf{X}^{(2)})\mathbf{b} = \mathbf{b}'\sum_{22}\mathbf{b} \\ \mathrm{Cov}(\boldsymbol{U},\boldsymbol{V}) &= \mathbf{a}'\mathrm{Cov}(\mathbf{X}^{(1)},\mathbf{X}^{(2)})\mathbf{b} = \mathbf{a}'\sum_{12}\mathbf{b} \end{aligned} \tag{6}$$

We shall seek coefficient vectors **a** and **b** such that

$$\mathrm{Corr}(\boldsymbol{U},\boldsymbol{V}) = \frac{\mathbf{a}'\sum_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\sum_{11}\mathbf{a}}\sqrt{\mathbf{b}'\sum_{22}\mathbf{b}}} \tag{7}$$

is as large as possible. The first pair of canonical variables are the pair of linear combinations $\boldsymbol{U}_1, \boldsymbol{V}_1$ having unit variances, which maximize the correlation (7). The kth pair of canonical variables are the pair of linear combinations $\boldsymbol{U}_k, \boldsymbol{V}_k$ having unit variances, which maximize the correlation (7) among all choices uncorrelated with the previous k-1 canonical variable pairs. Then $\max_{\mathbf{a},\mathbf{b}} \mathrm{Corr}(\boldsymbol{U},\boldsymbol{V}) = \rho_1^*$ attained by the linear combinations (first variate pair)

$$\boldsymbol{U}_1 = \underbrace{\mathbf{e}_1'\sum_{11}^{-1/2}}_{\mathbf{a}_1'}\mathbf{X}^{(1)} \text{ and } \boldsymbol{V}_1 = \underbrace{\mathbf{f}_1'\sum_{22}^{-1/2}}_{\mathbf{b}_1'}\mathbf{X}^{(2)}.$$

The k-th pair of canonical variates, k = 2, 3, 4, ……., p,

$$\boldsymbol{U}_k = \mathbf{e}_k'\sum_{11}^{-1/2}\mathbf{X}^{(1)};\ \boldsymbol{V}_k = \mathbf{f}_k'\sum_{22}^{-1/2}\mathbf{X}^{(2)}$$

maximizes $\mathrm{Corr}(\boldsymbol{U}_k,\boldsymbol{V}_k) = \rho_k^*$ among those linear combinations uncorrelated with the proceeding 1, 2, 3, ….., k–1 canonical variables. Here $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$ are the eigenvalues of $\sum_{11}^{-1/2}\sum_{12}\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1/2}$ and $\mathrm{e}_1, \mathrm{e}_2, \cdots, \mathrm{e}_\mathrm{p}$ are the associated $(p\times 1)$ eigenvectors. (The quantities $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$ are also the p largest eigenvalues of the matrix $\sum_{22}^{-1/2}\sum_{21}\sum_{11}^{-1}\sum_{12}\sum_{22}^{-1/2}$ with corresponding $(q\times 1)$ eigenvectors $\mathrm{f}_1, \mathrm{f}_2, \cdots, \mathrm{f}_p$. Each f_i is proportional to $\sum_{22}^{-1/2}\sum_{21}\sum_{11}^{-1/2}\mathrm{e}_i$.) The canonical variates have the properties

$$\begin{aligned} &\mathrm{Var}(\boldsymbol{U}_k) = \mathrm{Var}(\boldsymbol{V}_k) = 1 \\ &\mathrm{Cov}(\boldsymbol{U}_k,\boldsymbol{U}_l) = \mathrm{Corr}(\boldsymbol{U}_k,\boldsymbol{U}_l) = 0 \quad k \neq 1 \\ &\mathrm{Cov}(\boldsymbol{V}_k,\boldsymbol{V}_l) = \mathrm{Corr}(\boldsymbol{V}_k,\boldsymbol{V}_l) = 0 \quad k \neq 1 \\ &\mathrm{Cov}(\boldsymbol{U}_k,\boldsymbol{V}_l) = \mathrm{Corr}(\boldsymbol{U}_k,\boldsymbol{V}_l) = 0 \quad k \neq 1 \end{aligned}$$

for $k, l = 1, 2, ….., p$.

If the original variables are standardized with $\mathbf{Z}^{(1)} = \left[Z_1^{(1)}, Z_2^{(1)}, \cdots, Z_P^{(1)}\right]'$ and $\mathbf{Z}^{(2)} = \left[Z_1^{(2)}, Z_2^{(2)}, \cdots, Z_P^{(2)}\right]'$, from first principles, the canonical variates are of the form

$$\begin{aligned} \boldsymbol{U}_k &= \mathbf{a}_k'\mathbf{Z}^{(1)} = \mathbf{e}_k'\rho_{11}^{-1/2}\mathbf{Z}^{(1)} \\ \boldsymbol{V}_k &= \mathbf{b}_k'\mathbf{Z}^{(2)} = \mathbf{f}_k'\rho_{22}^{-1/2}\mathbf{Z}^{(2)}. \end{aligned} \tag{8}$$

Here,
$\mathrm{Cov}(\mathbf{Z}^{(1)}) = \rho_{11}$, $\mathrm{Cov}(\mathbf{Z}^{(2)}) = \rho_{22}$, $\mathrm{Cov}(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}) = \rho_{12} = \rho_{21}'$ and $\mathbf{e}_k$ and $\mathbf{f}_k$ are the eigenvectors of $\rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\ \rho_{11}^{-1/2}$ and $\rho_{22}^{-1/2}\rho_{21}\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1/2}$, respectively. The canonical correlations, ρ_k^*, satisfy

$$\mathrm{Corr}(\boldsymbol{U}_k,\boldsymbol{V}_k) = \rho_k^*, \qquad k = 1, 2, \cdots, p \tag{9}$$

where $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$ are the nonzero eigenvalues of the matrix $\rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1/2}$ (or $\rho_{22}^{-1/2}\rho_{21}\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1/2}$).

III. AN IMAGE CLASSIFICATION ALGORITHM USING CANONICAL CORRELATION ANALYSIS

1. Concept

The concept of algorithm using canonical correlation analysis is as follows.

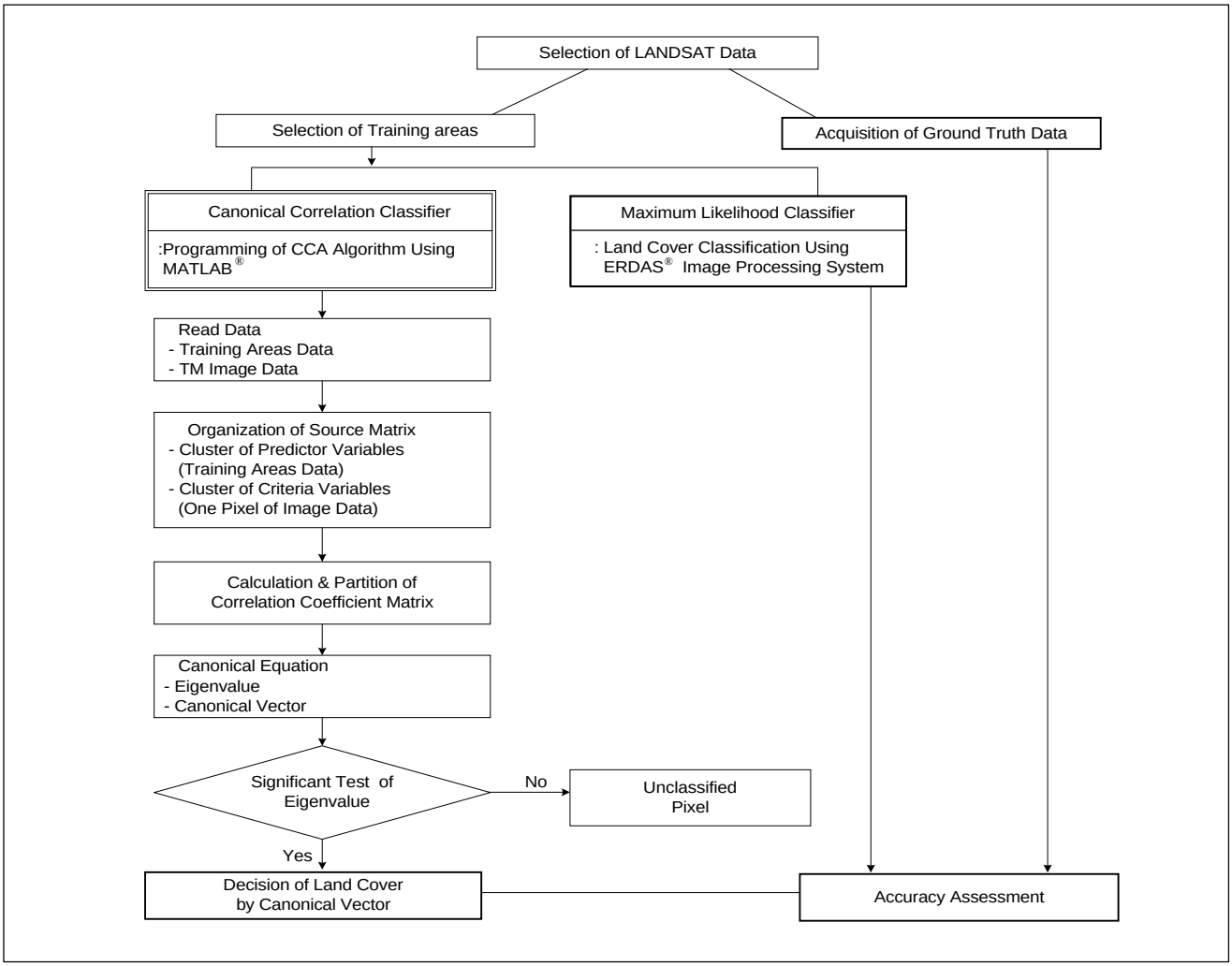


Fig. 1. The flowchart of canonical correlation classifier.

$X^{(2)}$ variables

T/A \ Band	1	2	.....	q
1	X_{11}	X_{12}	.....	X_{1q}
2	X_{21}	X_{22}	.....	X_{2q}
.	.	.	.	.
.	.	.	.	.
n	X_{n1}	X_{n2}	.....	X_{nq}

$X^{(1)}$ variable

Pixel \ Band	1
1	Y_1
2	Y_2
.	.
.	.
n	Y_n

Note:
T/A are training areas data.
Pixel is an individual pixel of LANDSAT image pixels for research area.
n is the number of band of LANDSAT image. MSS is 4, and TM is 7.

The $\mathbf{X}^{(2)}$ variables are the mean pixel values of the training areas, which are clearly identified from the aerial photo, a LANDSAT image and the topographic map. Each pixel in the LANDSAT image was used as $\mathbf{X}^{(1)}$ variable. In CCA, for convenience, the original variables ($X_{ji,}\, Y_j$) of data matrix are standardized with $\mathbf{Z}^{(1)} = \left[Z_1^{(1)}\right]'$ and $\mathbf{Z}^{(2)} = \left[Z_1^{(2)}, Z_2^{(2)}, \cdots, Z_q^{(2)}\right]'$.

In the case of the rectangular LANDSAT image with dimension x (row) × y (column), the classification per pixel was performed sequentially, scanning the image from upper-left corner to lower-right corner. Therefore, the number of the canonical correlation classifications for the study area was the same as the total pixel number. The classification class for each pixel of the study area image is decided as the class related to the largest value of the canonical weights, which are the q elements of the canonical vector $\mathbf{b}_1$ in $V_1 = \mathbf{b}_1' \mathbf{Z}^{(2)} = \mathbf{f}_1' \rho_{22}^{-1/2} \mathbf{Z}^{(2)}$. The canonical weight means the weight that each pixel is assigned to the training area.

2. Procedure

The procedure of canonical correlation classification is shown in Fig. 1, and each step is as follows:

(1) Data Reading

An original image data is converted into the binary raster files without header bytes according to the bands. In the case of Thematic Mapper data with 7 bands, 7 files are necessary. Also, a text file is created with the n-th band and q-th class from the average pixel value per band of training areas to obtain the training sample file representing the digital number of the classification classes. In this study, five classes were chosen.

(2) Construction of Source Data Matrix

Using the image data obtained from (1), one pixel is chosen and $n \times 1$ matrix defined as the $\mathbf{X}^{(1)}$ variable set is then constructed. Also, $n \times q$ matrix defined as the $\mathbf{X}^{(2)}$ variables set is constructed from the training areas file. Finally, by combining the $\mathbf{X}^{(1)}$ variable set matrix and the $\mathbf{X}^{(2)}$ variables set matrix, $n \times (1+q)$ source data matrix is produced [3], [5].

(3) Standardization of Source Data Matrix

All variables of data matrix are standardized to carry out all processes easily and simply.

(4) Calculation of Correlation Coefficient Matrix

The $(1+q) \times (1+q)$ correlation coefficient matrix is constructed from the standardized data matrix, which is used as the source matrix to produce the eigenvalue and the eigenvector for CCA.

(5) Partition of Correlation Coefficient Matrix

The $(1+q) \times (1+q)$ correlation coefficient matrix is partitioned into individual independent matrix including 1×1, $1 \times q$, $q \times 1$ and $q \times q$ matrices. Then, the inverse matrix, the square-root matrix, and the inverse matrix of the square-root matrix for each individual matrix can be respectively obtained.

(6) Solution of Canonical Equation

Considering the theoretical model, the newly created matrices in (5) are reconstructed to the source matrix (M). Using this matrix, the canonical equation is formulated and then the eigenvalue and the eigenvector can be obtained. This eigenvector produces the column vector **a**, **b** which indicate the canonical vectors of the two sets of variables. The number of the eigenvalue is determined as one, the number of $\mathbf{X}^{(1)}$ variable in this study. The element number of the column vector **b** corresponds to the number of land cover classes and each value of **b** vector elements indicates the weight of each pixel to the land cover class to the canonical correlation. Therefore, the main theme of the image classification process by CCA is that the land cover class corresponding to the largest element value is determined as the land cover class of a classified pixel.

(7) Statistical Significant Test of Eigenvalue

To determine the land cover types from CCA, the covariance $\sum_{12}$ or the canonical correlation coefficient between two sets of variables must not be zero. The distribution function of eigen value proposed by Fisher is adopted for the statistical significant test of eigenvalue [4] with a 95 % or 99 % confidence level to find the canonical correlation between two sets of variables. The statistical significance of one pixel, being less than 0.05 (or 0.01), rejects the null hypothesis, and then the class is classified with a 95 % (or 99 %) confidence level. Conversely, if the null hypothesis is adopted, the pixel cannot be assigned to one of the classes and the pixel is defined as an unclassified pixel. The unclassified pixel indicates that there is an area of a different feature from the selected training sites within the research area.

IV. THE APPLICATION OF CANONICAL CORRELATION CLASSIFIER FOR LANDSAT IMAGE CLASSIFICATION

1. Research Area

The research area for this study is a rectangular area of about 2.85 km × 2.14 km. It includes various land covers which consist of bareland, forest, grass, urban, water areas and includes the national cemetery, Tongjak Grand Bridge, Panp'o-Dong in Seoul. The topographic map (1994, revised) of the research area is shown in Picture 1 and the aerial photo of the same area, taken in October 1995, is shown in Picture 2. Picture 3 shows the LANDSAT image of the research area expressed by the natural color composite with RGB color assignments. The satellite image data for classification was obtained from the LANDSAT-5 TM with a 116–35 path-row and was the metropolitan area image including Seoul on June 2, 1992.

2. Image Registration to the Ground

The larger area, including the whole research area, was cut from the full scene. The twenty ground control points were distributed evenly among all. And, the research area image was geometrically transformed using the affine transformation equation to correspond to the map coordinate system [1]. The root mean square error (RMSE) of image registration to the ground was under 0.5 pixels. Then, a new image was generated through resampling using the bilinear interpolation method with the pixel values of the original image [26]. In the geometric correction and resampling, the dimension of one pixel was determined as 28.5 m × 28.5 m. Finally, the rectangular area with 100 pixels (column) × 75 pixels (row) was obtained from the geometrically corrected and resampled image. The map coordinate of the upper-left corner of this data is (X, Y) = (196,855.3 m, 445,245.5 m).

Picture 1. The topographic map of research area.

Picture 2. The aerial photo of research area.

3. Result of Application

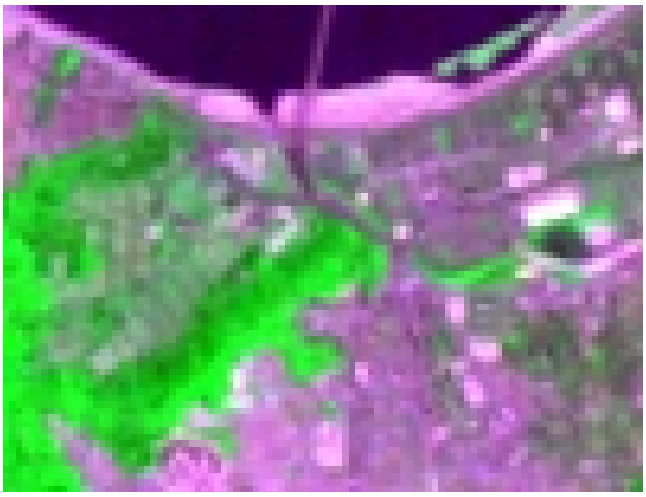
A. Selection of Class

In this study, the U.S. Geological Survey Land Use/Land Cover Classification System [13], which are widely used class codes for classifying a remotely sensed image, was adopted to apply the CCC method and the MLC method. Also, the classification level I was employed considering the resolution of TM data. Table 1 shows the class name in this study and U.S.G.S. reference comparatively.

B. Selection of Training Areas

The training areas have to be selected to apply CCC to LANDSAT image and to carry out MLC using the ERDAS® software for the same image. The training areas are the representative sample sites of known cover type. The training areas are used to compile a numerical "interpretation key" that describes the spectral attributes for each feature type of interest. Nevertheless, the specific criterion for selection of training areas was not adopted because it was more important to evaluate the efficiency and accuracy of CCC for LANDSAT TM image. Therefore, the positions of the training areas have only to be respectively clear when comparing the image of the research area with the aerial photo and the topographic map. Picture 7 shows the positions of the training areas. Each training area was selected based on the above-mentioned classes, and consists of the five following representative components.

① Bareland: In general, it consists of a schoolyard and soil with little plants.
② Forest: Its area consists of trees taller than 2 m.
③ Grass: It consists of shrubbery, grass, field, etc.
④ Urban: It has apartment sites, roads, small buildings, residences, concrete facilities, artificial constructions, etc.



Picture 3. Natural color image of research area.

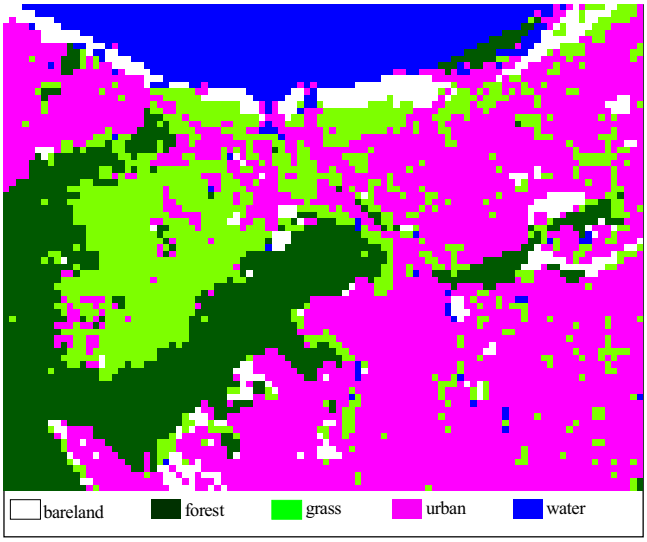
Table 1. Class name of this study & U.S.G.S. classification system.

Class Name in This Study	U.S.G.S. Classification System (Level I)
Bareland	Barren Land
Forest	Forest Land
Grass	Range land
Urban	Urban or Built-up Land
Water	Water

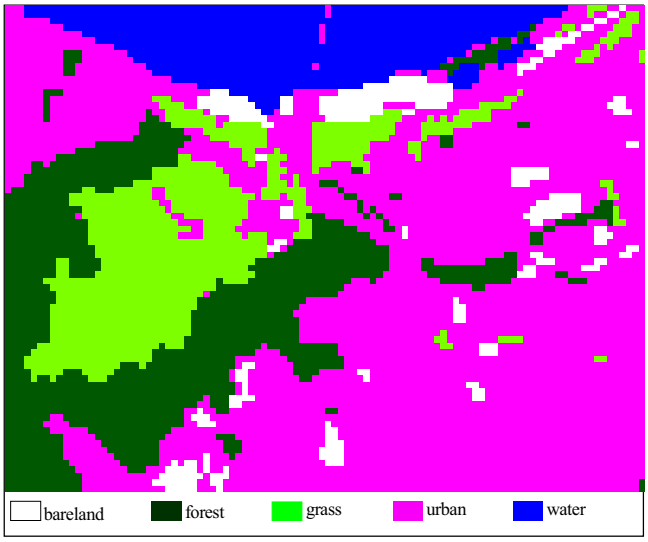
⑤ Water: It has rivers, water in rice fields, reservoirs, pools, etc.

C. Results of Canonical Correlation Classifier & Maximum Likelihood Classifier

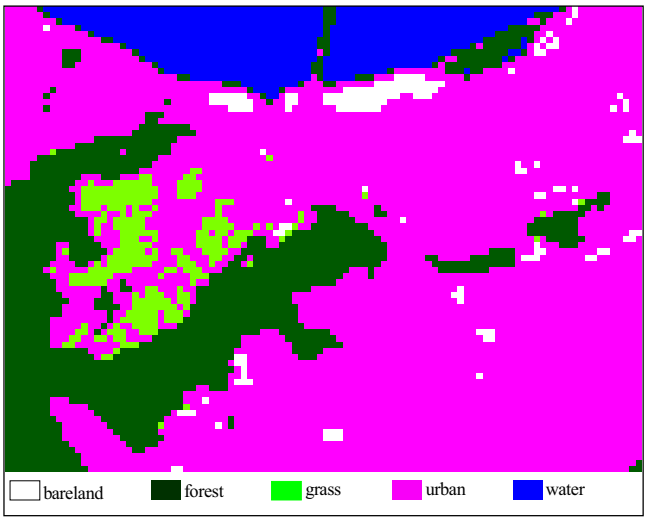
The classified results using the selected training areas are



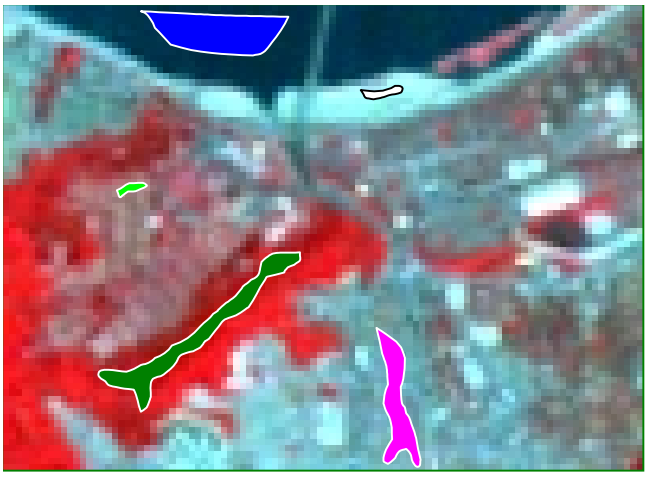
Picture 4. Result image by CCC.



Picture 6. Ground truth data.



Picture 5. Result image by MLC.



Picture 7. The positions of training areas.

Table 2. Result of each classification method.

(Units: Pixels)

Class / Method	Bareland	Forest	Grass	Urban	Water
CCC	424	1,440	1,374	3,437	825
MLC	170	1,497	289	4,873	671

shown in Table 2 (Note that the 'CCC' is the Canonical Correlation Classifier and the 'MLC' is the Maximum Likelihood Classifier.). Picture 4 and 5 show the displays of Table 2 (Note that in the pictures, Bareland, Forest, Grass, Urban, and Water are expressed by white, pine green, yellow green, magenta, and blue, respectively.). The processing time of CCC and MLC, respectively, is about 87 seconds and 4 seconds for 75× 100 sizes image with a Pentium 133 MHz PC. It is difficult to compare the processing time of two techniques with each other. The MLC using ERDAS® is a commercial software but the CCC using MATLAB® is an imperfect program coded by authors. It is predicted that the processing time of CCC will take more than that of MLC in the same condition.

4. Assessment & Analysis of Classification Accuracy

A. Acquisition of Ground Truth Data

The ground truth data were used to evaluate the classification accuracy of the two methods. It was obtained from the 1: 10,000 topographic map of the research area with a digitizer, and stored in a file of the same format as the classification results image. The aerial photo was identified for more accurate digitizing of land

Table 3. Pixels by each class of ground truth data.

(Units: Pixels)

Class	Barela	Forest	Grass	Urban	Water
Number of Pixels	351	1,415	997	3,942	795
Percentage (%)	4.68	18.87	13.29	52.56	10.60

cover boundaries. In performing the digitization, some errors between the boundary lines of land covers might be included. Also, areas with land covers that are not found in the topographic map and the aerial photo exist. These were very small areas expressed by one or two pixels. In this paper, the acquisition of ground truth data contains some limitation. The limitations are; the disagreement of the year in which the LANDSAT image, the topographic map and the aerial photo were produced; the simplification of ground truth data by the omission of small land cover area; and the indistinct representation of land cover information in the topographic map. In addition, there are the digitizing errors of land cover boundary lines, the registering error of the image to the ground. However, despite the above problems, the assessment of producer's accuracy, user's accuracy, mapping accuracy and overall accuracy in CCC and MLC was useful. The ground truth data are shown in Picture 6. Table 3 shows the number of pixels belonging to each class of ground truth data.

B. Construction and Evaluation of Contingency Table

Tables 4 and 5 show the contingency tables for the accuracy comparison of CCC with MLC [15], [17]. The total number of pixels in the research area is 7,500 and the number of correctly classified pixels is equal to the sum of five diagonal element values of the contingency table as shown in Tables 4 and 5 [7]. From these values, overall accuracy, producer's accuracy, user's accuracy and mapping accuracy can be calculated [26]. Mapping accuracy for each class is stated as the number of correctly identified pixels within each ground class divided by that number plus error pixels of commission and omission [14], [18]. Also, the classification accuracy of CCC after extracting the unclassified pixels is calculated, and that of MLC is obtained through excluding the pixels with the same positions as the unclassified pixels produced by CCC.

Table 6 shows the comparison results of accuracies between CCC and MLC. The original image for both classifying methods consists of 7,500 pixels in total, respectively. The accuracies of the modified image with 7,015 pixels, excluding the 485 unclassified pixels, are obtained. Table 6 indicates that both classified results of the image, excluding the unclassified pixels,

Table 4. Contingency table of CCC.

(Units: Pixels)

Ground Class / LANDSAT Class	B	F	G	U	W	Sum (row)
B	227	0	13	174	10	424
F	1	1,261	88	84	6	1,440
G	61	95	693	521	4	1,374
U	57	59	201	3,097	23	3,437
W	5	0	2	66	752	825
Sum (column)	351	1,415	997	3,942	795	7,500

Table 5. Contingency table of MLC.

(Units : Pixels)

Ground Class / LANDSAT Class	B	F	G	U	W	Sum (row)
B	145	0	0	25	0	170
F	0	1,278	86	61	72	1,497
G	0	0	278	11	0	289
U	206	137	633	3,843	54	4,873
W	0	0	0	2	669	671
Sum (column)	351	1,415	997	3,942	795	7,500

are more accurate than the original image. Therefore, the extraction of unclassified pixels by CCC is useful for classifying the image. The user's accuracy of bareland through extracting the unclassified pixels was reduced compared with that of the original image, since the bareland area of ground truth data has a high variation in itself and a new land cover type may exist.

Since the research area of this study was more than 50 % urban area, and the overall accuracy was mainly affected by the accuracy of the urban area, the overall accuracy of MLC was better than that of CCC as shown in Table 6. The land cover type and the size of land cover area can have an influence on the overall accuracy. To analyze each accuracy of land covers, producer's accuracy, user's accuracy and mapping accuracy were calculated. In the case of bareland, grass and water, the producer's accuracies of CCC were much higher than MLC. Considering the results of MLC, the accuracies of bareland and grass were very low, while that of urban area was very high. It is because MLC tends to classify the areas that are not urban area into urban area. Since there are various land cover types in an urban area, the variation of its numerical data is generally large. Considering

Table 6. The comparison results of accuracies between CCC and MLC.

(Units: %)

Classification Method \ Type of Accuracy		Overall Accuracy	Producer's Accuracy				
			B	F	G	U	W
CCC	Original Image	80.4	64.7	89.1	69.5	78.6	94.6
	Excluding Unclassified Pixel	82.2	71.1	89.5	69.3	80.5	96.5
MLC	Original Image	82.8	41.3	90.3	27.9	97.5	84.2
	Excluding Unclassified Pixel	84.4	48.5	91.6	28.5	97.6	87.9

Classification Method \ Type of Accuracy		User's Accuracy					Mapping Accuracy				
		B	F	G	U	W	B	F	G	U	W
CCC	Original Image	53.5	87.6	50.4	90.1	91.2	41.4	79.1	41.3	72.3	86.6
	Excluding Unclassified Pixel	47.0	88.6	53.5	90.1	94.2	N/A	N/A	N/A	N/A	N/A
MLC	Original Image	85.3	85.4	96.2	78.9	99.7	38.6	78.2	27.6	77.3	83.9
	Excluding Unclassified Pixel	81.8	86.0	97.5	80.8	99.7	N/A	N/A	N/A	N/A	N/A

*N/A: Not Available

Table 7. Individual error matrix kappa analysis results.

Error Matrix	KHAT	Variance	Z statistic
Canonical Correlation Classifier	0.7136	0.000044562	106.89
Maximum Likelihood Classifier	0.7164	0.000048647	102.72

Table 8. Kappa analysis results for the pairwise comparison of the error matrices.

Pairwise Comparison	Z statistic
CCC vs. MLC	0.2994

that MLC method uses the variances of classes, the ambiguous pixels for classifying can be easily included in urban area. Therefore, to improve the accuracies of bareland and grass using MLC method, many trials are necessary with increased training areas and larger variance of each class.

CCC method can be useful for classifying LANDSAT image data, considering that the producer's accuracy of CCC is not less than that of MLC. Especially, CCC for areas including grass and water can produce good results. The user's accuracy of MLC is lower than that of CCC in urban and forest areas. It shows that most of the pixels with omission errors in MLC were classified into urban area and forest area. The very meaningful accuracy for the comparison of classification methods is mapping accuracy. The mapping accuracy of each class in CCC is better than that of the same class in MLC excluding urban class. This point shows that CCC is considerably efficient to classification of LANDSAT TM image data compared with MLC.

For more complete comparison and accuracy assessment between the two classification methods, the kappa analysis as a statistical test of significance between the two error matrices was used [6], [8]. For the performance of kappa analysis, the KHAT statistic (the estimate of the Kappa statistic), Z statistic, for testing the significance of a single error matrix, and Z statistic for testing, if two independent error matrices are significantly different, were computed. Table 7 shows the kappa analysis results of individual error matrix. Table 8 shows the kappa analysis results for the comparison of the two error matrices.

The KHAT values for the two error matrices in Table 7 are 0.7136 and 0.7164, respectively, and so both classifications represent moderate agreement. Also the Z statistic values for the two error matrices in Table 7 are both more than 100, and so both classifications are significantly better than random. The Z statistic for pairwise comparison of the two classifications in Table 8 is 0.2994, and so this value reveals that the two matrices are not significantly different. Therefore, it could be concluded that CCC might be useful because both CCC and MLC produce approximately equal classifications.

V. CONCLUSIONS

Conclusively, the principal difference in classification criterion between MLC and CCC is that MLC uses normal probability density function as the discriminant function while CCC

uses the canonical vector of linear combination of $\mathbf{X}^{(2)}$ variables, when the correlation between two linear combinations is largest, as the discriminant function. In other words, MLC is concerned about the covariance of image data within each class in training areas, but CCC is concerned about the covariance of the average values of each class and each band in training areas. Actually, the spatial distribution of satellite image data is not usually a normal distribution. Therefore, CCC does not have the defect of MLC, which assumes the normal distribution of data.

Especially, the characteristics and advantages of the CCC method compared with the MLC method are as follows;

(1) Using the significance test for eigenvalue, unclassified areas can be extracted.

(2) Training areas can be easily selected without trials and errors. In other words, even though only the distinct areas for training areas are selected, the classification results of CCC are not degraded as much as the classification results of MLC.

(3) In the case of grass and water areas, the CCC method can be more accurate and effective. That is, the small grass areas and water within urban areas can be classified better. Therefore, CCC method can be used for harvest forecasting and surface water detection.

VI. FURTHER STUDIES AND RECOMMENDATIONS

Since the proposed CCC method makes independent classi fication criterion for each pixel, the process requires a long processing time when using a personal computer. The confidence level for significance test to extract unclassified pixels must be decided through more experiments and analyses for various areas. Also, the selection of effective bands is required to adopt the most efficient band number for CCC.

REFERENCES

[1] Yong Il Kim, *Improving Correctness in the Satellite Remote Sensing Data Analysis-Laying Stresses on the Topographic Spectral Reflectance Correction and the Statistical Classification Techniques*, doctoral dissertation, Seoul National University, 1991.

[2] Hyun Woo Nam, *Quantitative Geography*, Peop Mun Sa, Seoul, Korea, 1992.

[3] Hee Yeon Lee, *Geographic Statistics*, Peop Mun Sa, Seoul, Korea, 1991.

[4] M. S. Bartlett, "The Statistical Significance of Canonical Correlations," *Biometrica*, Vol. 32, 1941.

[5] David Clark, *Understanding Canonical Correlation Analysis*, Norwich: Geo Abstracts Ltd., 1975.

[6] R. G. Congalton, "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data," *Remote Sensing of Environment*, Vol. 37, 1991, pp. 35–46.

[7] R. G. Congalton, Mary Balogh, Cindy Bell, Kass Green, J. A. Milliken, and Robert Ottman, "Mapping and Monitoring Agricultural Crops and Other Land Cover in the Lower Colorado River Basin," *Photogrammetric Engineering & Remote Sensing*, 64(11): 1107–1113, 1998.

[8] R. G. Congalton, and K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, CRC/Lewis Press, 1999, pp. 43–53.

[9] *ERDAS Field Guide, Fourth Edition*, ERDAS Inc., Atlanta, Georgia, 1997.

[10] Franklin D. Wilson, *Institute for Research on Poverty—Canonical Correlation and the Relation between Sets of Variables*, doctoral dissertation, University of Wisconsin-Madison, 1975.

[11] Hanaizumi Hiroshi, Chino Shinji, and Fujimura Sadao, "Binary Division Algorithm for Clustering Remotely Sensed Multispectral Images," *IEEE Transactions on Instrumentation and Measurement*, Vol. 44, No. 3, 1995, pp. 759–763.

[12] M. E. Jakubauskas, "Canonical Correlation Analysis of Coniferous Forest Spectral and Biotic Relations," *International Journal of Remote Sensing*, Vol. 17, No. 12, 1996, pp. 2323–2332.

[13] John R. Jensen, *Introductory Digital Image Processing—A Remote Sensing Perspective*, Prentice-Hall, 1996.

[14] C. P. Lo, L. J. Watson, "The Influence of Geographic Sampling Methods on Vegetation Map Accuracy Evaluation in a Swampy Environment," *Photogrammetric Engineering & Remote Sensing*, 64(12): 1189–1200, 1998.

[15] R. D. Macleod, R. G. Congalton, "A Quantitative Comparison of Change-Detection Algorithms for Monitoring Eelgrass from Remotely Sensed Data," *Photogrammetric Engineering & Remote Sensing*, 64(3): 207–216, 1998.

[16] Mark J. Carlotto, "Spectral Shape Classification of Landsat Thematic Mapper Imagery," *Photogrammetric Engineering & Remote Sensing*, 64(9) : 905–913, 1998.

[17] S. V. Muller, D. A. Walker, F. E. Nelson, N. A. Auerbach, J. G. Bockheim, S. Guyer, and D. Sherba, "Accuracy Assessment of a Land-Cover Map of the Kuparuk River Basin, Alaska: Considerations for Remote Regions," *Photogrammetric Engineering & Remote Sensing*, 64(6) : 619–628, 1998.

[18] Nicholas M. Short, *Remote Sensing and Photo Interpretation Tutorial*, Goddard Space Flight Center, NASA, 1997, P.13-3.

[19] Niemeyer Irmgard, Canty Morton, and Klaus Dieter, "Possibilities and Limits of Remote Sensing for the Verification of International Agreements: Algorithms to Detect Changes at Nuclear Plants," *International Geoscience and Remote Sensing Symposium*, Vol. 2, IEEE, Piscataway, 1998, pp. 819–821.

[20] Okumura Hiroshi, Sugita Tadashi, Matsumoto Hironori, and Takeuchi Nobuo, "Noise Reduction Method for Lidar Echo Data Based on Multivariate Analysis Method," *Better Understanding of Earth Environment International Geoscience and Remote Sensing Symposium*, Vol. 2, IEEE, Piscataway, 1993, pp. 454–456.

[21] Padoong Torranin, *Aplicability of Canonical Correlation in Hydrology*, hydrology papers, Colorado State University, 1972.

[22] Richard A. Johnson and Dean W. Wichern, *Applied Multivariate Statistical Analysis*, Third Edition, Prentice Hall, 1992.
[23] R. M. Korobov and V. Ya Railyan, "Canonical Correlation Relationships among Spectral and Phytometric Variables for Twenty Winter Wheat Fields," *Remote Sensing of Environment*, Vol. 43, No. 3, 1993, pp. 1–10.
[24] K. M. Sharma, S. A. Sarkar, "A Modified Contextual Classification Technique for Remote Sensing Data," *Photogrammetric Engineering & Remote Sensing*, 64(4): 273–280, 1998.
[25] K.V. Shettigara, "Linear Transformation Technique for Spatial Enhancement of Multispectral Images Using a Higher Resolution Data Set," *Digest International Geoscience and Remote Sensing Symposium*, Vol. 4, 1989, pp. 2615–2618.
[26] Thomas M. Lillesand and Ralph W. Kiefer, *Remote Sensing and Image Interpretation*, Third Edition, 1994.

Jong-Hun Lee received the Bachelor of Science degree in civil engineering in 1981 from Yonsei University, Korea. After military service, he entered Yonsei Graduate School, receiving the Master's degree in civel engineering. He continued his studies at Cornell University, Ithaca, NY, where he completed the Master's and Ph.D. degrees in remote sensing. He is currently a Project Leader of Geographic Information Systems (GIS) Research Team of ETRI and is Responsible for coordinating research activities in GIS at ETRI.

Min-Ho Park received the B.S., M.S. and Ph.D. degrees from the Department of Civil Engineering at Seoul National University, Seoul, Korea in 1986, 1988 and 1996, respectively. Currently, he is an associate professor in the Department of Land Administration, Mokpo National University, Chonnam, Korea. His main research interests are the satellite image classification and analysis in the remote sensing community, and the information system construction using GIS/LIS.

Yong-Il Kim was educated at Seoul National University in the Rep. of Korea, where he received B.S. degrees in urban engineering in 1986. And he got the M.S. and Ph.D. degrees in the field of remote sensing in 1988 and 1991, respectively. He joined the faculty of Seoul National University in 1993, where he is currently working as an assistant professor at the school of civil, urban, and geosystem engineering. He stayed at Cornell University for one year as a visiting researcher in 1997. His major research interests include Remote Sensing, Global Positioning Systems (GPS), Geographic Information Systems, etc. And during past 10 years, he has been involved as project leader in several large projects such as Standardization of Digital Road Map Databases, Development of Feature Extraction Algorithms for Remote Sensing, etc. At present, he is a Member of Surveying Committee of National Geographic Institute, and also a Director and Editor of Journal of the Korean Society for Geo-Spatial Information Systems, Journal of the Korean Society of Geodesy, Photogrammetry, and Cartography, and Journal of the Korean Society of Remote Sensing.