1-1-1976

# Signature Extension through the Application of Cluster Matching Algorithms to Determine Appropriate Signature Transformations

Peter F. Lambeck

Daniel P. Rice

Reprinted from

# Symposium on Machine Processing of Remotely Sensed Data

**June 29 - July 1, 1976**

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
76CH1103-1 MPRSD

# SIGNATURE EXTENSION THROUGH THE APPLICATION OF CLUSTER MATCHING ALGORITHMS TO DETERMINE APPROPRIATE SIGNATURE TRANSFORMATIONS*

Peter F. Lambeck and Daniel P. Rice

Environmental Research Institute of Michigan
Ann Arbor, Michigan

## ABSTRACT

Signature extension is a process intended to increase the spatial-temporal range over which a set of training statistics can be used to classify data without significant loss of recognition accuracy. The goal of signature extension is to minimize the requirements for collecting ground truth and extracting training statistics, thus reducing the costs and time delays associated with those procedures. Signature extension would then help to provide timely and cost-effective classification over extensive land areas, including remote areas for which ground truth information may not be readily available.

Many current signature extension techniques are based on a transformation of training statistics to compensate for changes in sun angle, atmospheric conditions, etc., between a training area and a recognition area. Although preprocessing techniques which minimize or eliminate the need for altering training statistics are also potential solutions to the problem of signature extension, this presentation is principally concerned with those algorithms which define signature transformations based on associations between training and recognition area statistics.

ERIM has shown that since causes in nature for variations in the measured radiance from a given material are in all cases multiplicative and/or additive, an appropriate signature transformation would be both multiplicative and additive in each data channel. In principle, this signature transformation should be unique for each material since bidirectional reflectance, influenced by such factors as sun angle, wind velocity, and soil variations, is a unique attribute of each type of ground cover. However, current signature transformation algorithms concentrate, with only a few exceptions, on defining an average transformation to be applied equally to all signatures. A first cluster matching algorithm (called MASC, for Multiplicative and Additive Signature Correction) was developed at ERIM to test the concept of using associations between training and recognition area cluster statistics to define an average signature transformation.

A more recent signature extension module, CROP-A (Cluster Regression Ordered on Principal-Axis), has shown evidence of making meaningful associations between training and recognition area cluster statistics, with the clusters to be matched being selected automatically by the algorithm. These associations have led to multiplicative and additive signature corrections producing classification results over recognition areas which were significantly improved relative to what would have been achieved without the signature transformation and without local training.

The manner in which a signature extension module such as CROP-A, is embedded in an overall signature extension system has been identified as an important consideration in determining its performance and value as a signature extension tool. In this regard, research is currently underway at ERIM to define an optimum signature extension system utilizing the current state of the art. Improved signature extension modules are currently undergoing development, test, and evaluation.

Partitioning (i.e., defining the limits of regions over which a signature extension technique can reasonably be applied) has been identified as another major factor controlling signature extension utility. Hence, current research is also concerned with defining the necessary factors which limit the extent of a partition.

## INTRODUCTION

Signature extension is a process intended to increase the spatial-temporal range over which a set of training statistics can be used to classify data without significant loss of recognition accuracy. The training statistics which are required are extracted from multispectral scanner (MSS) data with the aid of

*This work is presently being performed for the Earth Observations Division of the NASA/Johnson Space Center under Contract NAS9-14123.

training information (ground truth) obtained from localized surveys on the ground or from interpretation of aerial photographs or MSS data images by trained analyst interpreters (AI's). Either of these procedures for acquiring ground truth information becomes costly and time consuming even for data processing over land areas of moderate size.

The goal of signature extension is to minimize the requirements for collecting ground truth and for extracting training statistics, thus reducing the costs and time delays associated with those procedures. Signature extension would then help to provide timely and cost-effective classification over extensive land areas, including remote areas for which ground truth information may not be readily available. This present signature extension effort has been concerned with the problem of performing large area agricultural surveys to estimate wheat production, using MSS data from the LANDSAT satellites.

Many current signature extension techniques are based on a transformation of training statistics to compensate for changes in sun angle, atmospheric conditions, etc., between a training area and a recognition area. Although preprocessing techniques which minimize or eliminate the need for altering training statistics are also potential solutions to the problem of signature extension, the following presentation is principally concerned with those algorithms which define signature transformations based on associations between training and recognition area statistics. Specific topics to be discussed below include (1) the underlying theory for the signature transformation, (2) the algorithms used to determine and to apply this transformation, and (3) improvements in signature extension which can be effected through procedures which are peripheral to the transformation itself.

## THEORY

The general form of the transfer equation representing the recorded MSS signal level within a specific spectral band for a given material $\alpha$ is expressed by

$$S_\alpha = G\,E\,T\,\rho_\alpha + G\,L_p + \delta \qquad (1)$$

G and $\delta$ represent gain and offset changes, respectively, in the response of the multispectral scanner instrument. E represents the irradiance through the atmosphere on the material, T represents the transmittance of the atmosphere over the path from the material to the scanner aperture, and $L_p$ represents the path radiance along this viewing path due to atmospheric scattering. $\rho_\alpha$ is the bidirectional reflectance of the material $\alpha$. All these variables are directly dependent on the wavelength of the signal being recorded, hence, there is no interaction between signals at different wavelengths, in principle, and each spectral band can be treated separately from the others.

Note that whenever the bidirectional reflectance of each material remains constant, the signals recorded are related to the reflectance of each material by a simple multiplicative and additive relationship, although to determine these multiplicative and additive factors by trying to estimate values for each variable in the transfer equation is by no means simple. If one postulates a reference condition in which the above multiplicative factors all equal unity and the additive factors all equal zero, and if one realizes that the inverse of a multiplicative and additive transformation (MAT) is itself multiplicative and additive and that the concatenation of two MAT's is likewise, overall, multiplicative and additive, one can conclude that the data transformation needed to compensate for any or all of the effects above (with bidirectional reflectance held constant) will also be multiplicative and additive. Furthermore, since there is no interaction between signals for different wavelengths, the required transformation may be determined separately for each spectral band.

One should be aware, however, that bidirectional reflectance does not, in general, remain constant for each material throughout a scene. Rather, reflectance is to be expected to vary differently for each material according to changes in illumination conditions (sun angle, relative intensities of direct and diffuse illumination), viewing angle, topography, crop or soil conditions (health of crop, density of ground cover, soil type, soil moisture content), crop orientation (due to wind), and cropping practice (methods of planting or harvesting). These effects, having a unique influence on the reflectance of each material, and varying sometimes from field to field or other times from county to county, cannot be fully compensated by a transformation applied indifferently to data from any and all materials in a scene. At best, one can devise a general transformation or means for data manipulation which treats these disparate effects only in an average way, or which takes advantage of some salient characteristics of the major materials of interest. (An example of the latter approach would be a classifier which takes advantage of multitemporal information and a knowledge of the characteristic growth cycle of a particular crop, e.g., winter wheat.) Variations in bidirectional reflectance should be recognized as one of the major potential stumbling blocks for signature extension. Other potential stumbling blocks are enumerated in the discussion below.

## SIGNATURE TRANSFORMATIONS

### Derivation

Signatures are usually represented by a gaussian probability density function of the form

$$P_\alpha = \frac{\exp\left[-\frac{1}{2}(x-\mu_\alpha)^t \theta_\alpha^{-1}(x-\mu_\alpha)\right]}{(2\pi)^{n/2}|\theta_\alpha|^{1/2}} \qquad (2)$$

$P_\alpha$ is the probability that a given signal x corresponds to the material $\alpha$, exclusive of any competing probability associated with other materials. x is the data vector representing the recorded signal levels in each spectral band of the MSS for a single measurement. $\mu_\alpha$ is the vector of mean values for the signature of material $\alpha$. $\theta_\alpha$ is the variance-covariance matrix for the signature of material $\alpha$. All the vectors have n components and the matrix has nxn components, with n being the number of

spectral bands used in the signature.

As a means to compensate for changes in bidirectional reflectance in an average way and to compensate for the multiplicative and additive effects arising from changes in the other variables of the transfer equation (1), a signature transformation may be proposed which alters signatures derived from one scene to match, at least approximately, the conditions present within a second scene. If one assumes that the difference between observed signal levels in the two scenes are purely multiplicative and additive, then the signals are related by

$$x' = A\,x + B \tag{3}$$

$x'$ represents the observed signal from the second scene, while $x$ represents a corresponding signal from the first scene. $A$ is a diagonal matrix with nxn components, representing the multiplicative changes to the signals in each spectral band, and $B$ is a vector with n components, representing the additive changes. The signature transformation corresponding to this multiplicative and additive change in signal levels is given by

$$\mu'_\alpha = A\,\mu_\alpha + B \tag{4}$$

and

$$\theta'_\alpha = A\,\theta_\alpha\,A \tag{5}$$

One should note that Eq. (5) applies only for data containing purely scenic information. In general, MSS data also contains non-scenic information, i.e., measurement noise inherent in the scanner instrument. When a signature is extracted from a scene, this measurement noise becomes a part of the variance-covariance statistics for the signature, changing those statistics from their purely scenic values in a strictly additive fashion. Ordinarily, signature extension is attempted between scenes recorded with the same MSS instrument, hence the measurement noise for each scene should be nearly the same, regardless of any changes in the variables of Eq. (1). Equation (5) should only apply to that portion of the variance-covariance statistics which excludes measurement noise. Depending on the source of the measurement noise, some other form of transformation may or may not be appropriate for the noise statistics. Since the nature of the measurement noise for LANDSAT data has not been determined, and since transforming the variance-covariance matrix produces little change in the results of signature extension applications, the policy at ERIM and at some other research laboratories so far has been not to use Eq. (5), leaving the variance-covariance statistics unchanged, and to use only Eq. (4) for signature transformations.

## Implementation

Given that a signature transformation is desired to compensate for multiplicative and additive changes between two scenes, the task is next to determine the appropriate coefficients, A and B, for Eq. (4). In general, one needs for this purpose some effective way to compare the data from the two scenes. One method for accomplishing this is to compare cluster statistics for the scenes. Clusters are multivariate gaussian probability density functions which, when weighted according to the amount of data in a scene which generated the statistics of each cluster, and when summed together, closely approximate the multivariate histogram distribution for the scene. Clusters are generally assumed to be equivalent to signatures for more or less unknown but distinct materials, which represent modes of the data distribution from which the clusters were generated. The extent to which clusters actually represent modes of the data distribution depends to a great extent on the nature of the clustering algorithm which is used, however, whatever algorithm is used, the clusters when taken together generally do represent adequately the variability to be found in the scene. The advantage in using cluster statistics for comparing data from scenes recorded under different conditions is that distinct materials by their presence give rise to representative clusters, but do not appreciably alter those clusters (mean values, variance, or covariance) according to the frequency of occurrence of the material within the scene. Hence, a valid comparison of recording conditions for two scenes requires only that clusters for similar materials be compared, rather than that the frequency of occurrence of the materials compared between scenes also be similar.

Once one has obtained a valid association between pairs of clusters from two scenes, a least squares estimate may be determined for the coefficients A and B of equation (4) by solving the following two equations once for each spectral band to be used.

$$\frac{\partial}{\partial A}\left[\sum_i (\mu'_i - A\,\mu_i - B)^2\right] = 0 \tag{6}$$

$$\frac{\partial}{\partial B}\left[\sum_i (\mu'_i - A\,\mu_i - B)^2\right] = 0 \tag{7}$$

$i$ is an index for identifying each cluster pair. The summations are over all cluster pairs. $\mu_i$ represents the mean value for the $i^{th}$ training scene cluster in the spectral band being considered, while $\mu'_i$ represents the mean value for the $i^{th}$ recognition scene cluster in the same spectral band. These equations lead to a pair of simultaneous linear equations which can be solved for the coefficients A and B in each spectral band, yielding

$$A = \frac{N\sum_i \mu_i\mu'_i - \sum_i \mu_i \sum_i \mu'_i}{N\sum_i \mu_i^2 - (\sum_i \mu_i)^2} \tag{8}$$

$$B = \frac{\sum_i \mu_i^2 \sum_i \mu'_i - \sum_i \mu_i \sum_i \mu_i\mu'_i}{N\sum_i \mu_i^2 - (\sum_i \mu_i)^2} \tag{9}$$

N is the total number of cluster pairs used in the regression. Again it should be realized that Equations (8) and (9) produce scalar values for A and B which are appropriate for the specific spectral band chosen. These equations need to be solved again for each additional spectral band used, to obtain the final A and B coefficient matrix and vector, respectively, indicated in equation (4).

Since the clusters which are paired in the regression to calculate A and B must be finite in number, there is a practical limit to the accuracy with which the A and B coefficients can be determined, even with all cluster pairs being valid. Of course the multiplicative and additive transformation sought cannot compensate perfectly for all the real physical causes of the change between the training scene and the recognition scene anyway, however in principle it is best to try to use as many valid cluster pairs in the regression as possible. Current signature extension tests at ERIM have tended to use between 10 and 20 cluster pairs for obtaining the A and B coefficients, out of a maximum of from 15 to 30 cluster pairs which were possible.

A first basic cluster matching algorithm, called MASC (for Multiplicative and Additive Signature Correction), was developed at ERIM to test the cluster regression approach to determining the A and B coefficients. While this algorithm achieved some occasional successes at signature extension, it did not include a means to adequately select only valid cluster pairs, a serious requirement for achieving practical results. The task was then to automate a procedure for selecting those few valid cluster pairs which might exist among the great many arbitrary pairs which were possible.

The difficulty involved in identifying valid cluster pairs may perhaps be partly appreciated by considering Figure 1, which shows a matrix representing all possible cluster pairs between a set of training scene clusters and a set of recognition scene clusters.

| Recognition Scene Clusters \ Training Scene Clusters | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | |
| 2 | 0 | | | | | | | | | |
| 3 | | | | | 0 | | | | | |
| 4 | | | | 0 | | | | | | |
| 5 | | | | | | 0 | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | 0 | | | |
| 8 | | | | | | | | | 0 | |
| 9 | | | | | | | | 0 | | |
| 10 | | | | | | | | | | 0 |

Figure 1. Myriad Potential Cluster Pairs

For the purpose of better illustrating a point to be brought up later, an equal number of training clusters and recognition clusters has been assumed, although the number of clusters obtained from each scene in practice turns out to be equal only occasionally. Also, for simplicity, a smaller than usual number of clusters has been assumed. The 0's in the matrix represent a hypothetical set of valid cluster pairs for this illustration. By ordering the sequence of the training scene and recognition scene clusters appropriately, these valid pairs may be made to fall close to the diagonal of the matrix, about as shown. If one tries to examine all possible sets of 10 cluster pairs to find which is best, one finds that there are 10! (3,628,800) sets of pairs to be considered, assuming that there are no multiple pairings with the same cluster. If one happens to guess that there will be only 8 valid pairs possible, then the number of sets of pairs to be considered increases by a factor of 45 (10!/8!/2!).

Obviously there are two basic difficulties to be dealt with in finding the valid cluster pairs from which to derive the required signature transformation. The first is to reduce the number of different sets of cluster pairs which need to be examined, and the second is to determine which among those several candidate sets of cluster pairs are most likely to be valid. The first attempt at ERIM toward solving the first of these two difficulties was to sort the training scene and recognition scene clusters according to their mean values in some designated spectral band, then to consider only those sets of cluster pairs which preserved that linear ordering. This procedure occasionally led to situations such as that shown in Figure 2.

| Recognition Scene Clusters \ Training Scene Clusters | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | X | | | | | | | | | |
| 2 | 0 | X | | | | | | | | |
| 3 | | | X | | 0 | | | | | |
| 4 | | | 0 | X | | | | | | |
| 5 | | | | | X | 0 | | | | |
| 6 | | | | | | X | | | | |
| 7 | | | | | | | ⊠ | | | |
| 8 | | | | | | | | X | 0 | |
| 9 | | | | | | | | 0 | X | |
| 10 | | | | | | | | | | ⊠ |

Figure 2. Limited Potential Cluster Pairs after Linear Ordering Constraint (example)

The X's indicate the one set of 10 cluster pairs that is permitted, subject to the cluster ordering constraint, when there is an equal number of training and recognition clusters from which to choose. The 0's again indicate the hypothetical set of valid cluster pairs specified in Figure 1. When the number of clusters in the training set differs from the number in the recognition set, the linear ordering constraint becomes less restrictive, as will be shown below. Note that of the 8 valid cluster pairs available, only two are within the candidate match indicated in Figure 2.

An improved cluster matching algorithm, called CROP-A (for Cluster Regression Ordered on Principal-Axis), was developed at ERIM and has evolved to include a partial remedy for the linear ordering constraint difficulty indicated in Figure 2. The name for this algorithm comes from its choice of the principal eigenvector of the covariance of the training signature means as the linear direction for the cluster ordering constraint. Cluster positions along this ordering axis are determined from an apparent mean value for each cluster, given by a

dot product between the cluster mean vector and a unit vector aligned with the principal eigenvector. Improvements in signature extension performance due to using this cluster ordering direction instead of using a particular spectral band appear to be mostly inconsequential, however the other new features contained in the algorithm appear to reap substantial benefits. In particular, the algorithm contains a provision to force a difference to occur in the number of training clusters and recognition clusters which are to be paired. For this purpose the algorithm keeps track of the number of data values used to generate each cluster. First, clusters generated from less than 1% of the data used to generate all clusters in the same set are excluded from being paired at all. This eliminates some of the "false alarm" clusters derived from minority constituents of a scene, which may be less likely to have counterparts in another scene. The percentage threshold for excluding clusters is then increased above 1% for one of the two sets of clusters (whichever requires the least number of additional exclusions) until a desired difference in the number of clusters remaining in the two sets is reached. Ordinarily the increased threshold is less than 2% when this condition is obtained. For cluster sets of between 15 and 30 clusters, a forced difference of 4 in the number of clusters is currently used, producing between 1000 and 30,000 candidate sets of cluster pairs. This situation is simulated in miniature in Figure 3.

| | | Training Scene Clusters | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | 1E | | | | | | | | | | |
| | 2 | ⊠ | X | X | X | | | | | | |
| | 3 | | X | X | X | ⊠ | | | | | |
| Recognition | 4E | | | O | | | | | | | |
| Scene | 5 | | | X | X | X | ⊠ | | | | |
| Clusters | 6E | | | | | | | | | | |
| | 7 | | | | X | X | X | ⊠ | | | |
| | 8 | | | | | X | X | X | X | O | |
| | 9 | | | | | | X | X | ⊠ | X | |
| | 10 | | | | | | | X | X | X | ⊠ |

Figure 3. Less Limited Potential Cluster Pairs after CROP-A Forced Difference

Recognition clusters eliminated by the requirement for a forced difference of 3 in the number of clusters in the two sets are designated (hypothetically) by the letter "E". The candidate cluster matches available from Figure 3, subject to the cluster ordering constraint, consist of sets of pairs designated by X's, one from each row, such that the chosen X's can be joined in sequence by a monotonic broken line segment. This requirement is equivalent to matching all possible subsets of 7 training clusters with the 7 retained recognition clusters, in sequence. In this simple case one obtains 120 (10!/7!/3!) candidate sets of 7 cluster pairs, rather than the single candidate (with 10 pairs) indicated in Figure 2. Also, one of the available candidates now contains 5 valid cluster pairs, compared to only 2 for the candidate in Figure 2. This new candidate is shown in Figure 4.

| | | Training Scene Clusters | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | 1E | | | | | | | | | | |
| | 2 | ⊠ | | | | | | | | | |
| | 3 | | | | | ⊠ | | | | | |
| Recognition | 4E | | | O | | | | | | | |
| Scene | 5 | | | | | | ⊠ | | | | |
| Clusters | 6E | | | | | | | | | | |
| | 7 | | | | | | | ⊠ | | | |
| | 8 | | | | | | | | X | O | |
| | 9 | | | | | | | | O | X | |
| | 10 | | | | | | | | | | ⊠ |

Figure 4. Optimum Candidate Cluster Match after CROP-A Forced Difference

Note that the pairing of recognition cluster #9 with training cluster #8, although potentially allowed by the CROP-A forced difference (Figure 3), would by its choice in a candidate exclude from that candidate, due to the ordering constraint, the valid pairings with recognition clusters #3, #5, and #7. Hence, at best this alternate candidate could only contain 3 valid pairs. This sort of limitation is not uncommon when a linear ordering constraint is used. The result is that not all of the valid cluster pairs can be selected by the algorithm at one time.

As a potential solution to the somewhat severe restrictions occasionally imposed by the CROP-A linear ordering constraint, another cluster matching algorithm, called CROWN (for Cluster Regression Ordered With N channels), is currently undergoing development and testing at ERIM. This algorithm uses a matrix of merit figures, one figure for each possible cluster pair, to allow apparent optimum cluster associations to be chosen one by one until a specified number of candidate sets of a fixed number of cluster pairs become available. The merit figures for the matrix are determined on the basis of similarities in the location of each training and recognition cluster within its respective overall cluster distribution. This technique appears to be satisfactory for reducing the complexity of the cluster matching problem without excluding any significant number of valid pairs from consideration.

Having devised a means to select a practical number of candidate cluster matches, one next needs to find the best candidate among those chosen and to determine which of the cluster pairs from that candidate are most likely to be valid. Both CROP-A and CROWN use the regression procedure itself to perform this selection. Presuming that invalid cluster pairs will tend not to match as closely as the valid pairs, these algorithms delete from the regressions performed for each candidate match those cluster pairs which appear to match the most poorly. This is accomplished by comparing the transformed training cluster mean values to the untransformed recognition cluster mean values for each cluster

pair. The mean values are first compared within the individual spectral bands as each separate regression is performed (equations (8) and (9)), since this is computationally the earliest opportunity to delete a cluster pair from the subsequent calculations. The cluster pair deleted after each iteration through the regression is the one among those with a difference in mean values in excess of a specified threshold, which has the largest difference in mean values. This iterative procedure continues until a stable situation is reached, with the regression in each spectral band updated to reflect deletions caused by the thresholding in any of the spectral bands. The RMS distance between the remaining cluster mean values is then tested, using an average over all spectral bands. If the greatest RMS distance is more than a second threshold, all cluster pairs with RMS distances greater than the average of the greatest RMS distance with this second threshold are deleted. The regressions are then updated accordingly and the test is repeated until once again the situation becomes stable. If at this point any of the deleted pairs now matches with an RMS distance less than a third threshold, the pair is restored and the regressions are updated just once more. This procedure has seemed to be quite effective. Candidate matches, with poorly matching cluster pairs deleted, are then compared to select the final result. The final result selected is that which has the minimum RMS mismatch between clusters, comparing averages over a specific fixed number of the "best" pairs from each candidate. Typically for CROP-A, this final selection is based on the best 67% of the cluster pairs in each match (whether deleted or not), while for CROWN it is based on the best 90%. Note, however, that the CROWN algorithm contains a provision to automatically select the number of cluster pairs which are reasonable to constitute a candidate, and that this number may sometimes be less than the number of pairs required for a CROP-A candidate, although the CROWN algorithm generally retains numerically more cluster pairs in its final result than does CROP-A.

Although the above candidate selection procedures and the subsequent iterated regressions with step by step deletions of poorly matched cluster pairs have seemed to be quite effective, it has for some time been apparent that the performance of cluster matching algorithms is limited by a fundamental difficulty somewhat allied with the problems caused by variations in bidirectional reflectance, mentioned earlier. This limitation occurs when there are an insufficient number of valid cluster pairs to be found, as happens when scenes contain dissimilar major constituents. Such major differences between scenes may arise simply from differences in crop varieties grown (different rates of growth), or from differences in crop treatment (fertilization or irrigation), as well as from more fundamental differences (different crops). Major differences between scenes constitute another potential stumbling block for signature extension. A method (partitioning) for partially alleviating this problem will be briefly discussed later.

## PERIPHERAL PROCEDURES

The manner in which a signature extension module, such as CROP-A or CROWN, is embedded in an overall signature extension system has been identified as an important consideration in determining its performance and value as a signature extension tool. In this regard research is currently underway at ERIM to define an optimum signature extension system, utilizing the current state of the art. Some particular techniques being tested are discussed below.

Since cluster matching algorithms in general use cluster statistics as their sole input, one might surmise that the manner in which cluster statistics are prepared may be an important consideration. Such is indeed the case. Since LANDSAT data is made up of many digitized data elements (commonly called pixels), each covering an area on the ground approximately 260 feet square, these pixels often contain a mixture of signals from more than one material. In fact, for scenes in Kansas which have many large fields one finds that 50% or more of the LANDSAT pixels straddle field boundaries and hence contain mixed signals. For cluster matching it is desirable to have cluster statistics which represent pure materials. Within a training scene, where the training field boundaries are known, one can cluster over pixels which are clearly within the field boundaries and thus obtain some relatively clean statistics, but within a recognition scene one is assumed not to have information on field boundaries, otherwise one could train locally and not need signature extension procedures. However, there are techniques for locating probable field boundaries in data for which there is no ground information. One of these techniques, which together with the subsequent clustering operation is called gradient filtered clustering, uses differences in the values of the 8 pixel neighbors to each pixel to compute a gradient value, indicating the amount of nonuniformity in the data adjacent to that pixel. A self adjusting threshold on the gradient value is then used to exclude a specified percentage (typically 75%) of the pixels, judged to be probable or possible mixtures, from clustering. While the remaining 25% of the pixels which are accepted may not represent all of the pure pixels which could be used, they generally represent a sufficient number of pure pixels for clustering and quite effectively exclude the mixtures. This technique permits the cluster matching algorithms to compare clusters for pure materials, increasing the validity of the good cluster pairs which can be found.

Still more improvement in signature extension performance might be expected to result from optimizing the way in which the transformed and untransformed clusters are used. With this in mind, ERIM has developed a technique called reverse transform labeling. This technique, rather than transforming training scene clusters to match the recognition scene, transforms the recognition scene clusters to match the training scene. The known training fields and the classification of the training scene by the transformed recognition clusters, together generate votes for labeling the recognition clusters. The

untransformed recognition clusters, with these labels, can then be used to classify the recognition scene. Since the recognition scene clusters (if gradient filtered) can be made to represent mostly pure materials, this technique only depends on obtaining a signature transformation accurate enough to obtain proper recognition cluster labels from the training scene information.

A third potential improvement in signature extension performance can be derived from developing the wisdom to know when and when not to try to use signature extension techniques. Earlier, the problem of training and recognition scenes with dissimilar major constituents was mentioned. The perhaps obvious solution to this problem is to use only training and recognition scenes which are sufficiently similar. The region of space and time over which one can successfully extend classification from a training scene, using signature extension techniques, is commonly called a stratum. The technique of estimating the number of strata and their boundaries in an area to be classified is called partitioning. The region of space and time which one uses to approximate a stratum is called a partition. Partitions may be static (if based on only general knowledge of an area, such as soil types and climate) or dynamic (if based on recent short term effects, such as the date of the last rainfall). The partitioning problem at present is highly complex and of course can vary substantially, depending on the signature extension techniques which are to be employed. Research is currently underway to determine to what extent the signature extension algorithms themselves can help to identify the boundaries of a partition.

## CONCLUSIONS

The preceding discussion has more or less followed the historical development of cluster matching techniques for signature extension at ERIM. An attempt has been made to indicate the theoretical boundaries which circumscribe signature extension efforts, and to indicate the step by step progress which has been achieved in cluster matching algorithms and in their use toward realizing the potential for timely, lower cost surveys over large areas, which the theory seems to offer. At this stage of its development, signature extension through the use of cluster matching algorithms appears to be a practical technique for economical and timely wheat surveys, using LANDSAT data, and certainly for other uses as well, provided that the reasonable limits to its use (partitions) can be adequately determined. All aspects of the signature extension problem are of course continually undergoing examination, testing, and development toward the goal of attaining a practical and fully operational implementation of a signature extension capability.