Hyperspectral Image Data Analysis

During the 1950s, the digital computer began to emerge as an indispensable tool for dealing with data. It was not long in this period before pattern recognition technology, the ability to discriminate between different patterns of numbers, began significant development. Then in 1957, Sputnik, the world's first artificial satellite, was launched, thus beginning the space age. It was the concurrence of these three developments, the possibility of spacecraft, pattern recognition technology, and the digital computer, that stimulated thought into how one might make observations from space to obtain information to better manage the Earth's renewable and nonrenewable resources.

This question began to be seriously addressed in the early 1960s. (Details of the early history of space-based land remote sensing are given in [1]. This issue of this journal was written in commemoration of the 25th anniversary of the launch of Landsat 1 in July, 1972.) Early work focused on what kind of measurements to make and how to process these measurements. The first thoughts quite naturally turned to imagery and the emerging image processing technology; however, it was not long before this approach was recognized as having substantial limitations. To be viable, the technology had to be economical. The desired information had to become available to the user at minimal cost. The great advantage of space-based technology was the economy of scale. Large areas could be covered very quickly and at low per unit cost. But, for example, to identify corn and its condition by direct image means would require spatial resolution of the order of centimeters so that the shape of a corn leaf could be discerned. Sensors with such resolution would be very expensive to build and operate, but the big problem would be the unreasonable volume of data that would be generated to cover even a county-sized area. Spatial resolution is one of the most expensive parameters to achieve in a space system. A more economical approach that did not require such high spatial resolution was needed, aside from the limitation that data processing technology of that day and the foreseeable future would impose.

The Multispectral Concept

What was hit upon to solve this problem came to be known as the multispectral approach. The fundamental basis for space-based remote sensing is that information is potentially available from the electromagnetic energy field

1053-5888/02/$17.00©2002IEEE

The basis for space-based remote sensing is that information is available from the electromagnetic energy field arising from the Earth's surface and from the spatial, spectral, and temporal variations in that field.

arising from the Earth's surface and, in particular, from the spatial, spectral, and temporal variations in that field. Rather than focusing on the spatial variations, which imagery perhaps best conveys, why not move on to look at how the spectral variations might be used. The idea was to enlarge the size of a pixel until it includes an area that is characteristic from a spectral response standpoint for the surface cover to be discriminated. For example, for corn this should be several meters so as to include an area involving several rows of corn. This composite of several rows of corn is assumed to have a response as a function of wavelength that is relatively unique to corn. For an urban area where the classes to be discriminated might be low density housing, high density housing, commercial, industrial, etc, the pixels should be perhaps several tens of meters so as to pick up a composite of the responses that go to make up those classes. The idea was not to "see" a house, but to sense the mixture that a collection of closely spaced houses and the intervening materials characteristic of high-density housing emits compared to the other classes. Then the discrimination between classes would be based upon the difference in distribution of the energy from a pixel in terms of the wavelength distribution. The fundamental assumption is that different classes of surface cover have families of spectral responses that are unique to them within a data set.

Thus, the focus of data collection moved from imagery per se, i.e., collecting measurements from every high resolution pixel location on the ground where pixels were to be immediately adjacent to one another with a proper geometric relationship between measurements, to one of making measurements of the power level emanating from each more moderate resolution pixel in each of several bandwidths. In this case, pixels did not need to be immediately adjacent to each other to facilitate identification of the pixel contents, since the identification of a pixel's contents could be based on the spectral response of that pixel only. This greatly reduces the number of pixels that must be measured to survey a given area, and since data volume increases as the square of the spatial resolution, but only linearly as the number of spectral bands, upon reducing the spatial resolution while increasing the number of spectral bands, the data volume is greatly reduced.

The early research on this approach in the 1960s was done with aircraft-mounted sensors that were optical-mechanical line scanning devices capable of making pixel measurements in less than 20 spectral bands over the visible, reflective infrared (IR) and thermal IR regions of the spectrum. However, when the time arrived to design and build a space sensor of this type, the space-based sensor technology would only permit a four-band system with 80 m pixels and a S/N justifying a 6-bit data system. This system, called MSS (multispectral scanner), was first launched in July 1972 aboard the Landsat 1 satellite (originally called Earth Resource Technology Satellite or ERTS 1). This sensor system proved to be very successful, but its rather crude spectral detail did limit the number and detail of ground cover classes that could be mapped in this way.

The success of MSS resulted in consideration for a second-generation system to begin in 1975. The resulting system, called Thematic Mapper, has seven spectral bands, 30 m pixels, and a SNR justifying an 8-bit data system. This system first flew in 1982 and, with relatively minor augmentations, is the current Landsat instrument, having been most recently launched onboard Landsat 7 on 1 April 1999. Several other land-oriented sensor systems are now in orbit operated by commercial organizations and other countries. A number of aircraft-based systems are also in routine use.

In the mean time, sensor technology has advanced substantially, thus allowing multispectral sensors with several hundred spectral bands and S/N requiring 10+ bit data systems. The launch of the experimental NASA EO-1 spacecraft in November 2000 carrying a sensor system called Hyperion, with 220 bands, 30 m pixels and a 10-bit data system is a demonstration of what sensor technology is now capable of producing. Sensors with this many spectral bands are referred to as hyperspectral.

Signal Representation

The data that is supplied by such systems is best represented in the form of an N-dimensional vector for each pixel where N is the number of spectral bands. This viewpoint of the data is referred to as a feature space representation, as compared to the image space and spectral space presentation in Fig. 1. Typically there are several hundred thousand pixels per data set. The spectral space graph of Fig. 1(b) might lead one to believe that each ground cover material is appropriately represented by a single spectral curve; some use the term "spectral signature." To proceed from this assumption gives up a considerable amount of potential. The angle of the sun, and thus the time of day, season and latitude, the direction of view, the atmospheric condition, and a number of other such uncontrollable variables substantially affects the spectral response of any given material. From a scientific point of view, it has been of interest to try to make adjustments for these variables. However, this proves to be quite a daunting problem as it is difficult to accumulate the needed data

for each pixel and each column of atmosphere to enough precision to do more than have a cosmetic effect on the data in image space. Sound application of appropriate analysis algorithms are not usually much improved by such adjustments.

More significantly, beyond these observational variables, the Earth's surface itself is a highly variable and dynamic place from a spectral point of view. Consider the grassy areas of Fig. 1(a). Even in terms of the three bands used to generate this image, it is apparent to the unaided eye that the spectral response of the class "grass" varies significantly over that scene. From a data analysis point of view, it is important to recognize that this variation in the ground scene response is not all "noise." Some (most) of this variation is information bearing. Thus, from a data analysis standpoint, a more effective and complete representation of diagnostic spectral responses is in terms of class-conditional probability density functions in the N-dimensional vector space.

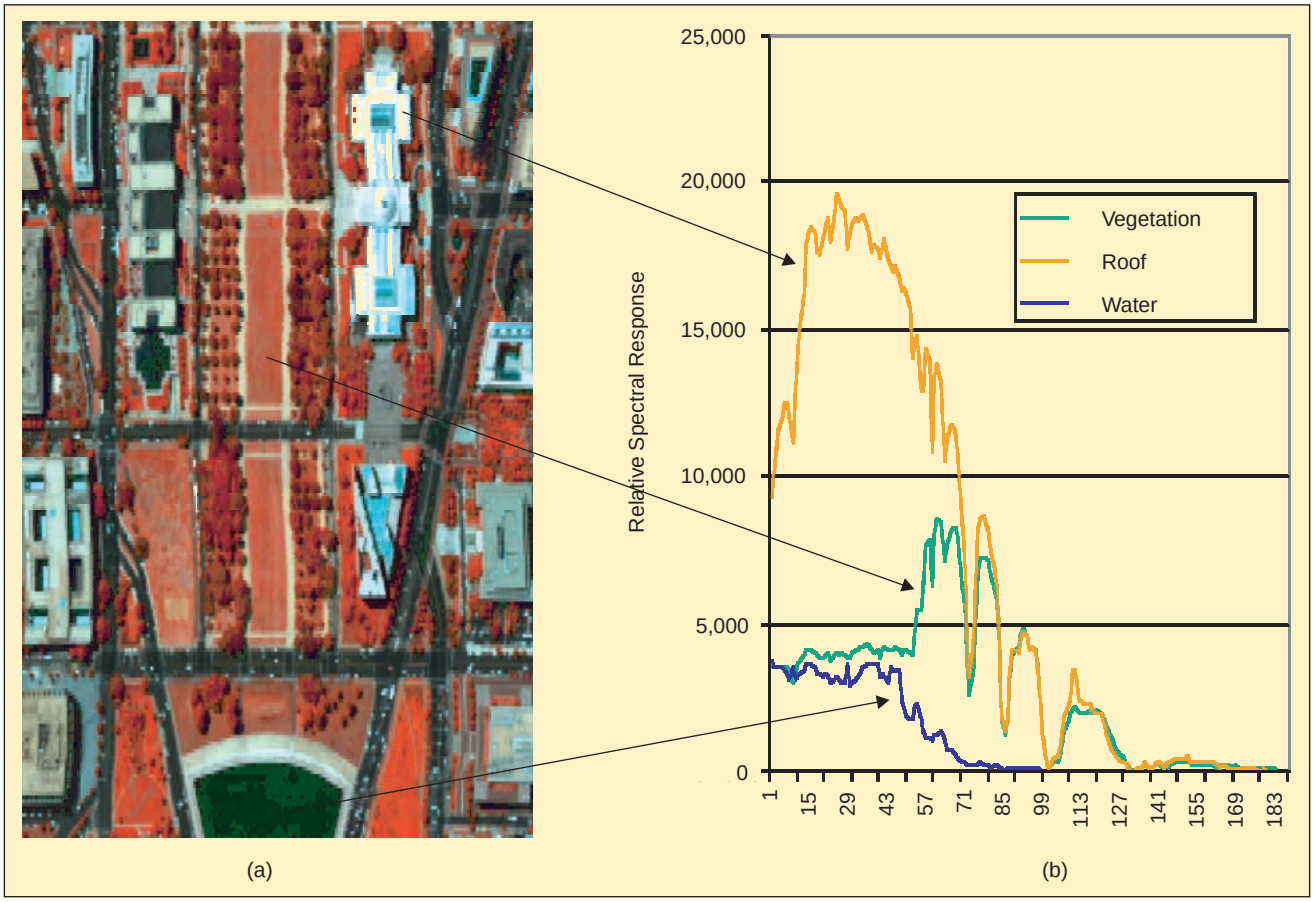
Class Discrimination

It is in such a representation, where not only the average spectral response but also the manner of variation of a material's response about its average exhibits, that is the most information bearing. To make clearer the value of this model in discriminating between two classes, one of the most common ways to predetermine the separability of two classes of materials is by the use of a statistical distance measure [2]. As an example, a commonly used one for this purpose is the Bhattacharyya distance, defined as

$$B = Ln \int_{-\infty}^{\infty} \sqrt{f(\mathbf{x}) g(\mathbf{x})}\, d\mathbf{x}$$

where $\mathbf{x}$ is the measured (vector) value of a pixel and $f(\mathbf{x})$ and $g(\mathbf{x})$ are the class-conditional density functions between which one wishes to discriminate. The form of this distance measure in terms of only the first two moments of these two density functions is

$$B = \frac{1}{8}[\mu_f - \mu_g]^T \left[\frac{\Sigma_f + \Sigma_g}{2}\right]^{-1} [\mu_f - \mu_g] + \frac{1}{2} Ln \frac{\left|\frac{1}{2}[\Sigma_f + \Sigma_g]\right|}{\sqrt{|\Sigma_f||\Sigma_g|}}$$



▲ *1. (a) A simulated color IR image of an urban area, the Washington, D.C., mall. This image is made using three bands of the 210 bands collected by the sensor system, one band from the visible green, one from the visible red, and one from the near infrared. Such displays are referred to as displays in image space. (b) A display of the data of pixels of three materials as a function of wavelength by spectral band number. The bands in this case are approximately 10 nm wide over the range of 0.4-2.4 μm. This type of data display is referred to as a display in spectral space.*

An advantage of the feature space representation is that its dimensionality is easily expanded, while that of the image space is not.

where μ_f and μ_g are the class mean values and Σ_f and Σ_g are the covariance matrices of the two classes. Note that the first term on the right measures the portion of the class separation due to the difference in means, while the second term measures the separation of the classes due to the covariances. Thus, to use only a single spectral curve to model a class (a "spectral signature?"), even if it is the average of a number of actual spectral responses makes use of only the separability measured by the first term on the right of the above equation. Further, even from this partial modeling of the class densities, it is clear that, though two classes might have the same mean values, making that first term on the right zero, they may still be quite separable.

Modeling each class in terms of a probability density function allows one to capture the information about a class also by the "shape" of the class in feature space, as quantified by all higher order statistics. Then classification can conveniently be implemented via the discriminant function concept. That is, for m classes, determine a set of m functions $\{g_1(\mathbf{x}), g_2(\mathbf{x}),\ldots, g_m(\mathbf{x})\}$ such that $g_i(\mathbf{x})$ is larger than all others whenever $\mathbf{x}$ is from class i. The class density functions can conveniently serve as the discriminant functions, and the appropriate classification rule then is

Let ω_i denote the ith class.

Decide $\mathbf{x} \in \omega_i$ iff (i.e., $\mathbf{x}$ is in class ω_i if and only if) $g_i(\mathbf{x}) \geq g_j(\mathbf{x})$ for all $j = 1,2,\ldots m$.

In this way, the process of designing the specific classifier is reduced to quantifying the class conditional density function that applies for each class. All such classifiers are thus maximum likelihood classifiers, and one may define a hierarchy of classifiers by making different assumptions about the relationship among the classes. For example,

Minimum distance to means classifier:

$$g_i(\mathbf{x}) = -(\mathbf{x} - \mu_i)^T (\mathbf{x} - \mu_i)$$

The classes all have unit variance in all features and the features are all uncorrelated to one another.

Fisher's Linear Discriminant classifier:

$$g_i(\mathbf{x}) = -(\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i)$$

The classes do not have the same variance in all features, the features are not necessarily uncorrelated, but all classes have the same variance and correlation structure.

Quadratic (Gaussian) classifier:

$$g_i(\mathbf{x}) = -(1/2)\ln|\Sigma_i| - (1/2)(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)$$

The classes are not assumed to have the same covariance, each being specified by Σ_i.

If the prior probabilities of each class are known, using them to multiply the class density functions as a part of the discriminate functions results in algorithms which derive directly from Bayes Rule and are sometimes called minimum error classifiers, because they result in the theoretically minimum overall error. In practice, these are not often used, because the prior class probabilities are often not known and/or though they minimize the overall error, they may do so by reducing the error of large classes at the expense of less frequently occurring classes.

Perhaps the next step up in classifier complexity would be to use a classifier utilizing third or even higher order statistics, up to a so-called nonparametric scheme. However, this quickly reaches a point of diminishing returns, since, as will be more apparent shortly, a key variable controlling the selection of algorithm complexity is the number of training samples available by which to define each of the classes quantitatively in feature space. It is characteristic of the remote sensing situation that the number of training samples available is always less than might be desirable. Thus, parameter estimation error quickly becomes dominant in limiting performance. Other types of algorithms are possible and in use, including ones which incorporate both spatial and spectral variations.

The question of what algorithm to use for classification is a well-studied matter. The quadratic pixel classifier is perhaps the most common algorithm. For classes with more complex distributions, something common in multispectral data, a typical approach is to define several subclasses, each with a quadratic distribution. So-called nonparametric schemes and iteratively trained algorithms such as neural network approaches have been highly studied. They can usually be made to perform well on individual data sets, but have the disadvantage of the need for extensive amounts of computation and larger training sets, both of which do not fit the practical remote sensing situation as well.

Another factor in deciding what algorithm to use for classification is the nature of the information being sought. The algorithms above are most commonly used where the intention is to make a thematic map of the entire scene, such that each pixel is labeled as to the class of its contents. In this case the classifier makes a decision as to which class a pixel should be assigned to after considering the exhaustive list of possibilities. Another type of situation might be to search a data set for a specific class of material. An application of this type might be military target identification. There are other types of algorithms that are sometimes used in such cases. Algorithms such as those based on matched filtering are examples.

The Potential of High-Dimensional Data

Consider the following. A two-channel feature space plot for the area marked by the dashed rectangle in the three-channel image space figure is shown in Fig. 2. From

the image space presentation, which utilizes three of the 12 bands available in this data set, it appears that there are two fairly distinct classes of ground cover in the rectangular area, but this is not so apparent from a visual observation of the two-dimensional feature space presentation. For these two classes in these two bands, the data appears to be heavily overlapped, and the two classes do not appear to be spectrally distinct.

However, an advantage of the feature space representation is that its dimensionality is easily expanded, while that of the image space is not. If one adds a third dimension to this feature space or a fourth, one might well be able to visualize that spreading these same data points over the larger volume of the higher dimensional space would allow for greater potential separability. Increasing the dimensionality further would spread the data over an even greater volume, thus reducing overlap and enhancing the potential for discrimination, so long as the fundamental assumption that different materials do have diagnostically different characteristics remains valid. (We note in passing that for multispectral data of Earth observational scenes, like the case illustrated above, classes of data in N-dimensional feature space usually do not occur in distinct clusters. Rather they occur in a sparse continuum, making the process of quantitatively specifying to considerable precision the classes to be discriminated a key to successful data analysis.)

As an extreme illustration of this, consider that one has 10-bit data in 100-dimensional space, a very feasible circumstance today. The 10-bit data implies 1024 possible discrete values in each of the 100 dimensions, or that there are approximately $(10^3)^{100} = 10^{300}$ discrete locations in this feature space. The volume of this space is so great that even for a data set of 10^6 pixels, the probability of any two pixels landing in the same digital cell or even fairly adjacent cells is vanishingly small. Thus there is no overlap, and in theory, anything is separable from anything. However, there are complexities that must be dealt with effectively in such a space in order to approach this potential.

High-dimensional vector spaces have been found by mathematicians to have some rather unusual and unintuitive characteristics [3]. Consider the following. It has been shown [4] that the volume of a hypersphere of radius r in d dimensions is given by

$$V_s(r) = \frac{2r^d}{d}\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$$

and that the volume of a hypercube in $[-r, r]^d$ is given by

$$V_c(r) = \text{volume of a hypercube} = (2r)^d.$$

The fraction of the volume of a hypersphere inscribed in a hypercube of the same dimension then is

$$f_d = \frac{V_s(r)}{V_C(r)} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)}$$

where d is the number of dimensions. Fig. 3 shows how f_d decreases as the dimensionality increases.

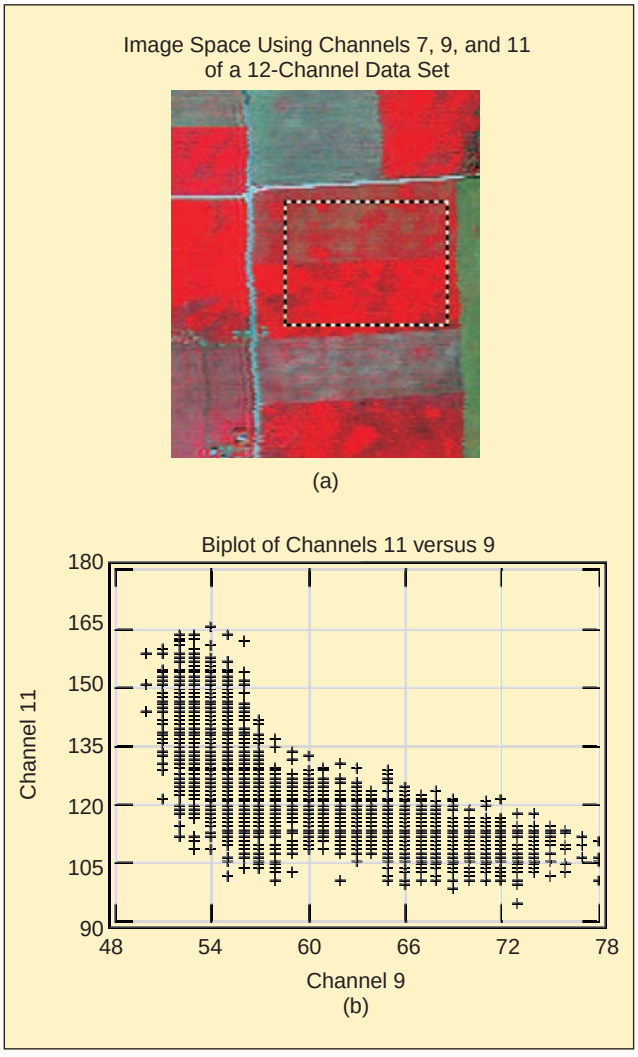
Note that $\lim_{d\to\infty} f_d = 0$, which implies that the volume of the hypercube is increasingly concentrated in the corners as d increases. Notice also that the dimensionality does not need to be very high, less than ten and certainly less than 100, for this effect to be significant.

These and other such characteristics have two important consequences for high dimensional data. The first one is that

▲ High-dimensional space is mostly empty, which implies that multivariate data in R^d is usually in a lower dimensional structure. As a consequence, for any given analysis task, high-dimensional data can be projected to a lower dimensional subspace without losing significant information in terms of separability among the different statistical classes. However, the specific subspace will surely be different for each different data set and analysis task.

A second consequence of the foregoing, is that

▲ Normally distributed data will have a tendency to concentrate in the tails; similarly, uniformly distributed data will be more likely to be collected in the corners, making



▲ *2. Two agricultural species in (a) three-dimensional image space and (b) two-dimensional feature space.*

Sensor systems must be built with a large number of spectral bands, so that they will provide suitable data for a broad spectrum of tasks and circumstances.

density estimation more difficult. Local neighborhoods are almost surely empty, producing the effect of losing detailed density estimation.

It turns out that this difficulty in density estimation is one of the chief challenges facing the data analyst. Due to the large number of parameters of the scene and its observation, one must expect to have to train a classifier for each new data set that is to be analyzed. The labeling of training samples and accumulation of the information by which to do so nearly always means that there will be a paucity of training samples with which to model each of the class density functions. Thus, one must determine the parameters of a high dimensional density function with a relatively small number of samples.

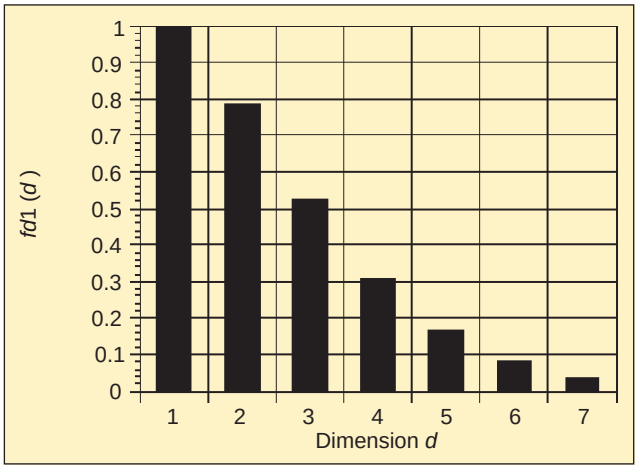
In a very general context, Hughes was able to demonstrate the impact of this problem on a theoretical basis some years ago [5]. One of his results is displayed in Fig. 4, which shows the mean recognition accuracy averaged over the ensemble of possible classifiers, versus the measurement complexity. Here, measurement complexity is related to the number of discrete cells in the feature space, and therefore the number of spectral bands and the bit precision in each. The parameter, m, of the individual graphs of the figure is the number of training samples available to define the classes.

It is seen that the expected accuracy starts at 50% for this two-class case, i.e., chance performance. For the case of an infinite number of training samples, the curve proceeds upward to the right as measurement complexity increases, rapidly at first but then more slowly, becoming asymptotic to its final value. However, for any finite number of training samples, the result has a maximum value. This is because there will then be estimation error in determining the values of the parameters of the classifier, and for a given number of training samples, the greater the measurement complexity the greater the estimation error and the poorer the performance. This may be the explanation for less complex classifiers sometimes outperforming more complex ones. The maximum value of each curve does increase with increasing numbers of training samples, and in this case, occurs at a higher measurement complexity. Thus on average, to achieve higher accuracy will require increased numbers of features and/or an increase in SNR reflected in the number of bits or discrete values per feature. Thus the number of spectral features and the SNR are interrelated with the number of training samples available per class.

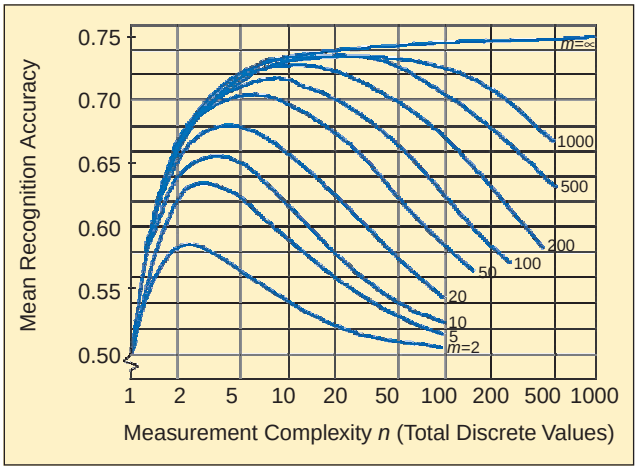
Feature Extraction

The combined implication of this is that a larger number of spectral bands may potentially make the discrimination between more detailed classes possible, but to do so will require an increasingly precise specification of the classes desired, sort of the inverse of the computer user's mantra, "garbage in, garbage out." Sensor systems must be built with a large number of spectral bands, so that they will provide suitable data for a broad spectrum of tasks and circumstances. The realization pointed out above, that high-dimensional spaces are mostly empty and a subspace will contain the significant structure for a given classification problem, points to the value of having a means for finding the most appropriate subspace as soon as the specific classes have been quantified. Algorithms for accomplishing this are referred to as feature extraction algorithms. Two examples are discriminate analysis [6] and decision boundary [7] feature extraction.

Discriminate analysis feature extraction (DAFE) is based on the following concept. In seeking the optimal subspace, the primary axis of this transformation should be oriented such that the classes have the maximum sepa-



▲ *3. Fractional volume of a hypersphere inscribed in a hypercube as a function of dimensionality.*



▲ *4. Mean recognition accuracy versus measurement complexity for the finite training case.*

ration between their means on this new axis, while at the same time they should appear as small as possible in their individual spreads. If the former is characterized by σ_B, the distance between the means, and the latter in terms of σ_{W1} and σ_{W2}, the spread of the classes about their means, then it is desired to find new axes such that

$$\frac{\sigma_B^2}{\sigma_W^2} = \frac{\text{between} - \text{class variance}}{\text{average within} - \text{class variance}}$$

is maximized, where σ_W^2 is the average of σ_{W1}^2 and σ_{W2}^2. In matrix form the within-class scatter matrix Σ_W and the between-class scatter matrix Σ_B may be defined as

$$\Sigma_W = \sum_i P(\omega_i)\Sigma_i \quad \text{(within class scatter matrix)}$$

$$\Sigma_B = \sum_i P(\omega_i)(\mu_i - M_o)(\mu_i - M_o)^T \quad \text{(between class scatter matrix)}$$

$$M_o = \sum_i P(\omega_i)\mu_i.$$

Here μ_i, Σ_i, and $P(\omega_i)$ are the mean vector, the covariance matrix, and the prior probability of class ω_i, respectively. The criterion for optimization may be defined as

$$J_1 = tr(\Sigma_W^1 \Sigma_B).$$

New feature vectors are selected to maximize the criterion.

Decision boundary feature extraction (DBFE) is based directly upon the decision boundary in feature space and the training samples that define it. Discriminately informative features have a component that is normal to the decision boundary at least at one point, while discriminately redundant features are orthogonal to a vector normal to the decision boundary at every point on the boundary. Based upon this, a decision boundary feature matrix (DBFM) may be defined to extract discriminately informative and discriminately redundant features from the decision boundary. The rank of the DBFM is the smallest dimension where the same classification accuracy can be obtained as from the original feature space, and the eigenfunctions of the DBFM corresponding to nonzero eigenvalues are the necessary features to achieve the same accuracy as in the original feature space. The calculation process uses the training samples themselves, rather than statistics from them, to determine the location of the decision boundary, and then from that, the DBFM. The details are contained in the referenced work.

A Data Analysis Paradigm

The major question that the analyst must deal with is how to choose and train a suitable sequence of algorithms by which to accomplish the desired analysis, given the circumstances found in the remote sensing situation. The problem of optimally training a classifier comes down to how completely and precisely one models the data set and the specific classes one wishes to discriminate between. The classification process ordinarily involves assigning each pixel to one of a list of classes. Thus one must set up an exhaustive list of classes, so that there is a logical class to which to assign each pixel of the data set, even though one may be interested in only one or a small number of classes in the scene. The rule for establishing the list of classes then is that the classes must be:

- Of informational value. The list must contain all of the classes of interest to the information consumer.
- Exhaustive. In addition to those desired by the user, it must contain enough additional classes so that there is a logical class to which to assign each pixel in the data set.
- Separable. The classes must be separable in terms of available spectral features.

Further, each class must be modeled to adequate completeness and precision. As pointed out earlier, one must specify not only the mean response of a given class, but also how the response for that class varies about its mean, since this variation is often quite diagnostic of the class. Modeling the class response in terms of a multidimensional probability density function is perhaps the most effective way of doing this. However, as the measurement complexity, defined by the number of features and the bit precision of the data reflecting S/N, increases, this becomes more daunting. There will usually only be a limited number of samples that can be made available for defining a class density function model. The number of training samples needed varies greatly with the specific situation. A number many times as large as the number of features is highly desirable, although there are algorithms becoming available to mitigate this condition to some extent.

In addition, the analyst has the further challenge that the samples used for training the classifier must be truly representative of the class intended. If one wishes to define a class to be called "corn" in an agricultural problem, how thick must be the stand of corn in a pixel for it to be desired to call the pixel corn, how much weeds should be allowed, what range of varieties (include popcorn?) what range of planting dates and thus maturity level, and many other variables must be considered. Clearly the process is not in any sense "automatic" as it must reflect the specific requirements of the user.

Data flow through a system to a final analysis generally requires the application of a sequence of algorithms. Fig. 5

High-dimensional vector spaces have been found by mathematicians to have some rather unusual and unintuitive characteristics.

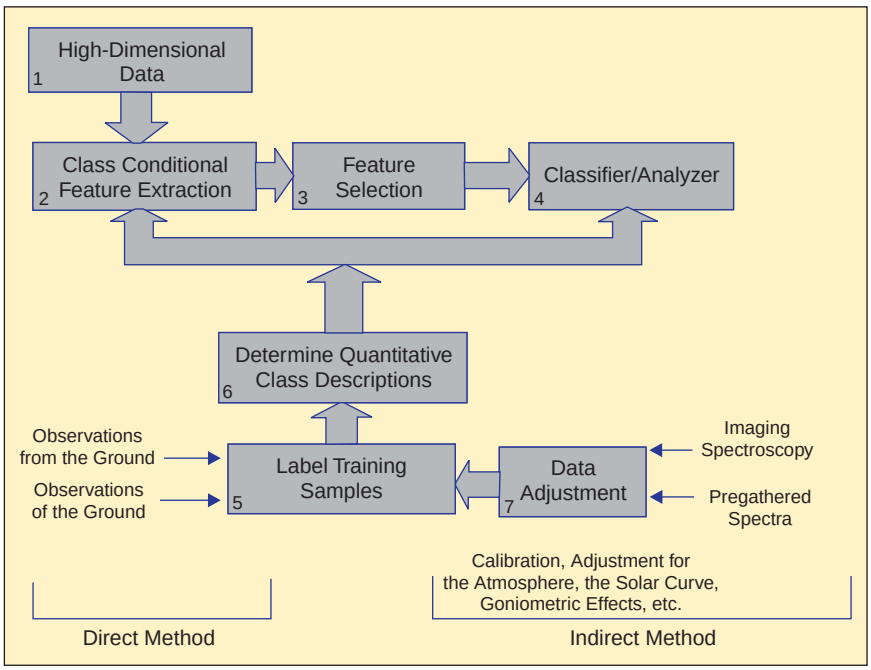
Data flow through a system to a final analysis generally requires the application of a sequence of algorithms.

outlines such a sequence. The numbered paragraphs refer to the numbered boxes in the diagram.

1) Multispectral data consists of data gathered in more than one spectral band. There is no accepted definition for where the boundary is between data termed multispectral and hyperspectral. It is well established that the geometry of vector spaces changes continually as the dimensionality of the space increases, and indeed that it is materially different from the familiar three-dimensional geometry by the time dimensionality reaches seven to ten. Further, it usually requires a dimensionality of the order of ten or more to satisfactorily accomplish many practical analysis tasks. Thus it will be assumed that the data to be analyzed contains at least ten and perhaps as many as several hundred spectral bands.

2) Again assuming that the data were gathered in a larger number of bands than is necessary or desirable for the particular analysis at hand, an important early step is to form the feature subset that is to be used in the analysis. This should be done in a situation-specific way, that is, using the description of the specific classes desired. Thus, a feature extraction algorithm such as those described above is applied at this point.

3) Given box 2, there may still remain the decision as to how many of the generated features to utilize. The choice here and that in box 4 will depend to some extent upon the individual classes and the precision with which they have been modeled.



▲ *5. A schematic diagram of the hyperspectral data analysis process.*

4) There remains, then, the application of the specific classification algorithm to be used. Again, the choice of algorithm depends upon the class model precision and the level of detail of the classes.

5) As has been detailed above, the labeling of adequate sets of training samples is a key step, perhaps the most important step of the entire process.

6) Having labeled a set of samples for each class that are assumed to be truly representative of an exhaustive list of classes that includes the desired classes, the task here is to use those samples to define as precise an N-dimensional model of the classes in the feature space as possible. Except in very simple cases where a single point in feature space is adequate, this will nearly always consist of modeling the entire distribution of each class. This may involve use of an iterative scheme, or it may simply consist of computing first- and second-order statistics. However classes may require modeling in terms of more than one mode, with the training samples divided between the various modes. There are also additional algorithms that can further assist in mitigating the small training sample problem [8].

7) Box 7 suggests one option for labeling training pixels being an attempt to adjust all or a part of the data for the various observational variables that were present, depending on the precise conditions of the scene and the sensor system at the time each pixel measurement was made. If one could do this adequately, this would make possible the use of some additional sources of reference data on which to base the labeling, as indicated on the diagram. The adjustment of the data for all of these variables is a very complex task and is problematic. It often cannot be done with as much precision as needed. Because of this, the overall scheme above is designed to not necessarily require calibrated data that has been so adjusted. Rather, one would only need to do so to the extent necessary to label an adequate set of training samples. Of course, if one has available information such as that indicated by one of the direct methods, the need for this added complexity can be avoided.

There are a number of additional kinds of processing that can be done to treat various circumstances that arise in practical problems. For example, having labeled a small set of samples of a presumably exhaustive list of classes, one cannot be sure that the labeled samples are indeed typical of all of the occurrences of classes in the entire scene, since no information about the quantitative nature of those other occurrences has been available. In short, the classifier may not be able to generalize well from the training samples to the nontraining occurrences of pixels.

An example scheme to mitigate this problem, which might be applied in box 6, is to use a systematic sampling of nontraining samples, in conjunction with the training samples, to modify or "enhance" the class statistics [9]. Since this process in effect increases the number of training samples in the process, it can also mitigate the Hughes phenomenon, i.e., the estimation error problem due to having too few training samples. Assume there are J classes in the feature space denoted by $S_1 ...,S_J$. Each class can have several Gaussian components. Let m denote the total number of the Gaussian components. Write $i \in S_j$ to indicate that component i belongs to class S_j. The probability density function of the feature can then be written as a mixture of m Gaussian components where the set of components can be partitioned into m classes:

$$f(x|\theta) = \sum_{i=1}^{m} \alpha_i f_i(x|\phi_i)$$

where

$$\phi_i = (\mu_i, \Sigma_i), \quad \theta = (\alpha_1,...,\alpha_m, \mu_1,...,\mu_m, \Sigma_1,...,\Sigma_m).$$

From each class S_j, N_j training samples are assumed to be available. Denote these samples by z_{jk} where $j = 1,..., J$ indicates the class of origin and $k = 1,..., N_j$ is the index of each particular sample. The training samples here are known to come from a particular class without any reference to the exact component within that class. In addition to the training samples, N unlabeled samples denoted by x_k, $k = 1,..., N$, are also assumed to be available from the mixture. The log likelihood to be maximized for obtaining the ML estimates can be written in the following form:

$$L(\theta) = \sum_{k=1}^{N} \log f(x_k|\theta) + \sum_{j=1}^{J} \sum_{k=1}^{N_j} \log\left(\frac{1}{\sum_{t \in S_j} \alpha_t} \sum_{l \in S_j} \alpha_l f_l(z_{jk}|\phi_l) \right).$$

The first term in the above log likelihood function is the likelihood of the unlabeled samples with respect to the mixture density, and the second term indicates the likelihood of the training samples with respect to their corresponding classes of origin. The EM equations for obtaining the ML estimates are the following:

$$\alpha_i^+ = \frac{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})}{N\left(1 + \frac{N_j}{\sum_{r \in S_j} \sum_{k=1}^{N} P^c(r|x_k)}\right)}$$

$$\mu_i^+ = \frac{\sum_{k=1}^{N} P^c(i|x_k) x_k + \sum_{k=1}^{N_j} P_j^c(i|z_{jk}) z_{jk}}{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})}$$

$$\Sigma_i^+ = \frac{\sum_{k=1}^{N} P^c(i|x_k)(x_k - \mu_i^+)(x_k - \mu_i^+)^T}{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})} + \frac{\sum_{k=1}^{N_j} P_j^c(i|z_{jk})(z_{jk} - \mu_i^+)(z_{jk} - \mu_i^+)^T}{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})}.$$

The equations are applied iteratively with respect to the training and unlabeled samples where "c" and "+" refer to the current and next values of the respective parameters, $i \in S_j$, and $P^c(.|.)$ and $P_j^c(.|.)$ are the current values of the posterior probabilities:

$$P^c(i|x_k) = \frac{\alpha_i^c f_i(x_k|\mu_i^c, \Sigma_i^c)}{f(x_k|\theta^c)}$$

$$P_j^c(i|z_{jk}) = \frac{\alpha_i^c f_i(z_{jk}|\mu_i^c, \Sigma_i^c)}{\sum_{t \in S_j} \alpha_t^c f_t(z_{jk}|\mu_t^c, \Sigma_t^c)}.$$

Thus as the iteration proceeds, successively revised values for the mean, covariance, and weighting coefficient of each component of each class are arrived at which steadily approach the values for a maximum of the expected likelihood value for the mixture density. The result can be class statistics that better model the entire data set, better classifier generalization, and a reduction of the estimation error of class statistics in the face of high dimensionality.

An Example Analysis

We conclude with one specific example. Fig. 6 shows an image space representation, using three bands to simulate a color IR photograph of an airborne hyperspectral data set over the Washington, D.C., mall. The data set was collected with an airborne sensor system delivering approximately 3 m pixels containing 210 spectral bands from the 0.4 to 2.4 μm region of the visible and infrared spectrum. This data set contains 1208 scan lines with 307 pixels in each scan line. It totals approximately 150 Mbytes. The analysis was done on an inexpensive personal computer. Thus, though by some standards, this is "a lot of data, calling for a powerful computer," this is not the case as will be seen. The specific steps for this analysis are briefly described as follows.

▲ *1) Display Image.* The first step is to present a view of the data set in an image space so that the analyst can select

▲ *6. A simulated color infrared image of the Washington, D.C., mall.*

and mark training samples, examples of each class desired in the final thematic map. A simulated color infrared photograph form is often convenient for this purpose; to do so, bands 60, 27, and 17 are used for the red, green, and blue colors, respectively. The result is shown in Fig. 6. Note that this particular data has not yet been adjusted for geometric distortion that arose due to air turbulence affecting the aircraft stability during data collection. Since the analysis is done on a pixel-by-pixel basis, this has no effect on the analysis process at this point. Rectification or geometric adjustments may be made either before or after spectral analysis so long as such adjustments do not affect the radiometric values of the pixels.

▲ *2) Define Classes.* Use the image display of the data to mark training samples for each desired class. The classes of informational value desired in this case are "Roofs," "Road," "Grass," "Trees," "Trail," "Water," and "Shadow." The class Shadow is not necessarily desired by the user, but is an example of the need to satisfy the requirement for the class list to be exhaustive, since areas in the scene in deep shadow are spectrally substantially different from the other areas.

The significant challenge for this analysis task stems from the fact that though the user would like to discriminate between the classes "Roof" and "Road," the materials used in some roofs are very similar to that used in roads, a mixture of gravel and asphalt. Further, there are many different types of roofs. Thus, one must carefully train for a number of subclasses of Roof, so that all of the various spectral subclasses for Roof and to a lesser extent for Road are represented properly in the training data. In this case six subclasses of Roof were defined to account for the difference in spectral response of the various types of roof. After classification, the subclasses were combined for display of the results. The number of samples for each class/subclass varied in this case from a few 10s to several hundred.

▲ *3) Feature Extraction.* After designating an initial set of training areas, a feature extraction algorithm is applied to determine a feature subspace that is optimal for discriminating between the specific classes defined. The algorithm used here was discriminate analysis feature extraction (DAFE). The result is a linear combination of the original 210 bands to form 210 new features that automatically occur in descending order of their value for producing an effective discrimination. From the DAFE output, it is seen that the first nine of these features will be adequate for successfully discriminating between the classes.

▲ *4) Reformatting.* The new features defined above are used to create a nine-band data set consisting of the first nine of the new features, thus reducing the dimensionality of the data set from 210 to nine.

▲ *5) Initial Classification.* Having defined the classes and the features, next an initial classification is carried out. An algorithm called ECHO (extraction and classification of homogeneous objects [10], [11]) was used here. This algorithm is a spectral/spatial quadratic maximum likeli-

hood classifier that first segments the scene into spectrally homogeneous objects. It then classifies the objects on a quadratic maximum likelihood basis.

▲ *6) Finalize Training.* An inspection of the initial classification result indicates that some improvement in the training of the set of classes is called for. To do so, two additional training fields were selected and added to the training set.

▲ *7) Final Classification.* The data were again classified using the new training set. The result is shown in Fig. 7. Rather than a continuous tone color image, this figure is a thematic map in which each pixel is displayed in a specific color, one of the seven colors in the legend indicating the class that to which the pixel was assigned.

The entire analysis process took less than three minutes of disk read and processing time on an inexpensive PC and about half an hour of analyst time. This example and additional information can be found in [12]. The example application analysis was done on a PC-based software system called MultiSpec, which is available at http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/. This site also has pointers to substantial additional amounts of technical details, including downloadable complete copies of some of the documents referred to.

So far as Earth observational multispectral sensing is concerned, the field is primarily limited by the availability of data. However, assuming this limitation will some day be mitigated sufficiently, the next layer of limiting factors will be the satisfactory means for data analysis. The process outlined here is, at best, a first step. Clearly, the designing of such an analysis procedure is a signal processing engineering problem rather than an Earth science problem; however, the signal processing engineer must keep in mind that the analysis process will no doubt need to be carried out by an individual with a background in the technology user disciplines rather than those of the technology producer disciplines. The procedure described above works well, but it is probably much too complicated to be adopted by many users. The designing of a suitable data analysis process is still ahead and is an engineering task in the true sense of the word.

In the mean time, multispectral data processing is finding increasingly broad applications that extend will beyond aerospace remote sensing. These methods are being increasingly explored in application areas such as medical diagnostics, manufacturing, materials processing, and many others.

▲ *7. A thematic map presentation of the analysis result.*

David Landgrebe holds B.S.E.E., M.S.E.E., and Ph.D. degrees from Purdue University. He is presently Professor of Electrical and Computer Engineering at Purdue University. His interests are communication science and signal processing, especially as applied to Earth observational remote sensing. He was President of the IEEE Geoscience and Remote Sensing Society and a member of its Administrative Committee. He re-

ceived that Society's Outstanding Service Award in 1988. He is a co-author of the textbook, *Remote Sensing: The Quantitative Approach*. He is an IEEE Life Fellow, a Fellow of the American Society of Photogrammetry and Remote Sensing, a Fellow of the American Association for the Advancement of Science, a member of the Society of Photo-Optical Instrumentation Engineers and the American Society for Engineering Education, as well as Eta Kappa Nu, Tau Beta Pi, and Sigma Xi honor societies. He received the NASA Exceptional Scientific Achievement Medal in 1973. In 1976, on behalf of the Purdue's Laboratory for Applications of Remote Sensing, he accepted the William T. Pecora Award. He was the 1990 recipient of the William T. Pecora Award and the 1992 recipient of the IEEE Geoscience and Remote Sensing Society's Distinguished Achievement Award.

References

[1] D. Landgrebe, "The evolution of landsat data analysis," *Photogrammetric Engineering and Remote Sensing,* vol. LXIII, no. 7, pp. 859-867, July 1997.

[2] J.A. Richards and X. Jia, *Remote Sensing Digital Image Analysis,* 3rd ed. New York: Springer, 1999.

[3] L. Jimenez and D. Landgrebe, "Supervised classification in high dimensional space:geometrical, statistical and asymptotical properties of multivariate data," *IEEE Trans. Syst., Man, Cybernet.,C*, vol. 28, pp. 39-54, Feb. 1998.

[4] M.G. Kendall, *A Course in the Geometry of n-Dimensions*. New York: Hafner, 1961.

[5] G.F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 55-63, Jan. 1968.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1972.

[7] C. Lee and D.A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 15, pp. 388-400, Apr. 1993.

[8] J.P. Hoffbeck and D.A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 18, pp. 763-767, July 1996.

[9] BM.. Shahshahani and D.A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. Geosci. Remote Sensing*, vol. 32, pp. 1087-1095, Sept. 1994.

[10] R.L. Kettig and D.A. Landgrebe, "Computer classification of remotely sensed multispectral image data by extraction and classification of homogeneous objects," *IEEE Trans. Geoscience Electron*., vol. GE-14, pp. 19-26, Jan. 1976.

[11] D.A. Landgrebe, "The development of a spectral-spatial classifier for earth observational data," *Pattern Recognit.*, vol. 12, no. 3, pp. 165-175,1980.

[12] D.A. Landgrebe, "Information extraction principles and methods for multispectral and hyperspectral remote sensing," in *Information Processing for Remote Sensing*, C.H. Chen, Ed. River Edge, NJ: World Scientific, 1999, pp. 3-37.