INFORMATION PROCESSING FOR REMOTE SENSING

Remote sensing is a technology through which information about an object is obtained by observing it from a distance. This article is specifically concerned with obtaining information about Earth through remote sensing.

Earth can be observed remotely in many ways. One of the earliest approaches to remote sensing was observing Earth from a hot air balloon using a camera, or just the human eye. Today, remotely sensed Earth observational data are routinely obtained from instruments onboard aircraft and spacecraft. These instruments observe Earth through various means, including optical telescopes and microwave devices at wavelengths from optical through microwave, including the visible, infrared, passive microwave, and radar.

Other articles in this series discuss the most widely employed approaches for obtaining remotely sensed data. This article discusses methods for effectively extracting information from the data once they have been obtained.

Most information processing of Earth remote sensing data assumes that Earth's curvature and terrain relief can be ignored. In most practical cases, this is a good assumption. It is beyond the scope of this article to deal with the special cases where it is not, such as with a relatively low flying sensor over mountainous terrain or when the sensor points toward Earth's horizon. This article deals with information processing of two-dimensional image data from down-looking sensors.

Remotely sensed image data can have widely varying characteristics, depending on the sensor employed and the wavelength of radiation sensed. This variation can be very useful, as in most cases this variation corresponds to information about what is being sensed on Earth. A key task of information processing for remote sensing is to extract the information contained in the variations of remotely sensed image data with changes in spatial scale, spectral wavelength, and the time at which the data are collected. Data containing these types of variations are referred to as multiresolution or multiscale data, multispectral data, and multitemporal data, respectively.

In some cases, Earth scientists may find useful a combined analysis of image data taken at different spatial scales and/or orientations by separate sensors. Such analysis will become even more desirable over the next several years as the number and variety of sensors increase under such programs as NASA's Earth Observing System. This type of analysis requires the determination of the correspondence of data points in one image to data points in the other image. The process of finding this correspondence and transforming the images to a common spatial scale and orientation is called *image registration.* More information on image registration can be found in REMOTE SENSING GEOMETRIC CORRECTIONS.

Multispectral data are often collected by an instrument that is designed to collect the data in such a way that they are already registered. In other cases, however, small shifts in location need to be corrected by image registration. Multitemporal data must almost always be brought into spatial alignment using image registration, as must multiresolution data when obtained from separate sensors.

Several approaches have been developed for analyzing registered multiscale/spectral/temporal data. Because most of these techniques were originally developed for analyzing multispectral image data, they will be discussed in terms of that context. However, many of these techniques can also be used in analyzing multiscale and/or multitemporal data. In the following discussion, each scale, spectral, or temporal

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering.* Copyright © 1999 John Wiley & Sons, Inc.

manifestation of the image data is referred to as an image band. Figure 1 gives an example of remotely sensed multispectral image data.

Sometimes important information identifying the observed ground objects is contained in the ratios between bands. Ratios taken between spectrally adjacent bands correspond to the discrete derivative of the spectral variation. Such band ratios measure the rate of change in spectral response and distinguish classes with a small rate of change in spectral response from those with a large rate of change. Other spectral ratios have been defined such that they relate to the amount of photosynthetic vegetation on the Earth's surface. These are called *vegetation indices.* Spectral ratios are also useful in the analysis of image data containing significant amounts of topographic shading. The process of spectral ratioing tends to reduce the effect of this shading.

The data contained in each band of multispectral image data are often correlated with the data from some of the other bands. When desirable to do so, this correlation can be reduced by transforming the data in such a way that most of the data variation is concentrated in just a few transformed bands. Reducing the number of image bands in this way not only may make the information content more apparent but also serves to reduce the computation time required for analy-

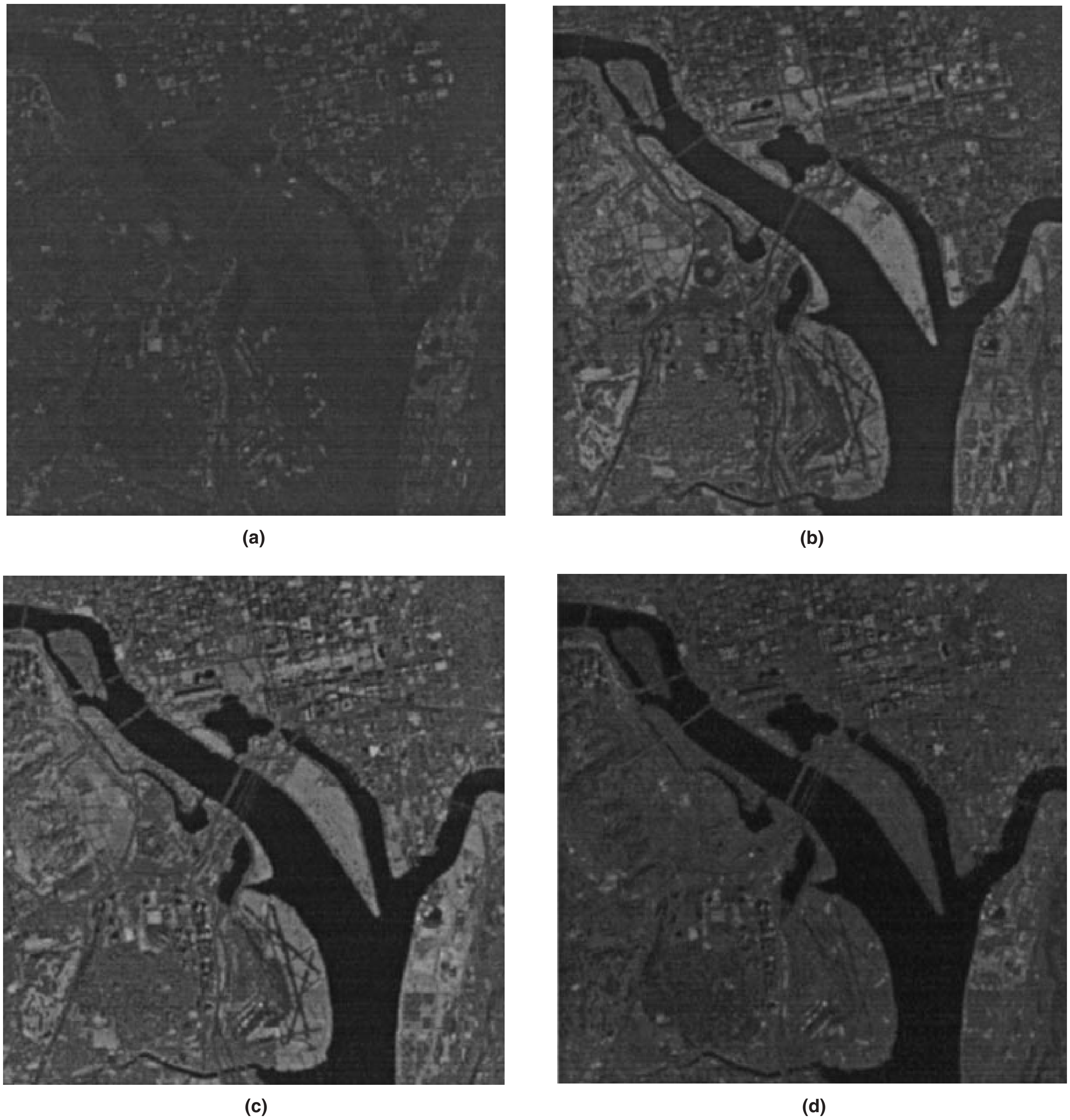

Figure 1. An example of remotely sensed multispectral imagery data. Displayed are selected spectral bands from a seven-band Landsat 5 Thematic Mapper image of Washington, DC: (a) spectral band 2 (0.52–0.60 μm), (b) spectral band 4 (0.76–0.90 μm), (c) spectral band 5 (1.55–1.75 μm), and (d) spectral band 7 (2.08–2.35 μm).

sis; it can be used, in effect, to "compress" the data by discarding transformed bands with low variation. There are many such transformations for accomplishing this concentration of variation. One is called Principal Component Analysis (PCA) or the Principal Component Transform (PCT). Other useful transforms are the Canonical Components Transform (CCT) and the Tasseled Cap Transform (TCT).

The process of labeling individual pixels in the image data as belonging to a particular ground cover class is called *image classification.* (An image data vector from a particular spatial location is called an image picture element or pixel.) This labeling process can be carried out directly on the remotely sensed image data, on image features derived from the original image data (such as band ratios or data transforms), or on combinations of the original image data and derived features. Whatever the origin of the data, the classification feature space is the n-dimensional vector space spanned by the data vectors formed at each image pixel.

The two main types of image classification are unsupervised and supervised. In unsupervised classification, an analysis procedure is used to find natural divisions, or clusters, in the image feature space. After the clustering process is complete, the analyst associates class labels with each cluster. Several clustering algorithms are available, ranging from the simple K-means algorithm, where the analyst must prespecify the number of clusters, to the more elaborate ISODATA algorithm, which automatically determines an appropriate number of clusters.

In supervised classification, the first step is to define a description of how the classes of interest are distributed in feature space. Then each pixel is given the class label whose description is closest to its data value. Determining the description of how the classes of interest are distributed in feature space is the training stage of supervised classification. An approach commonly used in this stage is to identify small areas throughout the image data that contain image pixels of the classes of interest. This is usually done using image interpretation combined with ground reference information (e.g., a map of the locations of areas of classes of interest obtained through a ground survey, knowledge from a previous time, or other generalized knowledge about the area in question). Then the classes are characterized according to the model used for the next step: the classification stage.

One of the simplest classification algorithms is the *minimum-distance-to-means classifier.* When this classifier is used, the vector mean value of each class is calculated in the training stage, and each data pixel is labeled as belonging to the closest class by some distance measure (e.g., the Euclidean distance measure). This classifier can work very well if all classes have similar variance and well-separated means. However, its performance may be poor when the classes of interest have a wide range of variance.

A relatively simple classification algorithm that can account for differing ranges of variation of the classes is the *parallelepiped classifier.* When this classifier is used, the range of pixel values in each band is noted for each class from the training stage, and image data pixels that do not fall uniquely into the range values for just one class are labeled as "unknown." This classifier gets its name from the fact that the feature space locations of pixels belonging to individual classes form parallelepiped-shaped regions in feature space. The number of pixels in the unknown class can be reduced by modeling each class by a union of several parallelepiped-shaped regions.

Table 1. Accuracy Comparison (Percent Correct Classification) Between Classifications of the Original and Presegmented Landsat Thematic Mapper Images [from (1)]

Ground Cover Class	Original Image (%)	Presegmented Image (%)
Water/marsh	73.7	79.3
Forest	74.8	75.6
Residential	54.4	64.9
Agricultural and domestic grasses	81.9	83.4
Overall	79.2	80.9

One of the most commonly used classification algorithms for remotely sensed data is the *Gaussian maximum likelihood classifier* (also called the ML classifier). The ML classifier often performs very well in cases where the minimum-distance-to-means classifier or the parallelepiped classifier perform poorly. This is because the ML classifier not only accounts for differences in variance between classes but also accounts for differences in between-band correlations. An even more general classification approach is a *neural network classifier.* The flexibility of the neural network classifier comes from its ability to generate totally arbitrary feature space partitions.

The analysis approaches discussed to this point have treated the data at each spatial location separately. This per-pixel analysis ignores the information contained in the spatial variation of the image data. One approach that can exploit the spatial information content in the data is *image segmentation.* Image segmentation is a partitioning of an image into regions based on the similarity or dissimilarity of feature values between neighboring image pixels. An image region is defined as a collection of image pixels in which, for any two pixels in this collection, there exists a spatial path connecting these two pixels, which travels only through pixels contained in the region. After an image is segmented into regions, the image can be labeled region by region using one of the classification approaches mentioned previously. The combination of image segmentation and image classification often produces superior results to per-pixel image classification (see Table 1).

A relatively recent development in remotely sensing instrumentation is imaging spectrometers, such as the Airborne Visible-InfraRed Imaging Spectrometer (AVIRIS). Imaging spectrometers produce *hyperspectral data,* consisting of hundreds of spectral bands taken at narrow and closely spaced spectral intervals. Two main types of specialized analysis approaches are currently under development for this type of data. One approach is an attempt to match laboratory or field reflectance spectra with remotely sensed imaging spectrometer data. The success of this approach depends on precise calibration of the remotely sensed data and careful compensation or corrections for atmospheric, solar, and topographic effects. The other approach depends on exploiting the unique mathematical characteristics of very high dimensional data. This approach does not necessarily require corrected data.

FEATURE EXTRACTION

The multispectral image data provided by a remote sensing instrument can be analyzed directly. However, in some cases, it may be beneficial to analyze features extracted from the original data. Such *feature extraction* commonly takes the form of subsetting and/or mathematically transforming the original data. It is used to compensate for one or more of the following problems often encountered with remotely sensed data: atmospheric effects, topographic shading effects, spectral band correlation, and lack of optimization for a particular application.

Atmospheric Effects

Most remote sensing data are collected from sensors on satellite platforms orbiting above Earth's atmosphere. Earth's atmosphere can have a significant effect on the quality and characteristics of such satellite-based remote sensing data.

For this article, it is sufficient to introduce the following first-order model for the input radiance to an Earth-orbiting sensor (2):

$$L(x, y, \lambda) = \frac{1}{\pi} T_{\mathrm{s}}(\lambda) T_{\mathrm{v}}(\lambda) E_0(\lambda) \cos[\theta(x, y)] \rho(x, y, \lambda) + L_{\mathrm{h}}(\lambda) \quad (1)$$

The solar irradiance from the sun $E_0(\lambda)$ provides the source radiation for the remote sensing process. This is the irradiance as it would be measured at the top of Earth's atmosphere and is referred to as the exo-atmospheric solar irradiance. The atmosphere affects the signal received by the sensor on two paths: (1) between the top of the atmosphere and Earth's surface (solar path) and (2) between the surface and the sensor (view path). The spectral transmittance of the atmosphere $T_s(\lambda)$ along the solar path or $T_v(\lambda)$ along the view path is generally high except in prominant molecular absorption bands attributable mainly to carbon dioxide and water vapor, as illustrated in Fig. 2. The cos[$\theta(x, y)$] term is the spatial variation of irradiance at the surface resulting from the solar zenith angle and topography, which determine the angle at which the incident radiation strikes the surface. The spatial and spectral variations in diffuse surface reflectance are modeled by the function $\rho(x, y, \lambda)$. A Lambertian, or perfectly diffuse, reflecting surface is assumed in Eq. (1). The atmospheric view path radiance $L_h(\lambda)$ is additive and increases at shorter visible wavelengths as a result of Rayleigh molecular scattering. This is the effect that causes the clear sky to appear blue. A related, second-order effect from down-scattered radiation (skylight) that is subsequently reflected at the surface into the sensor view path is not included in Eq. (1). This effect allows the surface-related signal in shadowed areas to be recovered, although with a spectral bias toward shorter wavelengths.

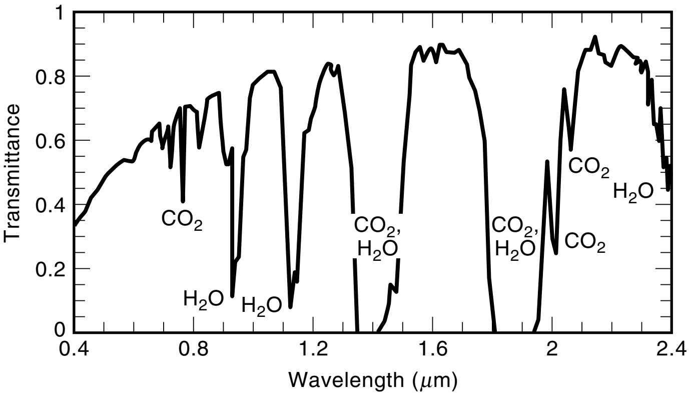

Figure 2. Atmospheric transmittance for a nadir path as estimated with the atmospheric modeling program MODTRAN (3). The transmittance is generally over 50% throughout the visible to short-wave infrared (SWIR) spectral region, except for prominent absorption bands resulting from atmospheric molecular constituents. Remote sensing of the Earth is not possible at wavelengths corresponding to the strongest absorption bands. The relatively lower transmittance below about 0.6 μm results from Rayleigh scattering losses.

Correction for atmospheric effects requires modeling or measurement of the various independent terms in Eq. (1), namely, $T_s(\lambda)$, $T_v(\lambda)$, $E_0(\lambda)$, and $L_h(\lambda)$ and, given the remotely sensed data measurements, $L(x, y, \lambda)$, solution of Eq. (1) for the surface spatial and spectral variations $\rho(x, y, \lambda)$. The cos[$\theta(x, y)$] term is a topographic effect that is described in the next section.

The path radiance term $L_h(\lambda)$ is primarily of concern at short, blue-green wavelengths, and the transmittance terms $T_s(\lambda)$ and $T_v(\lambda)$ are usually ignored for coarse multispectral sensing, such as with Landsat TM, where the bands are placed within atmospheric "windows" of relatively high and spectrally flat transmittance. For hyperspectral data, however, knowledge of and correction for transmittance is usually required if the data are to be compared to reflectance spectra measured in a laboratory.

Topographic Effects

Most areas of Earth have topographic relief. The irradiance from solar radiation is proportional to the cosine of the angle between the normal vector to the surface and the vector pointing to the sun. A surface element normal to the solar vector receives the maximum possible irradiance. Any element at some other angle will receive less. This spatially variant factor is the same in all solar reflective bands and, therefore, introduces a correlation across these bands.

Spectral Band Ratios. The pixel-by-pixel ratio of adjacent spectral bands corresponds to the discrete derivative of the spectral function. It therefore measures the rate of change in spectral signature and distinguishes classes with a small rate of change from those with a large rate of change. For example, the ratio of a near infrared (NIR) band to a red band will show a high value across the vegetation edge at 700 nm, whereas a ratio of a red band to a green band will show a small value for both vegetation and soil.

For bands where the atmospheric path radiance is small [e.g., in the NIR or short-wave infrared (SWIR) spectral regions], the spectral band ratio will be proportional to the surface reflectance ratio. In this case, the spectral band ratio is insensitive to topographic effects. If the path radiance is not small, then it should be reduced or removed using a technique such as Dark Object Subtraction (DOS) before spectral band ratios are calculated (2).

Vegetation Indices. A number of specific ratio formulae have been defined in attempts to obtain features that relate to the amount of photosynthetic vegetation on the Earth's surface. All depend on the red and NIR spectral reflectances (i.e., calibrated data). They are summarized in Table 2 and plotted as isolines in the NIR-red reflectance space in Fig. 3.

Table 2. Definition of Common Vegetation Indices

Index	Formula	Remarks
Ratio (R)	$\dfrac{\rho_{\mathrm{NIR}}}{\rho_{\mathrm{red}}}$	—
Normalized Difference Vegetation Index (NDVI)	$\dfrac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{red}}}$	—
Soil-Adjusted Vegetation Index (SAVI)	$\left(\dfrac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{red}} + L}\right)(1 + L)$	L is an empirical constant, typically 0.5 for partial cover.

Even though vegetation indices can be used as features in classifications, they are commonly produced as an end-product indicating photosynthetic activity, particularly on a global scale from Advanced Very High-Resolution Radiometer (AVHRR) data.

Spectral Band Correlation

Spectral band correlation can result from several factors. First, the sensor spectral sensitivities sometimes overlap between adjacent spectral bands. Second, the spectral reflectance of most natural materials on the earth, particularly over spectral bandwidths of 10 nm or greater, change slowly with wavelength. Therefore, the reflectance in one band will be similar to that in an adjacent band. A notable exception is the "vegetation edge" at about 700 nm where the reflectance of photosynthetic vegetation increases dramatically from the red to the NIR spectral regions. Finally, topographic shading can introduce into remotely sensed data an apparent spectral correlation because it affects all solar reflective bands equally.

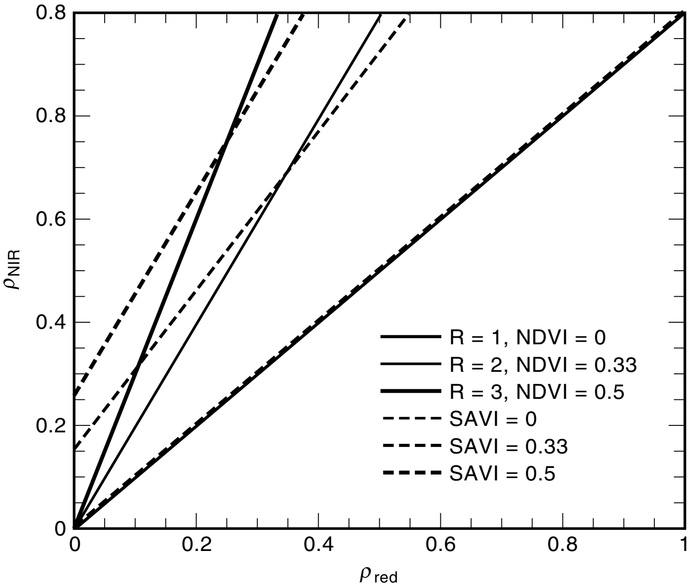

Figure 3. Isolines for three different vegetation indices in the NIR spectral reflectance space. The spectral ratio R and the NDVI are redundant in that either one can be expressed in terms of the other (see Table 2). The SAVI requires an empirically determined constant (4). A value of 0.5 is used for this graph and is appropriate under most conditions of partial vegetation cover with soil background. SAVI has a smaller slope than does NDVI in this graph. Therefore, SAVI is less sensitive to the ratio of the NIR reflectance than NDVI, reflecting the former's adjustment for soil background.

Principal Components. The Principal Component Transformation is often used to eliminate spectral band correlation. The PCT also produces a redistribution of spectral variance into fewer components, isolates spectrally uncorrelated signal components and noise, and produces features that, in some cases, align with physical variables. It is a data-dependent, linear matrix transform of the original spectral vectors into a new coordinate system that corresponds to a specific coordinate axes rotation in n-dimensions (2,5).

The PCT for a particular data set is derived from the eigenvalues and eigenvectors of the spectral covariance of the data, which is represented in matrix form as

$$\Sigma = \frac{1}{N-1}\sum_{j=1}^{N}(\mathbf{x}_j - \mu)(\mathbf{x}_j - \mu)^T \tag{2}$$

where N is the number of pixels in the image, $\mathbf{x}_j$ is the jth image data vector (pixel), the superscript T denotes the vector transpose, and μ is the vector mean value of the image given by

$$\mu = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j \tag{3}$$

The eigenvalues λ and eigenvectors ϕ of Σ are the solutions of the equation

$$\Sigma\phi = \lambda\phi \tag{4}$$

assuming ϕ is not the zero vector (6,7). The eigenvalues are ordered in decreasing order, and the corresponding eigenvectors are combined to form the eigenvector matrix

$$\Phi = [\phi_1\phi_2\cdots\phi_n] \tag{5}$$

The PCT is then given by

$$\mathbf{y} = \Phi^T\mathbf{x} \tag{6}$$

Each output axis is a linear combination of the input axes (e.g., the spectral bands) and is orthogonal to the other output axes (this characteristic can isolate uncorrelated noise in the original bands). The weights on the inputs $\mathbf{x}$ are the eigenvectors, and the variances of the output axes $\mathbf{y}$ are the eigenvalues. Because the eigenvalues are ordered in decreasing order, the PCT achieves a compression of data variation into fewer dimensions when a subset of PCT components corresponding to the larger eigenvalues is selected. A disadvantage of the PCT is that it is a global, data-dependent transform and must be recalculated for each image. The greatest computation burden is usually the covariance matrix for the input features. Figure 4 displays the first four principal components of the Landsat 5 TM scene displayed in Fig. 1.

Canonical Components. The Canonical Components Transform is similar to the PCT, except that the data are not lumped into one distribution in n-dimensional space when deriving the transformation matrix. Rather, training data for each class are used to find the transformation that maximizes the separability of the defined classes. A compression of significant information into fewer dimensions results, but it is not optimal as in the case of the PCT. Selection of the first three canonical components for a three-band color composite produces a color image that visually separates the classes better than any combination of three of the original bands.

The CCT is a linear transformation on the original feature space such that the transformed features are optimized and arranged in order of decreasing maximum separability of the classes. The optimization is accomplished based upon max-

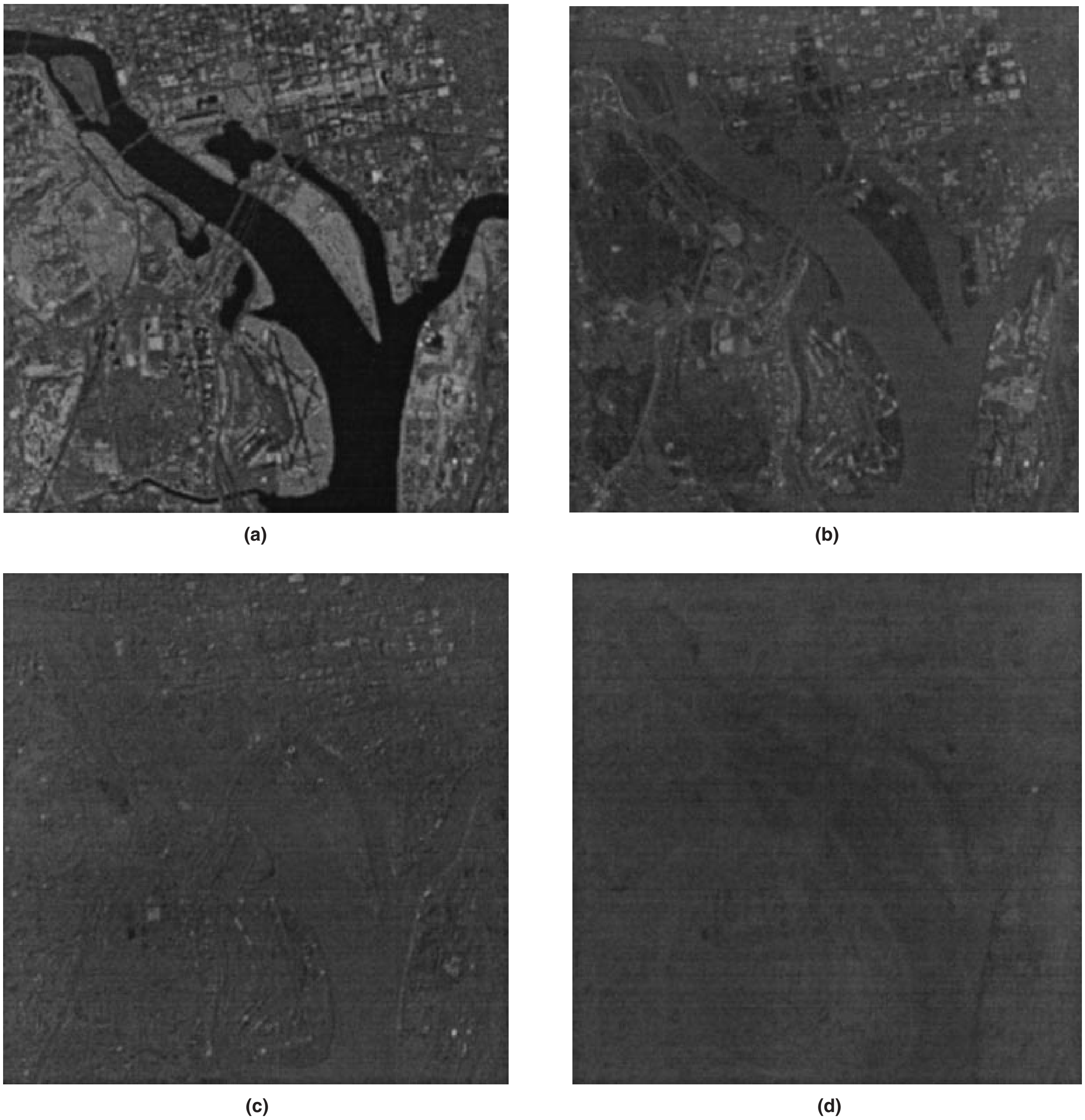

Figure 4. (a)–(d) The first four principal components from the PCT of the seven-band Landsat 5 TM image of Washington, DC (see Fig. 1). These four principal components contain 98.95% of the data variance contained in the original seven spectral bands.

imizing the ratio of the between-class variance to the within-class variance. The specific quantities are

$$\Sigma_{\mathrm{W}} = \sum_{i=1}^{n} P(\omega_i)\Sigma_i \qquad \text{(within-class scatter matrix)} \qquad (7)$$

$$\Sigma_{\mathrm{B}} = \sum_{i=1}^{n} P(\omega_i)(\mu_i - \mu_0)(\mu_i - \mu_0)^T \qquad \text{(between-class scatter matrix)} \qquad (8)$$

$$\mu_0 = \sum_{i=1}^{n} P(\omega_i)\mu_i \qquad (9)$$

where μ_i, Σ_i, and $P(\omega_i)$ are the mean vector, covariance matrix, and prior probability, respectively, for class ω_i. The optimality criterion then is defined as

$$J_1 = \mathrm{tr}(\Sigma_{\mathrm{W}}^{-1}\Sigma_{\mathrm{B}}) \qquad (10)$$

The transformation results in new features that are linear combinations of the original bands. The size of the feature eigenvalues indicates the relative class discrimination value. Thus, the size of the eigenvalues gives some idea as to how many features should be used.

Tasseled Cap Components. The Tasseled Cap Transform is a linear matrix transform, just as the PCT and CCT, but is fixed and independent of the data. It is, however, sensor dependent and must be newly derived for each sensor. The TCT produces a new set of components that are linear combinations of the original bands. The coefficients of the transformation matrix are derived relative to the "tasseled cap," which describes the temporal trajectory of vegetation pixels in the *n*-dimensional spectral space as the vegetation grows and matures during the growing season. The TCT was originally derived for crops in temperate climates, namely the U.S. Midwest, and is most appropriately applied to that type of data (8–11). For the Landsat MSS (Multispectral Scanner) data, four new axes are defined: soil brightness, greenness, yellow stuff, and non-such. For the Landsat TM (Thematic Mapper) data, six new axes are defined, soil brightness, greenness, wetness, haze, and an otherwise unnamed fifth and sixth axes. The transformed data in the tasseled cap space can be compared directly between sensors (e.g., Landsat MSS soil brightness and Landsat TM soil brightness).

Spectral Band Selection

Global satellite sensors must be designed to image a wide range of materials of interest in many different applications. The sensor design is thus a compromise for any particular application (continuous spectral sensing, such as that produced by hyperspectral sensors, is a way to provide data suitable for all applications, at the expense of large data volumes). A multispectral sensor may have bands in the red and NIR suitable for vegetation mapping but lack bands in the SWIR suitable for mineral mapping.

In spectral band selection, an optimal set of spectral bands are selected for analysis. The spectral characteristics of the material classes of interest must be defined before this technique is applied. The spectral characteristics are obtained from training data and may consist of the class mean vectors or the class mean vectors and covariances matrices (second-order statistics), depending on the metric to be used. Various band combinations can be compared to find the combination that best separates (distinguishes) the given classes.

Many separability metrics have been defined to measure separability. Each can be interpreted as a type of distance in spectral space (Table 3). The angular distance metric is particularly interesting because it conforms to the general shape of the scattergram between spectral bands in many cases. Topographic shading introduces a scatter of spectral signatures along a line through the origin of the spectral space. The angular metric directly measures angular separation of two distributions and is insensitive to the distance of a class distribution from the origin.

To select the optimum spectral bands from a sensor band set, an exhaustive calculation is performed to find the average interclass separability for each possible combination of bands. For example, bands 2, 3, and 4 of Landsat TM may show the highest average transformed divergence of any three-band combination of the seven TM bands for a vegetation and soil classification. The full classification can then be performed using only bands 2, 3, and 4.

MULTISPECTRAL IMAGE DATA CLASSIFICATION

Data classification is the process of associating a thematic label with elements of the data set. The data elements so labeled are typically individual pixels, but they may be groups of pixels that have been associated with one another, for example, by having previously segmented the scene into regions (i.e., spectrally homogeneous areas). Mathematically, the process of classification may be described as mapping the data from a vector-valued space (spectral feature space) to a scalar space that contains the list of final classes desired by the user (i.e., mapping from the data to the desired output).

Classification is carried out based upon ancillary information, often in terms of samples labeled by the analyst as being representative of each class of surface cover to be mapped. These samples are often called training samples or design samples. The development of an appropriate list of classes and the process of labeling these samples into these classes is a key step in the analysis process. A valid list of classes for a given data set must be, simultaneously,

1. exhaustive—There must be a logical and appropriate class to which to associate every pixel in the data set.
2. separable—It must be possible to discriminate accurately each class from the others in the list based on the spectral features available.
3. of informational value—The list of classes must contain the classes desired to be identified by the user.

Training Phase

Classification is typically carried out in two phases: the training phase and the analysis phase. During the training phase, ancillary information available to the analyst is used to define the list of classes to be used, and, from it, to determine the appropriate quantitative description of each of the classes. How this is done is situation-dependent, based on the form and type of ancillary information available and the desired classification output. In some cases, the analyst may have partial knowledge of the scene contents based upon observa-

Table 3. Separability Metrics for Classification (6,12)

Metric	Formula*	Remarks
City block	$L_1 = \lvert\mu_i - \mu_j\rvert$	Results in piecewise linear decision boundaries
Normalized city block	$\mathrm{NL}_1 = \sum_{b=1}^{n} \frac{\lvert m_{ib} - m_{jb}\rvert}{(\sigma_{ib} + \sigma_{jb})/2}$	Normalizes for class variance
Euclidean	$L_2 = \lVert\mu_i - \mu_j\rVert = [(\mu_i - \mu_j)^T(\mu_i - \mu_j)]^{1/2}$ $= \left[\sum_{b=1}^{n} (m_{ib} - m_{jb})^2\right]^{1/2}$	Results in linear decision boundaries
Angular	$\mathrm{ANG} = \mathrm{acos}\left(\frac{\mu_i^T \mu_j}{\lVert\mu_i\rVert\,\lVert\mu_j\rVert}\right)$	Normalizes for topographic shading
Mahalanobis	$\mathrm{MH} = \left[(\mu_i - \mu_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mu_i - \mu_j)\right]^{1/2}$	Assumes normal distributions; normalizes for class covariance; zero if class means are equal
Divergence	$D = \frac{1}{2}\,\mathrm{tr}[(\Sigma_i - \Sigma_j)(\Sigma_i^{-1} - \Sigma_j^{-1})]$ $+ \frac{1}{2}\,\mathrm{tr}\,[(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T]$	Zero if class means and covariances are equal; does not converge for large class separation
Transformed divergence	$D^t = 2[1 - e^{-D/8}]$	Asymptotically converges to one for large class separation
Bhattacharyya	$B = \frac{1}{8}\mathrm{MH} + \frac{1}{2}\ln\left[\frac{\lvert(\Sigma_i + \Sigma_j)/2\rvert}{(\lvert\Sigma_i\rvert\lvert\Sigma_j\rvert)^{1/2}}\right]$	Zero if class means and covariances are equal; does not converge for large class separation
Jeffries-Matusita	$\mathrm{JM} = [2(1 - e^{-B})]^{1/2}$	Asymptotically converges to one for large class separation

In the formulae, m_{ib} is the mean value for class i and band b, σ_{ib} is the standard deviation for class i and band b, μ_i is the mean vector for class i, Σ_i is the covariance matrix for class i, and n is the number of spectral bands.

tions from the ground and photointerpretation of air photographs of a part of the scene or from generalized knowledge of the area that is to be made more quantitative and specific by the analysis.

For example, the data set may be of an urban area in which the analyst is partially familiar and can designate in the data areas which are used for classes such as high-density housing, low-density housing, commercial, industrial, and recreational. The analyst would use this generalized knowledge to mark areas in the data set that are typical examples of each class. These then become the training areas from which the quantitative description of each class is calculated.

Examples of other types of ancillary data from which training samples may be identified are so-called signature banks, which are databases of spectral responses of the materials to be identified that were collected at another time and location with perhaps different instruments. In this case, the additional problem exists of reconciling the differences in data collection circumstances for the database with those of the data set to be analyzed. Examples of these circumstances are the differences in the instruments used to collect the data, the spatial and spectral resolution, the atmospheric conditions, the time of day, the illumination and direction of view variables, and the season.

Another example of an ancillary data source that might be used for deriving training data is more fundamental knowledge about the materials to be identified. For example, in a geological mapping problem, it might be known that certain minerals of interest have molecular absorption features as used by chemical spectroscopists to identify specific molecules. If such spectral features can be extracted from the data to be analyzed, they can be used to label training samples for such classes.

Analysis Phase

During the second phase of classification, the analysis phase, the pixel or region features are compared quantitatively to the class descriptions derived during the training phase to accomplish the mapping of each of the data elements to one of the defined classes. Classifiers may be of two types: relative and absolute. A relative classifier is one that assigns a data element to a class after having compared it to the entire list of classes to see to which class it is most similar. An absolute classifier compares the data element to only one class description to see if it is sufficiently similar to it. Generally speaking, in remote sensing, relative classifiers are the more common and more powerful.

Many different algorithms are used for classification in the analysis phase (6,12). A common approach for implementing a relative classifier is through the use of a so-called discriminant function. Designate the data element to be classified as vector **X**, in which the elements of the vector are the values measured for that pixel in each spectral band. Then, for a k-class situation, assume that we have k functions of **X**, $\{g_1(\mathbf{X}), g_2(\mathbf{X}), \ldots, g_k(\mathbf{X})\}$ such that $g_i(\mathbf{X})$ is larger than all others whenever **X** is from class i. Let ω_i denote the ith class. Then the classification rule can be stated as

$$\text{Decide } \mathbf{X} \text{ is in } \omega_i \text{ if and only if}$$
$$g_i(\mathbf{X}) \geq g_j(\mathbf{X}) \text{ for all } j = 1, 2, \ldots, k \quad (11)$$

The functions $g_i(\mathbf{X})$ are referred to as discriminant functions. An advantage of using this scheme is that it is easy to implement in computer software or hardware.

A common scheme for defining discriminant functions is to use the class probability density functions. The classification process then amounts to evaluating the value of each class density function at **X**. The value of a probability density function at a specific point is called the likelihood of that value. Such a classifier is called a maximum likelihood classifier because it assigns the data element to the most likely class.

Another example for a classification rule is the so-called Bayes rule strategy (13). Bayes' Theorem from the theory of probability states that

$$p(\omega_i|\mathbf{X}) = \frac{p(\mathbf{X}|\omega_i)}{p(\mathbf{X})} p(\omega_i) = \frac{p(\mathbf{X}, \omega_i)}{p(\mathbf{X})} \quad (12)$$

where $p(\omega_i|\mathbf{X})$ is the probability of class ω_i given the data element valued **X**, $p(\mathbf{X}|\omega_i)$ is the probability density function for class ω_i, $p(\omega_i)$ is the probability that class ω_i occurs, $p(\mathbf{X}, \omega_i)$ of the value **X** and the class ω_i, and $p(\mathbf{X})$ is the probability density function for the entire data set. Then, to maximize the probability of correct classification, one must select the class that maximizes $p(\omega_i|\mathbf{X})$. Because $p(\mathbf{X})$ is the same for any i, one may use as the discriminant function, just the numerator of Eq. (12), $p(\mathbf{X}|\omega_i)\, p(\omega_i)$. Thus, the classification rule becomes

$$\text{Decide } \mathbf{X} \text{ is in } \omega_i \text{ if and only if}$$
$$p(\mathbf{X}|\omega_i)\, p(\omega_i) \geq p(\mathbf{X}|\omega_j)\, p(\omega_j) \text{ for all } j = 1, 2, \ldots, k \quad (13)$$

This classification strategy leads to the minimum error rate. Note that if all the classes are equally likely, the $p(\omega_i)$ terms may be canceled and the Bayes rule strategy reduces to the maximum likelihood strategy. Because, in a practical remote sensing problem, the prior probabilities $p(\omega_i)$ are not known, it is common practice to assume equal priors.

Other factors that are significant in the analysis process are the matter of how the class probability density functions are modeled and, related to this, how many training samples are available by which to train the classifier. Parametric models, assuming that each class is modeled by one or a combination of Gaussian distributions, are very common and powerful. Within this framework, one can also make various simplifying assumptions. Some common ones, in parametric form, and the corresponding discriminant functions follow:

- Assume that all classes have the same covariance, in which there is no correlation between bands, and that all bands have unit variance:

  $$g_i(\mathbf{X}) = (\mathbf{X} - \mu_i)^T (\mathbf{X} - \mu_i) \quad (14)$$

  The decision boundary that results is linear in spectral feature space and is oriented perpendicular to the line connecting the class mean values at the midpoint of the line. This is the minimum-distance-to-means classifier.
- Assume that all classes have the same covariance but account for correlation between bands and for different variances in each:

  $$g_i(\mathbf{X}) = (\mathbf{X} - \mu_i)^T \Sigma^{-1} (\mathbf{X} - \mu_i) \quad (15)$$

  The resulting decision boundary in spectral feature space is linear, but its orientation and location are dependent upon the common covariance Σ.
- Assume that classes have different covariances:

  $$g_i(\mathbf{X}) = -\tfrac{1}{2} \ln |\Sigma_i| - \tfrac{1}{2} (\mathbf{X} - \mu_i)^T \Sigma_i^{-1} (\mathbf{X} - \mu_i) \quad (16)$$

  The resulting decision boundary in spectral space is a second-order hypersurface whose shape and location is dependent upon the individual mean vectors μ_i and covariance matrices Σ_i. This is the maximum likelihood classifier.
- Assume that the class densities have a more complex structure such that a combination of a small number of Gaussian densities is not adequate:

  $$g_i(\mathbf{X}) = \frac{1}{N_i} \sum K\left(\frac{\mathbf{X} - \mathbf{X}_{ji}}{\lambda}\right) \quad (17)$$

  The resulting decision boundary in spectral space can be of nearly arbitrary shape.

It can be seen that this list of discriminant functions has a steadily increasing generality and a steadily increasing complexity, such that a rapidly increasing number of training samples is required to adequately estimate the rapidly growing number of parameters in each. The latter one, for example, which, though it is still parametric in form, is referred to as a nonparametric Parzen density estimator with kernel K. The kernel function K, as well as the number of kernal terms to be used N_i, is selectable by the analyst. For example, one possible selection is a Gaussian-shaped function, thus making this discriminant function a direct generalization of the previous ones.

There are many additional variations to this list of discriminant functions. There are also additional variations to the possible training procedures. For example, one variation that is popular at the present time is the neural network method. This method uses an iterative scheme for determining the location of the decision boundary in spectral feature space. A network is designed, consisting of as many inputs as there are spectral features, as many outputs as there are classes, and threshold devices with weighting functions connecting the inputs to the outputs. Training samples are applied to the input sequentially, and the resulting output for each is observed. If the correct classification is obtained for a given sample, as evidenced by the output port for the correct class being the largest, the weights for correct output are augmented, and incorrect class output weights are diminished. The training set is reused for as many times as necessary to obtain good classification results.

The advantage of this approach is its generality and that it can be essentially automatic. Characteristics generally regarded as disadvantages are that it is nearly entirely heuristic, thus making analytical calculations and performance predictions difficult, its generality means that very large training sets are required to obtain robust performance, and there is a great deal of computation required in the training process. Because, in practical circumstances, classifiers must be retrained for every new data set, characteristics affecting the training phase are especially significant.

Unsupervised Classification

A second form of classification that finds use in remote sensing is unsupervised classification, also known as clustering. In this case, data elements, usually individual pixels, are assigned to a class without the use of training samples, thus the "unsupervised" name. The purpose of this type of classification is to assign pixels to a group whose members have similar spectral properties (i.e., are near to one another in spectral feature space). There are, again, many algorithms to accomplish this. Generally speaking, three capabilities are needed.

1. A measure of distance between points. Euclidean distance is a common choice.
2. A measure of distance or separability between the sets of points comprising each cluster. Any separability measure, such as listed in Table 3 could be used, but usually simpler measures are selected.
3. A cluster compactness criterion. An example might be the sum of squared distances from the cluster center for all pixels assigned to a cluster.

The process typically begins when one selects (often arbitrarily) a set of initial cluster centers and then assigns each pixel to the nearest cluster center using step 1. After assigning all the pixels, one computes the new cluster centers. If any of the cluster centers have moved, all the pixels are reassigned to the new cluster centers. This iterative process continues until the cluster centers do not move or the movement is smaller than a prescribed threshold. Then steps 2 and 3 are used to test if the clusters are sufficiently distinct (separated from one another) and compact. If they are not adequately distinct, the two that are the closest are combined, and the process is repeated. If they are not sufficiently compact, an additional cluster center is created within the least distinct cluster, and the process is repeated.

Clustering is ordinarily not useful for final classification as such because it is unlikely that the data would be clustered into classes of specific interest. Rather it is primarily useful as an intermediate processing step. For example, in the training process, it is often used to divide the data into spectrally homogenous areas that might be useful in deciding on supervised classifier classes and subclasses and in selecting training samples for these classes and subclasses.

CLASSIFICATION USING NEURAL NETWORKS

Unlike statistical, parametric classifiers, Artificial Neural Network (ANN) classifiers rely on an interative error minimization algorithm to achieve a pattern match. A network consists of interconnected input (feature) nodes, hidden layer nodes, and output (class label) nodes. A wide range of network architectures have been proposed (14); here a simple three-layer network is considered to explain the basic operation. The input nodes do no processing but simply provide the paths for the data into the hidden layer. Each input node is connected to each hidden layer node by a weighted link. In the hidden layer, the weighted input features are summed and compared to a thresholding decision function. The decision function is usually "soft," with a form known as a sigmoid,

$$\text{output}(\text{input}) = \frac{1}{1 + \exp(-\text{input})} \tag{18}$$

The output from each hidden layer node is then fed through a weighted link to each output layer node. The same processing, summation and comparison to a threshold, is performed in each output node. The output node with the highest resulting value is selected as the label for the input feature vector.

The decision information of the ANN is contained in its weights. To adapt the weights to the data, an iterative algorithm is required. The classic example is the Back Propagation (BP) algorithm (15,16). The BP algorithm minimizes the output error over all classes for a given set of training data. It achieves this by measuring the output error and adjusting the ANN's link weights progressively backward through each layer to reduce the error. If local minima in the decision space of the ANN can be avoided, the BP algorithm will converge to a global minimum for the output error [although one is never sure that it is not in reality a local minimum (i.e., the algorithm cannot be proven to result in a global error minimum)]. Other convergence algorithms, such as Radial Basis Functions, have been used and are faster than BP.

One parameter that must be set for ANNs is the number of hidden layer nodes. A way to specify this is to relate the total number of Degrees-Of-Freedom (DOF) in the ANN to that of another classifier for comparison (2). For example, in a three-layer ANN, the DOF are

$$N_{\text{ANN}} = H(K + L) \tag{19}$$

where H is the number of hidden layer nodes, K is the number of input features, and L is the number of output classes. For the same number of features and classes, the ML classifier has the following DOF:

$$N_{\text{ML}} = \frac{LK(K+3)}{2} \tag{20}$$

Therefore, to compare the two classifiers, it is logical to set their DOF equal, obtaining

$$H = \frac{LK(K+3)}{2(K+L)} \tag{21}$$

for the number of hidden layer nodes in the ANN. This analysis yields only 20 hidden layer nodes for six bands of nonthermal TM imagery, even for as many as 20 classes. Fewer hidden layer nodes result in faster BP training.

Performance Comparison to Statistical Classifiers

The ANN type of classifier has some unique characteristics that are important in comparing it to other classifiers:

1. Because the weights are initially randomized, the final output results of the ANN are stochastic (i.e., they will vary from run to run on the same training data). It has been estimated that this variation is as much as 5% (17).
2. The decision boundaries move in the feature space to reduce the total output error during the optimization process. The network weights and final classification map that result will depend on when the process is terminated.

The ANN classifier is nonparametric (i.e., it makes no assumptions about an underlying statistical distribution for each class). In contrast, the ML classifier assumes a Gaussian distribution for each class. These facts make the feature space decision boundaries totally different. It appears that the boundaries from a three-layer ANN trained with the BP algorithm are often more similar to those from the minimum-distance-to-means classifier than to those from the ML classifier. Experiments with a land-use/land-cover classification involving heterogeneous class spectral signatures indicate that the nonparametric characteristic of the ANN classifier results in superior classifications (18).

IMAGE SEGMENTATION

Image segmentation is a partitioning of an image into regions based on the similarity or dissimilarity of feature values between neighboring image pixels. It is often used in image analysis to exploit the spatial information content of the image data. Most image segmentation approaches can be placed in one of three categories (19):

1. characteristic feature thresholding or clustering,
2. boundary detection, or
3. region growing.

Characteristic feature thresholding or clustering does not exploit spatial information. The unsupervised classification (clustering) approaches discussed previously are a form of this type of image segmentation. Boundary detection exploits spatial information by examining local edges found throughout the image. For simple noise-free images, detection of edges results in straightforward boundary delineation. However, edge detection on noisy, complex images often produces missing edges and extra edges that cause the detected boundaries to not necessarily form a set of closed connected curves that surround connected regions. Image segmentation through region growing uses spatial information and guarantees the formation of closed, connected regions. However, it can be a computationally intensive process.

Edge Detection

Edge detection approaches generally examine pixel values in local areas of an image and flag relatively abrupt changes in pixel values as *edge pixels*. These edge pixels are then extended, if necessary, to form the boundaries of regions in an image segmentation.

Derivative-Based Methods for Edge Detection. The simplest approaches for finding abrupt changes in pixel values compute an approximation of the gradient at each pixel. The mathematical definition of the gradient of the continuous function $f(x, y)$ is

$$\nabla f(x,y) = \left(\frac{\partial f}{\partial x}(x,y), \frac{\partial f}{\partial y}(x,y) \right) \tag{22}$$

where $\nabla f(x, y)$ is the gradient at position (x, y), and $(\partial f/\partial x)(x, y)$ and $(\partial f/\partial y)(x, y)$ are the first derivatives of the function $f(x, y)$ with respect to the x and y coordinates, respectively. The gradient magnitude is

$$|\nabla f(x,y)| = \sqrt{\left[\frac{\partial f}{\partial x}(x,y)\right]^2 + \left[\frac{\partial f}{\partial y}(x,y)\right]^2} \tag{23}$$

and the gradient direction (angle) is

$$\phi = \arctan \left[\frac{\dfrac{\partial f}{\partial x}(x,y)}{\dfrac{\partial f}{\partial y}(x,y)} \right] \tag{24}$$

In order to apply the concept of a mathematical gradient to image processing, $(\partial f/\partial x)(x, y)$ and $(\partial f/\partial y)(x, y)$ must be approximated by values on a discrete lattice corresponding to the image pixel locations. Such a simple discretization is

$$\frac{\partial f}{\partial x}(x,y) \cong f(x+1,y) - f(x,y) \tag{25}$$

for edge detection in the x direction and

$$\frac{\partial f}{\partial y}(x,y) \cong f(x,y+1) - f(x,y) \tag{26}$$

for edge detection in the y direction. These functions are equivalent to convolving the image with one of the two templates in Fig. 5, where (x, y) is the upper left corner of the window.

−1	1
0	0

$\frac{\partial f}{\partial x}(x, y)$

−1	0
1	0

$\frac{\partial f}{\partial y}(x, y)$

Figure 5. Convolution templates corresponding to the discretized first deriviative of the image function $f(x, y)$ in the x and y directions. These templates can be used as image edge detectors. However, their small 2 × 2 window size makes these templates very susceptible to noise.

–1	0	1
–2	0	2
–1	0	1

$\frac{\partial f}{\partial x}(x, y)$

1	2	1
0	0	0
–1	–2	–1

$\frac{\partial f}{\partial y}(x, y)$

Sobel template

–1	0	1
–1	0	1
–1	0	1

$\frac{\partial f}{\partial x}(x, y)$

1	1	1
0	0	0
–1	–1	–1

$\frac{\partial f}{\partial y}(x, y)$

Prewitt template

Figure 6. The Sobel and Prewitt edge detection templates. These 3 × 3 window templates are somewhat less susceptible to noise as compared to the 2 × 2 window templates illustrated in Fig. 5.

A disadvantage of this and other similar [e.g., Roberts template (20)] approximations of the gradient function is that the small 2 × 2 window size makes them very susceptible to noise. Somewhat less susceptible to noise are the 3 × 3 window templates devised by Sobel [see Duda and Hart (21)] and Prewitt (22), which are illustrated in Fig. 6.

The edge detection templates given in Figs. 5 and 6 are approximations of an image gradient or a discretation of the first derivative of the image function. The second derivative of the image function, called the Laplacian operator, can also be used for edge detection. Whereas the first derivative produces positive or negative peaks at and image edge, the second derivative produces a zero value at the image edge, surrounded closely by positive and negative peaks. Edge detection then reduces to detecting these “zero-crossing” values from the Laplacian operator. For a continuous function $f(x, y)$, the Laplacian operator is defined as

$$\nabla^2 f(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} \tag{27}$$

The usual discrete approximation is

$$\nabla^2 f(x,y) = 4f(x,y) - f(x-1,y) - f(x,y+1) - f(x+1,y) \tag{28}$$

This can be represented by convolving a two-dimensional image with the image template shown in Fig. 7. Note that the Laplacian operator is directionally symmetric.

Image Filtering for Edge Detection. All these methods for edge detection are intrinsically noise sensitive (some more than others) because they are based upon differences between pixels in local areas of the image. Marr and Hildreth (23) suggested the use of Gaussian filters with relatively large window sizes to remove noise in images. Combining the Gaussian filter with the Laplacian operator yields the Laplacian of Gaussian (LOG) function

$$\nabla^2 G(x,y) = \left\{\frac{1}{2\pi\sigma^4}\right\} \cdot \left\{\left(\frac{x^2+y^2}{\sigma^2}\right) - 1\right\} \exp\left\{\frac{-(x^2+y^2)}{2\sigma^2}\right\} \tag{29}$$

where σ controls the amount of smoothing provided by the filter.

0	–1	0
–1	4	–1
0	–1	0

Figure 7. The Laplacian edge detection template. This edge detection template is the discretized second derivative of the image function $f(x, y)$. This operator produces a zero value at image edges, which is surrounded closely by positive and negative peaks.

Edge detection through convolving the image with the LOG function and searching for zero-crossing locations is less sensitive to noise than the previously discussed methods. Even more sophisticated filtering and edge location techniques have been devised. These techniques were unified in a paper by Shen and Castan (24), in which they derive the optimal filter for the multi-edge case.

Region Growing

Region growing is a process by which image pixels are merged with neighboring pixels to form regions, based upon a measure of similarity between pixels and regions. The basic outline of region growing follows (25–27):

1. Initialize by labeling each pixel as a separate region.
2. Merge all spatially adjacent pixels with identical feature values.
3. Calculate a similarity or dissimilarity criterion between each pair of spatially adjacent regions.
4. Merge the most similar pair of regions.
5. Stop if convergence has been achieved; otherwise, return to step 3.

Beaulieu and Goldberg (25) describe a sequential implementation of this algorithm in which step 3 is kept efficient through updating only those regions involved in or adjacent to the merge performed in step 4. Tilton (26) describes a parallel implementation of this algorithm in which multiple merges are allowed in step 4 (best merges are performed in image subregions) and the (dis)similarity criterion in step 3 is calculated in parallel for all regions. Schoenmakers (27) simultaneously merges all region pairs with minimum dissimilarity criterion value in step 4.

The similarity or dissimilarity criterion employed in step 3 should be tailored to the type of image being processed. A simple criterion that has been used effectively with remotely sensed data is the Euclidean spectral distance (27), as in Table 3. Other criteria that have been employed are the Normalized Vector Distance (28), criteria based on minimizing the mean-square error or change in image entropy (29), and a criterion based on minimizing a polynomial approximation error (25).

Clear-cut convergence criteria have not been developed for region growing segmentation. Simple criteria that are satisfactory in some applications are the number of regions or a ratio of number of regions to the total number of image pixels. Direct thresholding on the dissimilarity criterion value (i.e., perform no merges between regions with a dissimilarity criterion value greater than a threshold) has also been used with mixed results. More satisfactory results have been obtained by defining convergence as the iteration prior to the iteration at which the maximum change in dissimilarity criterion value occurred.

Extraction and Classification of Homogeneous Objects

An image segmentation followed by a maximum likelihood classification is the basic idea behind the Extraction and Classification of Homogeneous Objects (ECHO) classifier (30,31). The segmentation scheme used by ECHO was designed for speed on the computers of mid-1970s and could be replaced by a segmentation approach of more recent vintage. However, the formalization of the maximum likelihood classification for image regions (objects) is still appropriate. For single pixels, the maximum likelihood decision rule is

$$\text{Decide } \mathbf{X} \text{ is in } \omega_i \text{ if and only if}$$
$$p(\mathbf{X}|\omega_i) \geq p(\mathbf{X}|\omega_j) \text{ for all } j = 1, 2, \ldots, k \qquad (30)$$

The rule is just Eq. (13) with $p(\omega_j) = 1$. Suppose that an image region consists of m pixels. To apply the maximum likelihood decision rule to this region, $\mathbf{X}$ must be redefined to include the entire region, that is, $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\}$. The evaluation of $p(\mathbf{X}|\omega_i)$, where $\mathbf{X}$ is redefined as a collection of pixels, is very difficult. However, this collection of pixels belongs to a homogeneous region. In this case, it is reasonable to assume that the pixels are statistically independent. This assumption allows the evaluation of $p(\mathbf{X}|\omega_i)$ as the product

$$p(\mathbf{X}|\omega_i) = p(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m|\omega_i) = \prod_{j=1}^{m} p(\mathbf{X}_j|\omega_i) \qquad (31)$$

Split and Merge

Seeking more efficient methods for region-based image segmentation has led to the development of split-and-merge approaches. Here the image is repeatedly subdivided until each resulting region has a minimum homogeneity. After the region-splitting process converges, the regions are grown as previously described. This approach is more efficient when large homogenous regions are present. However, some segmentation detail may be lost. See Cross et al. (32) for an example of split-and-merge image segmentation.

Hybrids of Edge Detection and Region Growing

A number of approaches have been offered for combining edge detection and region growing. Pavlidis and Liow (33) perform a split-and-merge segmentation such that an oversegmented result is produced and then eliminate or modify region boundaries based on general criteria including the contrast between the regions, region boundary smoothness, and the variation of the image gradient along the boundary. LeMoigne and Tilton (34) use region growing to generate a hierarchical set of image segmentations and make local selections of the best level of segmentation detail based on edges produced by an edge detector.

Hybrids of Spectral Clustering and Region Growing

Tilton (35) has recently demonstrated the potential of a hybrid of spectral clustering and region growing. In this approach, spectral clustering is performed in between each region growing iteration. The spectral clustering is constrained to merge regions that are at least as similar as the last pair of regions merged by region growing, and is not allowed to merge any spatially adjacent regions. This approach to image segmentation is very computationally intensive. However, practical processing times have been achieved by a recursive implementation on a cluster of 64 Pentium Pro PCs configured as a Beowulf-class parallel computer (35,36).

HYPERSPECTRAL DATA

Hyperspectral Data Normalization

Hyperspectral imagery contains significantly more spectral information than does multispectral imagery such as that from Landsat TM. Imaging spectrometers produce hundreds of spectral band images, with narrow (typically 10 nm or less) contiguous bands across a broad wavelength range (e.g., 400–2400 nm). Also, such new sensor systems are capable of generating more precise data radiometrically, with signal-to-noise ratios justifying 10 or more bit data systems (1024 or more shades of gray per band), as compared to 6 or 8 bit precision in previous systems. This potentially high precision requires concomitant substantially improved calibration for atmospheric, solar, and topographic effects, particularly if comparisons are to be made to laboratory or field reflectance spectra for classification.

To convert remote sensing data to reflectance, one must first correct for the additive and multiplicative factors in Eq. (1). Even though in some circumstances (e.g., multitemporal analysis) this correction may be useful for all spectral data, it is especially critical for hyperspectral imagery when the intention is to use narrow band spectral absorption features in a deterministic sense because

1. narrow atmospheric absorption bands have a severe effect on corresponding sensor bands, and
2. some algorithms for physical constituent estimation, either in the atmosphere or on the Earth's surface, require precise measurements of absorption band locations, widths, and depths.

The computation burden for calibration is, of course, much larger for hyperspectral imagery than it is for multispectral imagery.

In one effective calibration technique, the *empirical line* method (37), the sensor values are linearly correlated to field reflectance measurements. In this single process, all the coefficients in Eq. (1) are determined except for topographic shading. Obtaining field reflectance measurements is difficult and expensive at best, so a number of indirect within-scene approaches have also been used.

An example of the use of within-scene information to achieve a partial calibration of hyperspectral imagery is termed *flat-fielding* (38). An object that can be assumed spectrally uniform and with high radiance ("white" in a visual sense) must be located within the scene. Its spectrum as seen by the sensor contains the atmospheric transmittance terms of Eq. (1). If the data are first corrected for the haze level, or if it can be ignored (at longer wavelengths such as in the NIR and SWIR), then a division of each pixel's spectrum by the bright reference object's spectrum will tend to cancel the solar irradiance and atmospheric transmittance factors in Eq. (1). An example is shown in Fig. 8. The data are from an Airborne Visible-InfraRed Imaging Spectrometer (AVIRIS) flight over Cuprite, Nevada, in 1990. The mineral kaolinite contains a doublet absorption feature centered at about 2180 nm, which is masked in the at-sensor radiance data by the downward trend of the solar irradiance. An atmospheric carbon dioxide absorption feature can also be seen at 2060 nm. After the flat-field operation, the relative spectral reflectance closely matches a sample reflectance curve in shape, including no atmospheric absorption features. The reflectance magnitude does not agree because the flat-field correction does not correct for topographic shading [the cosine term in Eq. (1)].

After the hyperspectral data are normalized in this way, it is possible to characterize the surface absorption features by such parameters as their location, depth (relative to a *continuum,* which is a hypothetical curve with no absorption features), and width (Fig. 9). These features can be used to distinguish one mineral (or any other material with narrow absorption features) from another (39). The feature extraction algorithms first detect local minima in the spectral data using operations such as a spectral derivative and then calculate the depth and width of those minima. Zero crossings in second-order derivatives and a spectral scale-space can also be used to detect and measure significant spectral absorption features (40).

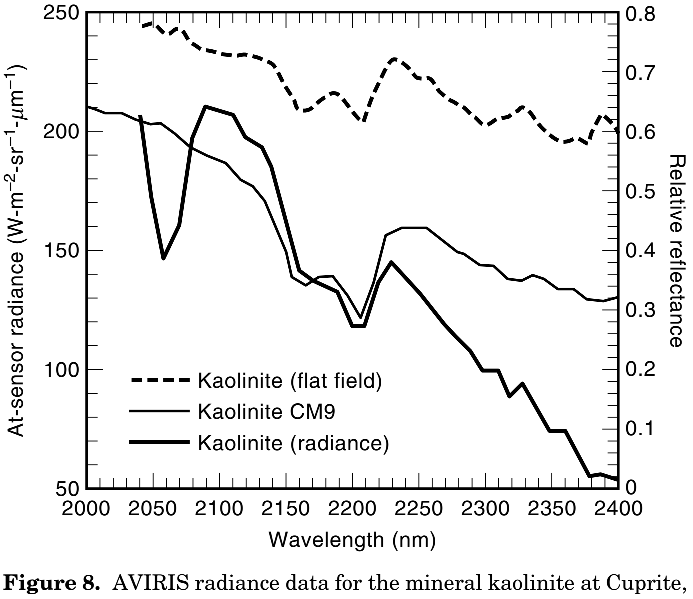

Figure 8. AVIRIS radiance data for the mineral kaolinite at Cuprite, Nevada, before and after flat-field normalization, compared to spectral reflectance data from a mineral reflectance library (sample designated CM9). It is evident that the normalization process produces a spectral signal from the image radiance data that more closely matches the shape of the spectral reflectance curve. If classification of image radiance data is performed with library spectral reflectance as the reference signal, either an empirical normalization of this type, or a difficult calibration of the sensor radiance data to reflectance would be required.

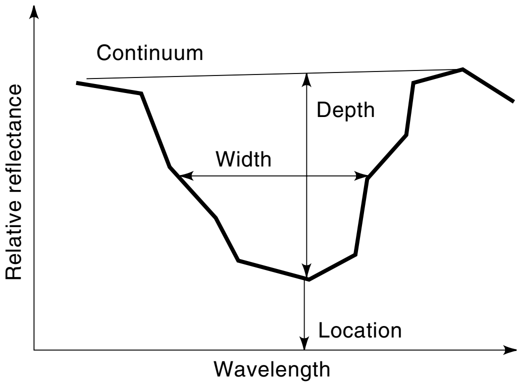

Figure 9. Definition of spectral absorption features in hyperspectral data. After an absorption band is detected in the spectral data, these features can be measured and compared with the same features derived either from labeled training pixels within the image itself or from library spectral reflectance data. For surface materials with characteristic absorption bands, this approach can considerably reduce the amount of computation required for classification of hyperspectral imagery.

Classification and Analysis of Hyperspectral Data

Data in higher-dimensional spaces have substantially different characteristics than that in three-dimensional space, such that the ordinary rules of geometry of three-dimensional space do not apply. For example, two class distributions can lie right on top of one another, in the sense of having the same mean values and yet they may be perfectly separable by a well-designed classifier.

Examples of these differences of data in high-dimensional space follow (41). As dimensionality increases,

1. The volume of a hypercube concentrates in the corners.
2. The volume of a hypersphere concentrates in an outside shell.
3. The diagonals are nearly orthogonal to all coordinate axis.

When data sets contain a large number of spectral bands or features, more than 10 or so, the ability to discriminate between classes with higher accuracy and to derive greater information detail increases substantially, but some additional aspects become significant in the data analysis process in order to achieve this enhanced potential. For example, as the data dimensionality increases, the number of samples necessary to define the class distributions to adequate precision increases very rapidly. Furthermore, both first- and second-order statistics are significant in achieving optimal separability of classes.

The fact that second-order statistics are significant, in addition to first-order statistics, tends to exacerbate the need for a larger number of training samples. For example, if one were to attempt to analyze a 200-dimensional data set at full dimensionality using conventional estimation methods, many thousands of samples may be necessary in order to obtain the full benefit of the 200 bands. Rarely would this number of samples be available.

Hyperspectral Feature Extraction

Quantitative feature extraction methods are especially important because of the large number of spectral bands on the one hand and the significantly enhanced amount and detail of information that is potentially extractable from such data on the other. Given the large number of spectral bands in such data, feature selection, choosing the best subset of bands of size m from a complete set of N bands, quickly becomes intractable or impossible. For example, to choose the best subset of bands of size 10 out of 100 bands, there are more than 1.7×10^{13} possible subsets of size 10 that must be examined if the optimum set is to be determined.

It is possible to avoid working directly at such high dimensionality without a penalty in classifier performance and with a substantial improvement in computational efficiency. This is the case because, as implied by the preceding geometric characteristics, the volume in a hyperdimensional feature space increases very rapidly as the dimensionality increases. A result of this is that, for remote sensing problems, such a space is mostly empty, and the important data structure in any given problem will exist in a subspace. The particular subspace is very problem-dependent and is different for every case. Thus, if one can determine which subspace is needed for the problem at hand, one can have available all the separability that the high-dimensional data can provide, but with reduced need for training set size and reduced amounts of computation. The problem then becomes focused on finding the correct subspace containing this key data structure.

Feature extraction algorithms may be used for this purpose. The CCT introduced earlier is one of several approaches that is suitable for extraction features from hyperspectral data. Each approach tends to have unique advantages and some disadvantages. The CCT, for example, is a relatively fast calculation. However, it does not perform well when the classes to be discriminated have only a small difference in their mean values, and it produces predictably useful features only up to one less than the number of classes to be separated. It has another disadvantage common to many such algorithms, namely that it depends on parameters, class means, and covariances, which must be estimated from the training samples at full dimensionality. Thus, it may produce a suboptimal feature subspace resulting from the imprecise estimation of the class parameters.

Another possible scheme is the Decision Boundary Feature Extraction algorithm (42). This scheme does not have the disadvantages of CCT. However, it tends to be a lengthy calculation because it is based directly upon the training samples instead of the class statistics derived from them.

A very fortunate characteristic of high-dimensional spaces is that for most high-dimensional data sets, lower-dimensional linear projections tend to be normally distributed or with a combination of normal distributions. This tends to add to the credibility of using a Gaussian model for the classification process, reduces the need to consider nonparametric schemes, and reduces the need to consider the use of higher-order statistics.

Thus, based upon the algorithms referred to previously, one can expect to do a very effective analysis of high-dimensional multispectral data and, in a practical circumstance, achieve a near to optimal extraction of desired information with performance substantially enhanced over that possible with more conventional multispectral data.

BIBLIOGRAPHY

1. J. C. Tilton, Image segmentation by iterative parallel region growing with applications to data compression and image analysis, *Proc. 2nd Symp. Frontiers Massively Parallel Computat.,* pp. 357–360, Fairfax, VA, 1988.
2. R. A. Schowengerdt, *Remote Sensing—Models and Methods for Image Processing,* 2nd ed. Chestnut Hill, MA: Academic Press, 1997.
3. A. Berk, L. S. Bernstein, and D. C. Robertson, *MODTRAN: A Moderate Resolution Model for LOWTRAN 7,* U.S. Air Force Geophysics Laboratory, No. GL-TR-89-0122, 1989.
4. A. R. Huete, A soil adjusted vegetation index (SAVI), *Remote Sens. Environ.,* **25**: 295–309, 1988.
5. J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction,* 2nd ed. Berlin: Springer-Verlag, 1993.
6. K. Fukunaga, *Introduction to Statistical Pattern Recognition,* 2nd ed., Boston: Academic Press, 1990.
7. G. W. Stewart, *Introduction to Matrix Computations,* New York: Academic Press, 1973.
8. R. J. Kauth and G. S. Thomas, The Tasselled Cap—A graphic description of spectral-temporal development of agricultural crops as seen by Landsat, *Proc. 2nd Int. Symp. Remotely Sensed Data,* **4B**: 41–51, Purdue University, West Lafayette, IN, 1976.
9. D. R. Thompson and O. A. Whemanen, Using Landsat digital data to detect moisture stress in corn-soybean growing regions, *Photogrammetric Eng. Remote Sens.,* **46**: 1087-1093, 1980.
10. E. P. Crist and R. C. Cicone, A physically-based transformation of Thematic Mapper data—the TM Tasseled Cap, *IEEE Trans. Geosci. Remote Sens.,* **GE-22**: 256–263, 1984.
11. E. P. Crist, R. Laurin, and R. C. Cicone, Vegetation and soils information contained in transformed Thematic Mapper data. *Proc. 1986 Int. Geosci. Remote Sens. Symp.,* pp. 1465–1470, Zurich, 1986.
12. P. H. Swain and S. M. Davis, *Remote Sensing: The Quantitative Approach,* New York: McGraw-Hill, 1978.
13. A. Papoulis, *Probability, Random Variables, and Stochastic Processes,* Tokyo: McGraw-Hill, 1984.
14. J. D. Paola and R. A. Schowengerdt, A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery, *Int. J. Remote Sens.,* **16**: 3033–3058, 1995.
15. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, eds., *Parallel Distributed Processing: Explorations in the Microstruction of Cognition,* Vol. I, pp. 318–362, Cambridge, MA: MIT Press, 1986.
16. R. P. Lippmann, An introduction to computing with neural nets, *IEEE ASSP Magazine,* **4** (2): 4–22, 1987.
17. J. D. Paola and R. A. Schowengerdt, The effect of neural network structure on a multispectral land-use classification, *Photogrammetric Eng. Remote Sens.,* **63**: 535–544, 1997.
18. J. D. Paola and R. A. Schowengerdt, A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification, *IEEE Trans. Geosci. Remote Sens.,* **33**: 981-996, 1995.

19. K. S. Fu and J. K. Mui, A survey on image segmentation, *Pattern Recognition,* **13**: 3–16, 1981.
20. L. G. Roberts, Machine perception of three dimensional solids, *Proc. Symp. Optical Electro-Optical Image Processing Technol.,* pp. 159–197, Cambridge, MA: MIT Press, 1965.
21. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York, Wiley, 1973.
22. J. M. S. Prewitt, Object enhancement and extraction. In B. S. Lipkin and Rosenfeld, eds., *Picture Processing and Psychopictorics,* pp. 75–149. New York: Academic Press, 1970.
23. D. Marr and E. Hildreth, Theory of edge detection, *Proc. Roy. Soc. London,* **B 207**: 187–217, 1980.
24. J. Shen and S. Castin, Towards the unification of band-limited derivative operators for edge detection, *Signal Process.,* **31**: 103–119, 1993.
25. J.-M. Beaulieu and M. Goldberg, Hierarchy in picture segmentation: A stepwise optimization approach, *IEEE Trans. Pattern Anal. Mach. Intell.,* **11**: 150–163, 1989.
26. J. C. Tilton, Image segmentation by iterative parallel region growing and splitting, *Proc. 1989 Int. Geosci. Remote Sens. Symp.,* pp. 2235–2238, Vancouver, Canada, 1989.
27. R. P. H. M. Schoenmakers, *Integrated Methodology for Segmentation of Large Optical Satellite Image in Land Applications of Remote Sensing,* Agriculture series, Catalogue number : CL-NA-16292-EN-C, Luxembourg: Office for Official Publications of the European Communities, 1995.
28. A. Baraldi and F. Parmiggiani, A neural network for unsupervised categorization of multivalued input parameters: An application to satellite image clustering, *IEEE Trans. Geosci. Remote Sens.,* **33**: 305–316, 1995.
29. J. C. Tilton, Experiences using TAE-Plus command language for an image segmentation program interface, *Proc. TAE Ninth Users' Conf.,* New Carrollton, MD, pp. 297–312, 1991.
30. R. L. Kettig and D. A. Landgrebe, Computer classification of remotely sensed multispectral image data by Extraction and Classification of Homogeneous Objects, *IEEE Trans. Geosci. Electron.,* **GE-14**: 19–26, 1976.
31. D. A. Landgrebe, The development of a spectral-spatial classifier for Earth observational data, *Pattern Recognition,* **12**: 165–175, 1980.
32. A. M. Cross, D. C. Mason, and S. J. Dury, Segmentation of remotely sensed image by a split-and-merge process, *Int. J. Remote Sens.,* **9**: 1329–1345, 1988.
33. T. Pavlidis and Y.-T. Liow, Integrating region growing and edge detection, *IEEE Trans. Pattern Anal. Mach. Intell.,* **12**: 225–233, 1990.
34. J. Le Moigne and J. C. Tilton, Refining image segmentation by integration of edge and region data, *IEEE Trans. Geosci. Remote Sens.,* **33**: 605–615, 1995.
35. J. C. Tilton, Image segmentation by region growing and spectral clustering with natural convergence criteria, *Proc. 1998 Int. Geosci. Remote Sens. Symp.,* Seattle, Washington, 1998.
36. D. J. Becker et al., Beowulf: A parallel workstation for scientific computation, *Proc. Int. Conf. Parallel Process.,* 1995.
37. F. A. Kruse, K. S. Kierein-Young, and J. W. Boardman, Mineral mapping at Cuprite, Nevada with a 63-channel imaging spectrometer, *Photogrammetric Eng. Remote Sens.,* **56**: 83–92, 1990.
38. M. Rast et al., An evaluation of techniques for the extraction of mineral absorption features from high spectral resolution remote sensing data, *Photogrammetric Eng. Remote Sens.,* **57**: 1303–1309, 1991.
39. F. A. Kruse, Use of Airborne Imaging Spectrometer data to map minerals associated with hydrothermally altered rocks in the northern Grapevine Mountains, Nevada and California, *Remote Sens. Environment,* **24**: 31–51, 1988.
40. M. A. Piech and K. R. Piech, Symbolic representation of hyperspectral data, *Appl. Opt.,* **26**: 4018–4026, 1987.
41. L. Jimenez and D. A. Landgrebe, Supervised classification in high dimensional space: geometrical, statistical, and asymptotical properties of multivariate data, *IEEE Trans. Syst. Man. Cybern.,* **286**: 39–54, 1998.
42. C. Lee and D. A. Landgrebe, Feature extraction based on decision boundaries, *IEEE Trans. Pattern Anal. Mach. Intell.,* **15**: 338–500, 1993.

JAMES C. TILTON
NASA's Goddard Space Flight Center

DAVID LANDGREBE
Purdue University

ROBERT A. SCHOWENGERDT
University of Arizona

INFORMATION PROCESSING, OPTICAL.

See OPTICAL NEURAL NETS.