Hyperspectral Image Data Analysis as a High Dimensional Signal Processing Problem

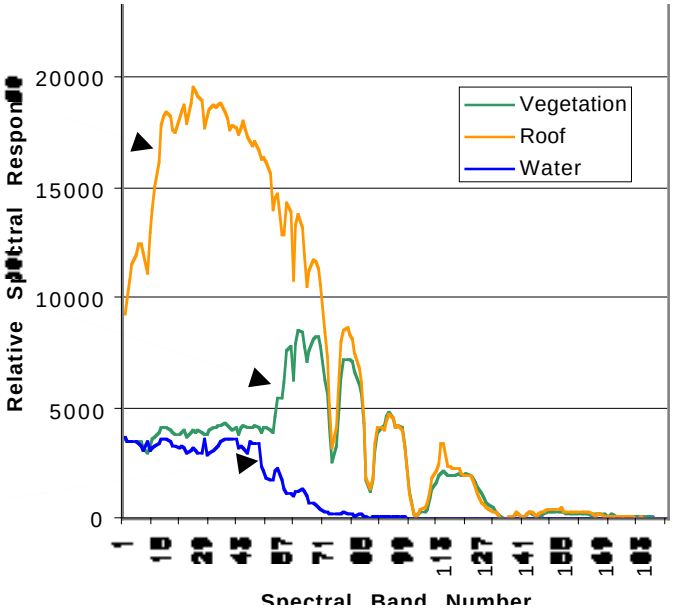
David Landgrebe
School of Electrical & Computer Engineering
Purdue University
West Lafayette IN 47907-1285
landgreb@ecn.purdue.edu

Background and Introduction. During the 1950's, the digital computer really 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 in order to better manage the Earth's renewable and nonrenewable resources.

This question began to be seriously addressed in the early 1960's[1]. 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, it was recognized that the technology had to be economical. The desired information had to become available to the user at minimal cost. The great advantage space-based technology had to offer 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,

[1] Details of the early history of space-based land remote sensing are given in Landgrebe, David, "The Evolution of Landsat Data Analysis," (Invited), Photogrammetric Engineering and Remote Sensing, Vol. LXIII, No. 7, July 1997, pp. 859-867. This issue of this journal was written in commemoration of the 25th anniversary of the launch of Landsat 1 in July, 1972.

A simulated color IR image of an urban area, the Washington DC mall. This image is made using 3 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*.



A display of the data of pixels of three materials as a function of spectral band number. This type of data display is referred to as a display in *spectral space.*

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, quite 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 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. 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, the data volume is greatly reduced.

The early research on this approach in the 1960's 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 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 4-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 7 spectral bands, 30 m pixels, and a S/N ratio justifying an 8-bit data system. This system first flew in 1982 and with relatively minor augmentations, it is the current Landsat instrument, having been launched on Landsat 7 on April 1, 1999. In the mean time, sensor technology has advanced substantially, thus allowing multispectral sensors with several hundred spectral bands and S/N requiring 10- or more 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 sensors.

The Signal Processing Problem. 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 above. Typically there are several hundred thousand pixels per data set. The spectral space graph above 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 may affect the spectral response of any given material. However, beyond that, the Earth's surface, itself, is a highly variable and dynamic place from a spectral point of view. Consider the grassy areas of the above view in image space. 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.

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 pre-determine the separability of two classes of materials is by the use of a statistical distance measure[2]. A commonly used one for this purpose is the Bhattacharyya Distance, defined as,

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

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,

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

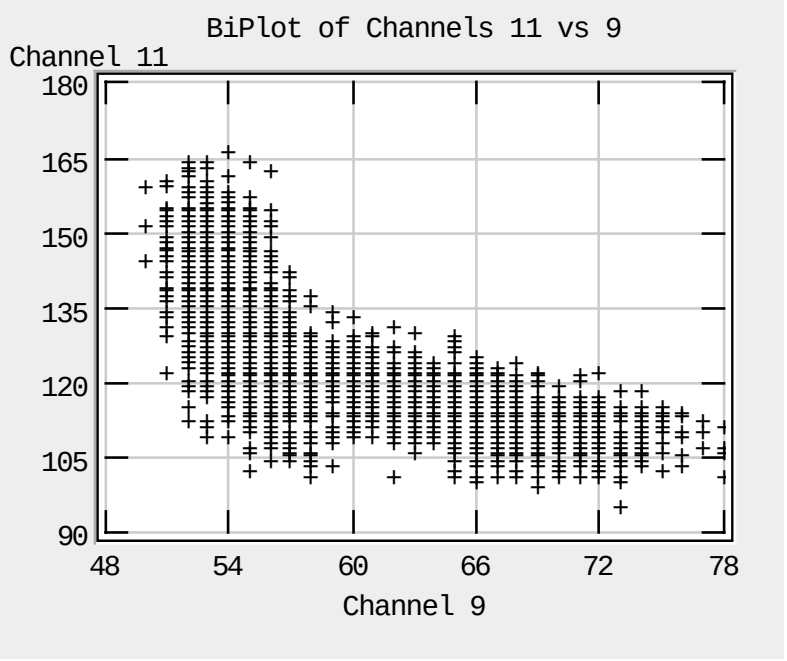
where $\boldsymbol{\mu}_f$ and $\boldsymbol{\mu}_g$ are the class mean values and $\boldsymbol{\Sigma}_f$ and $\boldsymbol{\Sigma}_g$ are the covariance matrices of the two classes. Note that the first term on the right measures the portion of the class separation due to the difference in means, while 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, 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 have the same mean values, they may still be quite separable.

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 image space figure is

[2] John A Richards and Xiuping Jia, *Remote Sensing Digital Image Analysis,* 3rd edition, Springer 1999.

shown below. From the image space presentation, which utilizes 3 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.

Image Space using channels 7, 9, and 11 of a 12 channel data set.



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 this same data 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.

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 value 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. However, there are complexities that must be dealt with effectively in such a space.

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 the equation:

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

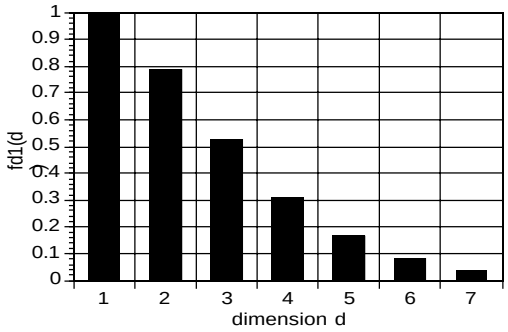
and that the volume of a hypercube in $[-r, r]^d$ is given by the equation:

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

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

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

where d is the number of dimensions. The following figure shows how f_d decreases as the dimensionality increases.



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

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.

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

[3] Luis Jimenez and David Landgrebe, "Supervised Classification in High Dimensional Space:Geometrical, Statistical and Asymptotical Properties of Multivariate Data," *IEEE Transactions on Systems, Man, and Cybernetics*, Volume 28 Part C Number 1, pp. 39-54, Feb. 1998.

[4] Kendall, M. G., A Course in the Geometry of n-dimensions, Hafner Publishing Co., 1961.

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.

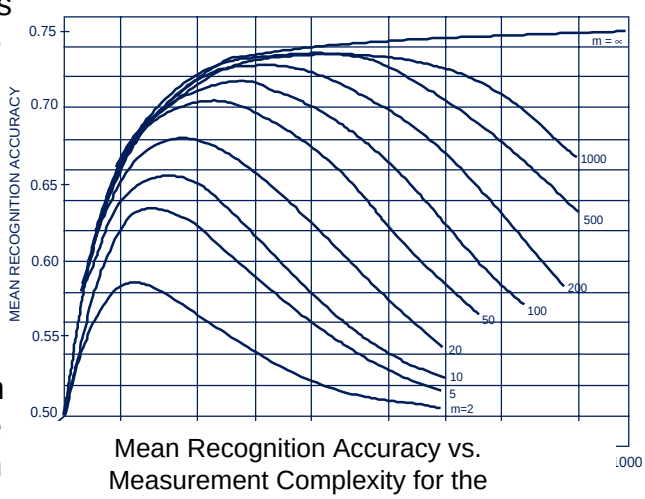
A second consequence of the foregoing, is that

- 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 density estimation more difficult. Local neighborhoods are almost surely empty, requiring the bandwidth of estimation to be large and 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 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, Gordon 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 the figure below, 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 due to the fact that there will then be estimation error in determining the parameters of the classifier and for a given



Mean Recognition Accuracy vs. Measurement Complexity for the finite training case.

[5] Hughes, G. F., "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory,* Vol. IT-14, No. 1, January 1968.

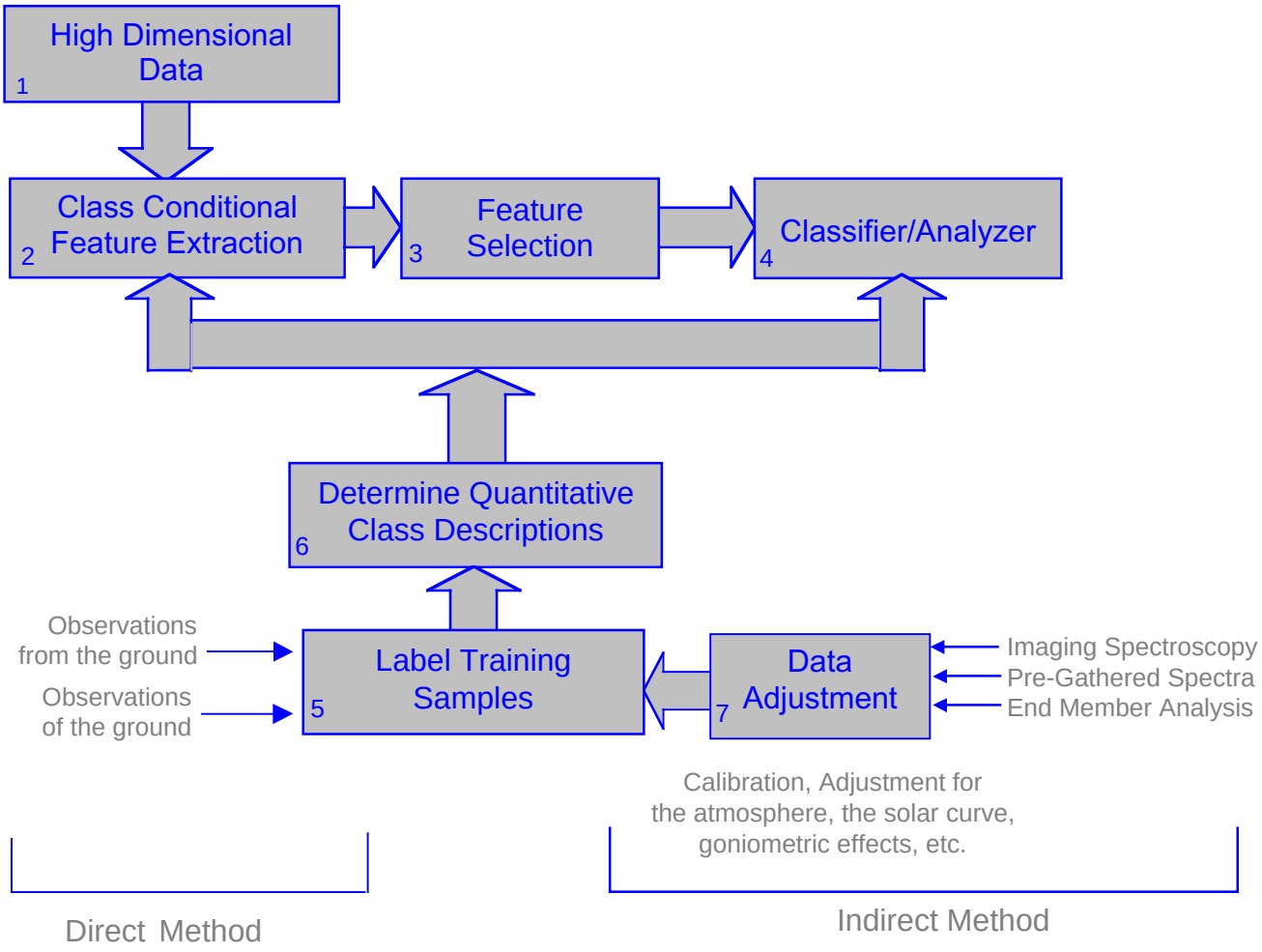
number of training samples, the greater the measurement complexity the greater the estimation error and the poorer the performance. The maximum value of each curve does increase with increasing numbers of training samples, and in this case occurs at a higher measurement complexity, implying that to achieve it will require increased number of features and/or an increase in S/N ratio reflected in the number of bits or discrete values per feature. Thus the number of spectral features and the S/N ratio are interrelated with the number of training samples available per class.

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.

A data analysis paradigm. The major question that the analyst must deal with is how to choose and implement a suitable sequence of algorithms by which to accomplish the desired analysis. The following diagram outlines such a sequence. This will be discussed in terms of the numbered boxes in the diagram below.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* Academic Press, 1972.

[7] Chulhee Lee and David A. Landgrebe, "Feature Extraction Based On Decision Boundaries," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 4, April 1993, pp 388-400.



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.
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.

A simulated color infrared image of the Washington, DC mall.

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.
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 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.
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.

An example analysis. We conclude with one specific example. The figure at left shows an image of an airborne hyperspectral data set over the Washington DC mall. The data set was collected with an airborne sensor system delivering approximately 3 m pixels containing

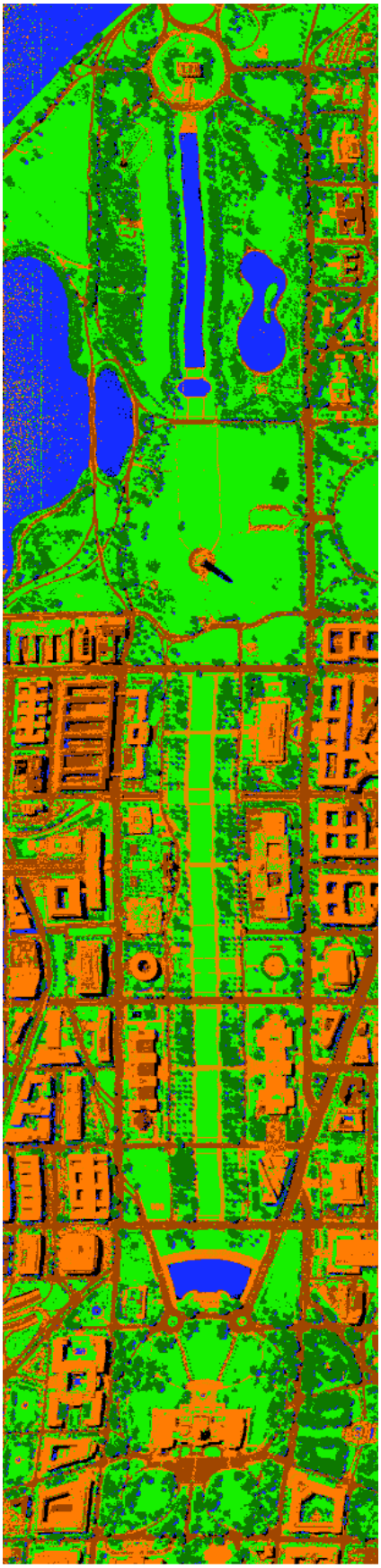
210 spectral bands from the 0.4 to 2.4 μm region of the visible and infrared spectrum. This data set contains 1208 scan lines with 307 pixels in each scan line. It totals approximately 150 Megabytes. The steps used on an inexpensive personal computer for this analysis are briefly described as follows.

1. Display Image. The first step is to present a view of the data set in image form so that the analyst can select and mark training samples, examples of each class desired in the final thematic map. A simulated color infrared photograph form is 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 the figure at left. Note that the data has not yet been adjusted for geometric distortion that arose due to turbulence 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 they 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 separate classes for "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.

3. Feature Extraction. After designating an initial set of training areas, a feature extraction algorithm is applied to



A thematic map presentation of the anlaysis result.

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 9 band data set consisting of the first nine of the new features, thus reducing the dimensionality of the data set from 210 to 9.

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[8,9]) was used here. This algorithm is a maximum likelihood classifier that first segments the scene into spectrally homogeneous objects. It then classifies the objects on a 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 the figure at right. 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.

This example and additional information about the material of this article can be found in C. H. Chen, editor, *Information Processing for Remote Sensing,* Chapter 1 by David Landgrebe, World Scientific Publishing Company, World Scientific Publishing Co., Inc., 1060 Main Street, River Edge, NJ 07661, USA, 1999. The example application analysis was done on a software system for personal computers called MultiSpec, which is available without charge at

http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/

[8] Kettig, R. L. and Landgrebe, D. A. "Computer Classification of Remotely Sensed Multispectral Image Data by Extraction and Classification of Homogeneous Objects," IEEE Transactions on Geoscience Electronics, Volume GE-14, No. 1, pp. 19-26, January 1976.

[9] Landgrebe, D.A.. "The Development of a Spectral-Spatial Classifier for Earth Observational Data," Pattern Recognition, Vol. 12, No. 3, pp. 165-175,1980.