Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects

R. L. KETTIG AND D. A. LANDGREBE, SENIOR MEMBER, IEEE

IEEE TRANSACTIONS ON GEOSCIENCE ELECTRONICS, VOL. GE-14, NO. 1, pp. 19-26, JANUARY 1976

Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects

R. L. KETTIG AND D. A. LANDGREBE, SENIOR MEMBER, IEEE[1]

Abstract—A method of classification of digitized multispectral image data is described. It is designed to exploit a particular type of dependence between adjacent states of nature that is characteristic of the data. The advantages of this, as opposed to the conventional "per point" approach, are greater accuracy and efficiency, and the results are in a more desirable form for most purposes. Experimental results from both aircraft and satellite data are included.

I. INTRODUCTION

An important subject before the engineering and scientific community at the present time is the processing of scenes which represent tracts of the earth's surface as viewed from above. A typical scene may consist primarily of regular and/or irregular regions arranged in a patchwork manner, each containing one "class" of surface cover type. These homogeneous regions are the "objects" in the scene. A basic processing goal is to locate the objects, identify (classify) them, and produce tabulated results and/or a "type-map" of the scene. As in other image processing applications, the locations and spatial features (size, shape, orientation) of objects are revealed by changes in average spectral properties that occur at boundaries. But unlike most other applications, these spatial features often enable only a rough categorization of the object. Therefore classification is more often based on its spectral features using statistical pattern recognition techniques, a task for which the digital computer is well adapted.

Computer classification of multispectral scanner (MSS) data collected over a region is typically done by applying a "simple symmetric" decision rule to each resolution element (pixel). This means that each pixel is classified individually on the basis of its spectral measurements alone. A basic premise of this technique is that the objects of interest are large compared to the size of a pixel. Otherwise a large proportion of pixels would be composites of two or more classes, making statistical pattern classification unreliable; i.e., the prespecified categories would be inadequate to describe the actual states of nature. Since the sampling interval is usually comparable to the pixel size (to preserve system resolution), it follows that each object is represented by an array of pixels. This suggests a statistical dependence between consecutive states of nature, which the simple symmetric classifier fails to exploit. To reflect this property, we shall refer to simple symmetric classification as "no-memory" classification.

One method for dealing with dependent states is to apply the principles of compound decision theory or sequential compound decision theory. Abend [1] points out that a sequential procedure can be implemented fairly efficiently when the states form a low-order Markov chain. However, the prospect is considerably less attractive when they form a Markov mesh, which is a more suitable model for two-dimensional scenes. Furthermore, estimation of the state transition probabilities could be another significant obstacle to implementation of such a procedure.

---

[1] Manuscript received April 10, 1975; revised July 9,1975. This work was supported by the National Aeronautics and Space Administration, in part under Grant NGL 15-005-112 and m part under Contract NAS 9-14016.

R. L. Kettig was with the Laboratory for Applications of Remote Sensing, Department of Electrical Engineering, Purdue University, West Lafayette, IN 47907. He is now with the Communication Sciences Division, U.S. Naval Research Laboratory, Washington, DC 20375.

D. A. Landgrebe is with the Laboratory for Applications of Remote Sensing, Department of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

The compound decision formulation is a powerful approach for handling very general types of dependence. This suggests that perhaps by tailoring an approach more directly to the problem at hand, one can obtain similar results with considerable simplification. A distinctive characteristic of the spatial dependence in MSS data is "redundancy;" i.e., the probability of transition from state i to state j is much greater if $j = i$ than if $j \neq i$, because the sampling interval is generally smaller than the size of an object. This suggests the use of an "image partitioning" transformation to delineate the arrays of statistically similar pixels before classifying them. Since each homogeneous array represents a statistical "sample" (a set of observations from a common population), a "sample classifier" could then be used to classify the objects. In this way, the classification of each pixel in the sample is a result of the spectral properties of its neighbors as well as its own. Thus its "context" in the scene is used to provide better classification. The acronym ECHO (extraction and classification of homogeneous objects) designates this general approach.

A characteristic of both no-memory and compound decision techniques is that the number of classifications which must be performed is much larger than the actual number of objects in the scene. When each classification requires a large amount of computation, even the no-memory classifier can be relatively slow. An ECHO technique would substantially reduce the number of classifications, resulting in a potential increase in speed (decrease in cost).

The recent literature contains numerous references to image partitioning algorithms. Robertson [2] divides them into two main categories. "Boundary seeking" algorithms characteristically attempt to exploit object contrast. Two of these have been implemented with MSS data [3], but they are incompatible with sample classifiers due mainly to their failure to produce boundaries that always close on themselves. The other category can be called "object seeking" algorithms, which characteristically exploit the internal regularity (homogeneity) of the objects. As the name implies, an object seeking algorithm always produces well-defined samples (and thus closed boundaries as well). There are two opposite approaches to object seeking, which we shall call conjunctive and disjunctive. A conjunctive algorithm begins with a very fine partition and simplifies it by progressively merging adjacent elements together that are found to be similar according to certain statistical criteria [4], [5]. A disjunctive algorithm begins with a very simple partition and subdivides it until each element satisfies a criterion of homogeneity. For example, Robertson's algorithm [2], [6] is based on the premise that if a region contains a boundary, splitting the region arbitrarily will usually produce two sub-regions with significantly different statistical characteristics.

We combined Rodd's [5] conjunctive partitioning algorithm with a minimum distance sample classifier and observed an improvement in classification accuracy over conventional no-memory classification, but processing time was increased [7] . Gupta and Wintz [8] added a test of second order statistics to Rodd's first order test, but obtained essentially the same results as the first order test at greater cost in processing time. Robertson [2], [6] implemented a disjunctive partitioning algorithm with the same minimum distance classifier. He obtained about the same classification accuracy as conventional no-memory classification with an order of magnitude increase in processing time.

The current investigation is devoted to further development and testing of the conjunctive approach. Major changes in both the classification and partitioning strategies have resulted in significant improvements in accuracy, stability, and speed.

II. SAMPLE CLASSIFICATION

A typical scene is assumed to consist primarily of objects whose boundaries form a partition of the scene. Each object in the partition belongs to one of K classes. Let W_i denote the event that an object belongs to class i. As previously indicated, we ignore any statistical dependence of this event on the size, shape, and location of the object. We rely instead on its spectral features. Each pixel in an object is a q-dimensional random variable, where q denotes the number of spectral measurements per pixel. It is commonly assumed that the q-variate, marginal, probability density

function (pdf) of a pixel, $\mathbf{X}$, depends only on the class of the object containing $\mathbf{X}$. This is due to the homogeneity of the types of objects typically encountered in remote sensing applications. $p(\mathbf{x}|W_i)$, $x \in R^q$, denotes this class-conditional density function for the i^{th} class. Another common assumption is that the classes can be defined such that $p(\mathbf{x}|W_i)$ is approximately multi-variate normal (MVN); i.e.,

$$p(\mathbf{x}|W_i) = N(\mathbf{X}; \mathbf{M}_i, \mathbf{C}_i) \equiv (|2\pi\mathbf{C}_i| \exp((\mathbf{x} - \mathbf{M}_i)^t\mathbf{C}_i^{-1}(\mathbf{x} - \mathbf{M}_i)))^{-1/2}$$

for some q-dimensional positive-definite, covariance matrix $\mathbf{C}_i$ and some mean vector $\mathbf{M}_i \in R^q$. Parametric estimates of these density functions are obtained by estimating $\mathbf{M}_i$ and $\mathbf{C}_i$ from sets (samples) of training data supplied for each class.

Two pixels in spatial proximity to one-another are unconditionally correlated, with the degree of correlation decreasing as the distance between them increases. Much of this correlation is attributable to the effect of dependent states mentioned in the previous section, which is the effect we wish to exploit. For simplicity we shall ignore other sources of correlation. Thus we assume class-conditional independence (as does the compound decision approach).

If $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n) \in R^{nq}$ represents a set of pixels in some object, then this set constitutes a "sample" from a population characterized by one of the class-conditional pdf's. A sample classifier is simply a strategy for deciding which one, based on the n observations. One popular approach is the "minimum distance (MD) strategy" [9]. In MD classification, the n data vectors are used to estimate the pdf of the population, and the class is chosen whose pdf is closest to this estimate as measured by some appropriately defined "distance measure" on the set of density functions. A popular distance measure is the Bhattacharyya distance, which for $N(\mathbf{X}; \mathbf{M}_i, \mathbf{C}_i)$ and $N(\mathbf{X}; \mathbf{M}, \mathbf{C})$ is given by:

$$B = \frac{1}{4}\left(\ln\frac{(|\mathbf{C}_i + \mathbf{C})/2|^2}{|\mathbf{C}_i|\,|\mathbf{C}|} + \mathrm{tr}\,((\mathbf{C}_i + \mathbf{C})^{-1}(\mathbf{M}_i - \mathbf{M})(\mathbf{M}_i - \mathbf{M})^t)\right) \qquad (1)$$

A drawback of the MD approach is that it fails for small n, because the density estimate becomes degenerate.

Our preference is the maximum likelihood (ML) strategy which assigns $\mathbf{X}$ to class i if

$$\ln p(\mathbf{X}|W_i) = \max_j \ln p(\mathbf{X}|W_j).$$

Due to the assumption of class-conditional independence, these quantities can be computed as:

$$\ln p(\mathbf{X}|W_i) = -\frac{1}{2}\,\mathrm{tr}\,(\mathbf{C}_i^{-1}\mathbf{S}_2) + \mathbf{M}_i^t\mathbf{C}_i^{-1}\mathbf{S}_1 - \frac{1}{2}\,n(\mathbf{M}_i^t\mathbf{C}_i^{-1}\mathbf{M}_i + \ln|2\pi\mathbf{C}_i|)$$

$$\mathbf{S}_1 = \sum_{i=1}^{n} \mathbf{X}_i\,, \qquad \mathbf{S}_2 = \sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i^t\,. \qquad (2)$$

Of course: $\mathbf{M}=\mathbf{S}_1/n$ and $\mathbf{C}=\mathbf{S}_2/n - \mathbf{M}\mathbf{M}^t$. Formula (2) is much faster to compute than formula (1) for each (S_1, S_2) pair, once the non-data-dependent constants have been initialized. Thus the ML strategy is computationally efficient. Another important property is that it does not fail for small n. On theoretical grounds, for the idealized conditions we have stated, it is the optimum strategy (for minimum error rate) when the a priori class probabilities are equal. Also, the Chernoff bound for

ML no-memory classification (n = l) can be extended to provide an error bound for ML sample classification that is a sum of exponentially decreasing functions of the sample size. Experimentally the two strategies appear about equal in terms of accuracy, with the ML strategy possibly having a slight advantage.

As a matter of theoretical interest, it can be shown that use of the ML strategy gives the same results (with less computation) as an MD strategy using one of the Kullback-Leibler numbers, if $|\mathbf{C}\underline{I}| > 0$. (If $|\mathbf{C}\underline{I}| = 0$, the K-L number is undefined, but the ML strategy is still valid.)

III. IMAGE PARTITIONING

The basic approach that we have adopted (due to Rodd [5]) consists of two "levels" of tests. Initially the pixels are divided, by a rectangular grid, into small groups of four (for example). At the first level of testing, each group becomes a unit called a "cell," provided that it satisfies a relatively mild criterion of homogeneity. Those groups that are rejected are assumed to overlap a boundary and their individual pixels are classified by the no-memory method. These groups are referred to as "singular" cells. At this level it is usually desirable to maintain a fairly low rejection rate to reflect the relatively high *a priori* probability of a group being homogeneous. The goal at this level is essentially the same as the goal of the boundary seeking techniques mentioned previously; i.e., to detect as many pixels as possible that lie along boundaries without requiring that the ones detected form closed contours or even be connected.

At the second level, an individual cell is compared to an adjacent "field," which is simply a group of one or more connected cells that have previously been merged. If the two samples appear statistically similar by some appropriate criterion, then they too are merged. Otherwise the cell is compared to another adjacent field or becomes a new field itself. By successively "annexing" adjacent cells, each field expands until it reaches its natural boundaries, where the rejection rate abruptly increases, thereby halting further expansion. The field is then classified by a sample classifier, and the classification is assigned to all its pixels.

This approach has the important advantage that it can be implemented "sequentially;" i.e., raw data need be accessed only once and in the same order that it is stored on tape. This is important for practical, rather than theoretical, considerations. The flow chart in Figure 1 indicates how it can be done. In this chart, the top of the scene is referred to as north, and the general processing sequence is from north to south.

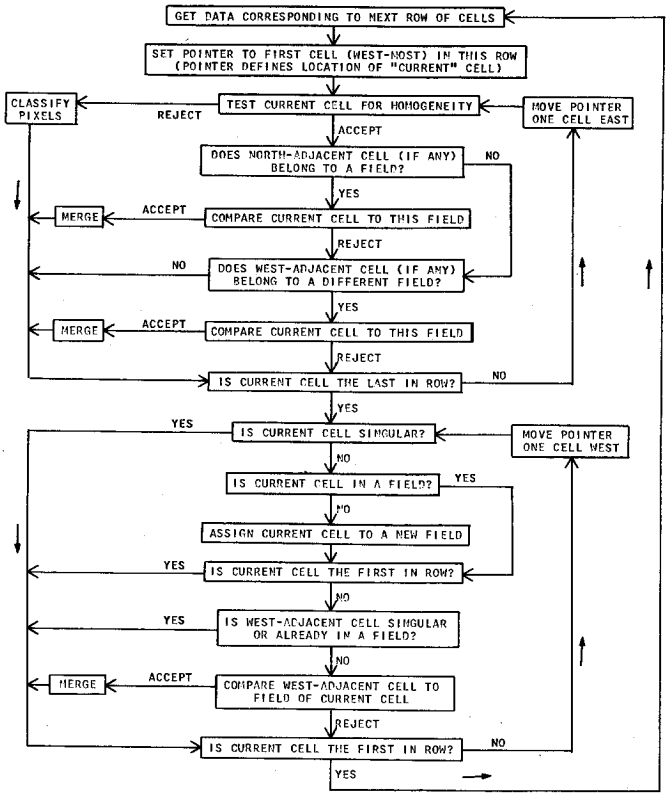


Fig. 1. Basic flow chart for a two-level, conjunctive, partitioning algorithm.

Many modifications to the basic flow chart are, of course, possible. One of the modifications we use involves comparing a cell to as many as three different fields at once (seeking the best "match"), instead of one-at-a-time.

Annexation Criterion

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ represent the pixels in a group of one or more cells which have been merged by successive annexations. Let $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$ represent the pixels in an adjacent, non-singular cell. Since both $\mathbf{X}$ and $\mathbf{Y}$ have satisfied certain criteria of homogeneity, we assume that each is a sample from a MVN population. Let f and g represent the corresponding density functions. It is desired to test the (null) hypothesis that f =g. This is a composite hypothesis, since it does not specify f and g. The "likelihood ratio procedure" [10] provides an effective statistic for testing this hypothesis. Van Trees [11] refers to it as the "generalized likelihood ratio." Let

$$H_o(\mathbf{X},\mathbf{Y}) = \{p(\mathbf{x},\mathbf{y}|f,g): g = f, \quad f \in \Omega\}$$

$$H_1(\mathbf{X},\mathbf{Y}) = \{p(\mathbf{x},\mathbf{y}|f,g): f \in \Omega, \quad g \in \Omega\}$$

where $p(\mathbf{x},\mathbf{y}|f,g)$ is the conditional joint density of $\mathbf{X}$ and $\mathbf{Y}$ evaluated at $\mathbf{x} \in R^{nq}$ and $\mathbf{y} \in R^{mq}$, and Ω is a set of MVN density functions. The assumption of class-conditional independence enables us to express the joint-density of pixels as the product of their marginal densities. Thus:

$$p(\mathbf{x},\mathbf{y}|f,g) = p(\mathbf{x}|f)p(\mathbf{y}|g) = \left(\prod_{i=1}^{n} f(\mathbf{x}_i)\right)\left(\prod_{i=1}^{m} g(\mathbf{y}_i)\right)$$

The generalized likelihood ratio is given by:

$$\Lambda = \frac{\sup H_o(\mathbf{X},\mathbf{Y})}{\sup H_1(\mathbf{X},\mathbf{Y})} = \frac{\max_{f\in\Omega} p(\mathbf{X}|f)\, p(\mathbf{Y}|f)}{\max_{f\in\Omega} p(\mathbf{X}|f)\ \max_{f\in\Omega} p(\mathbf{Y}|g)}$$

For an "unsupervised" approach to partitioning we take Ω to be the following set of functions of $\mathbf{x}\in R^q$:

$$\Omega = \{N(\mathbf{x};\mathbf{M},\mathbf{C}): \mathbf{M}\in R^q,\quad \mathbf{C} = \text{symmetric and positive-definite}\}$$

Anderson [12] shows that:

$$\Lambda = \Lambda_1 \cdot \Lambda_2 \tag{3}$$

where

$$\Lambda_1 = (|A| /|B|)^{N/2} \tag{4}$$

$$\Lambda_2 = (|A_x/n|^n\, |A_y/m|^m/|A/N|^N)^{1/2} \tag{5}$$

$$N = n + m$$

$$\overline{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{X}_i/n$$

$$\overline{\mathbf{Y}} = \sum_{i=1}^{m} \mathbf{Y}_i/m$$

$$A_x = \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^t$$

$$A_y = \sum_{i=1}^{n} (\mathbf{Y}_i - \overline{\mathbf{Y}})(\mathbf{Y}_i - \overline{\mathbf{Y}})^t$$

(In order to assure non-singular matrices with probability one, we need $n > q$, $m > q$ [12] .)

$$A = A_x + A_y$$

$$\mathbf{M} = (n\,\overline{\mathbf{X}} + m\,\overline{\mathbf{Y}})/N$$

$$B_x = \sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{M})(\mathbf{X}_i - \mathbf{M})^t = A_x + n(\overline{\mathbf{X}} - \mathbf{M})(\overline{\mathbf{X}} - \mathbf{M})^t$$

$$B_y = \sum_{i=1}^{m} (\mathbf{Y}_i - \mathbf{M})(\mathbf{Y}_i - \mathbf{M})^t = A_y + n(\overline{\mathbf{Y}} - \mathbf{M})(\overline{\mathbf{Y}} - \mathbf{M})^t$$

$$B=B_x + B_y = A + \frac{nm}{N}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})(\overline{\mathbf{X}} - \overline{\mathbf{Y}})^t$$

Anderson also suggests modifying Λ by replacing the number of pixels in each sample by the number of degrees of freedom; i.e., replace *n* by *n - I, m* by *m* - 1, and *N* by *N* - 2 in formulas (4) and (5). In either case, the statistics are invariant with respect to a linear transformation on the data vectors. It follows that their distributions under the null hypothesis are independent of the actual MVN population from which the samples are drawn.

Therefore we can construct a significance test of the null hypothesis. Λ_1 and Λ_2 are independent under the null hypothesis [12], so the procedure we use is to test Λ_1 at significance level α_1 and Λ_2 at level α_2 and reject the null hypothesis if either test produces a rejection. (Cooley and Lohnes [13] give transformations of Λ_1 and Λ_2 (the modified versions) with F-distributions under the null hypothesis.) The overall significance level is then $\alpha = 1 - (1 - \alpha_1)(1 - \alpha_2)$. Essentially, Λ_2 tests the hypothesis of equal covariance matrices (second order statistics), and α_1 tests the hypothesis of equal mean vectors (first-order statistics).

These multivariate (MV) tests have the same weakness as MD classification, namely the problem of estimating a MVN density from a relatively small sample (sometimes known as the "dimensionality" problem). This led to the constraint m > q, a condition which is often not met. Even when the condition is met, poor estimates can result, leading to decision errors. One approach to this problem is to reduce q by deleting features. It is well-known, for example, that a subset of features used to train a classifier from small training samples can sometimes produce better classification results than the full set. With this approach, however, one is faced with the problem of choosing the subset.

Another approach is to base the decision on the q, univariate, marginal distributions; i.e., simply consider the data in one spectral channel at a time. This has been termed a "multiple univariate" (MUV) approach. In each channel we test the univariate hypothesis that the means and variances of the two samples are equal. Since the boundaries may be strong in some spectral channels and weak in others, we accept the null hypothesis only if the univariate hypothesis is accepted in all q channels. Besides avoiding the dimensionality problem, the MUV procedure requires less computation and simpler distribution theory. However, it must be pointed out that in situations where class separability is primarily a multivariate effect, the MV procedure may be more advantageous.

For a "supervised" approach to partitioning we take Ω to be:

$$\Omega = \{p(x|W_i): i = 1, \cdots, K\}.$$

This greatly simplifies each hypothesis, but paradoxically the resultant test criterion is much more complicated:

$$\Lambda = \frac{\max_i p(\mathbf{X}|W_i)\ p(\mathbf{Y}|W_i)}{\max_i p(\mathbf{X}|W_i)\ \max_i p(\mathbf{Y}|W_i)} \qquad (6)$$

This is a multivariate statistic without the constraint m >q that was necessary in the unsupervised mode. However, the maxima in formula (6) cannot be expressed in a simple analytic form as in (3). They can only be obtained by exhaustive search. Furthermore, the distribution of (6) is unknown under either hypothesis, because it depends on the true classes of **X** and **Y**. But in return we gain a statistic which should be more "sensitive" to the presence or absence of a boundary.

This should produce better performance and make the specification of a decision threshold less critical. In fact, the experimental results indicate that the threshold need not be a function of n, the current size of sample **X**, in order to obtain good results. Furthermore, the results tend to be fairly stable over several orders of magnitude of threshold variation. Thus we will find it convenient to represent the decision threshold as

$$T = 10^{-t} \quad t \geq 0$$

In other words, we reject the null hypothesis if $\Lambda < T$ or equivalently $-\log \Lambda > t$. Otherwise we accept it. Experimentally we investigate the effect of different values of t on performance.

Cell Selection Criterion

"Cell selection" refers to the Level-1 test which is used to detect cells that overlap boundaries. Such cells frequently exhibit abnormally large variances. Thus, in the unsupervised mode, we say that a cell is singular if the ratio of the square root of the sample variance to the sample mean falls above some threshold, c, in any channel.

In the supervised mode we call a cell singular if $Q_j(Y) > c$, where:

$$Q_j(Y) = \mathrm{tr}\left(C_j^{-1}\sum_{i=1}^{m} \mathbf{Y}_i \mathbf{Y}_t^i\right) - 2\mathbf{M}_t^j \; C_j^{-1}\sum_{i=1}^{m} \mathbf{Y}_i + m\cdot\mathbf{M}_j^t \, C_j^{-1} \, \mathbf{M}_j$$

where j is such that:

$$\ln p(\mathbf{Y}|W_j) = \max_i \ln P(\mathbf{Y}|W_i)$$

$$= \max_i -\frac{1}{2}(m \ln |2\pi C_i| + Q_i(\mathbf{Y}))$$

The decision rule is to accept the hypothesis that **Y** is homogeneous if $Q_j(\mathbf{Y}) < c$, where c is a pre-specified threshold. Otherwise the hypothesis is rejected. This criterion has the particular advantage that it tends to reject not only inhomogeneous cells, but "unrecognizable" cells as well. (Unrecognizable cells are those which represent spectral classes that the classifier has not been trained to recognize.) Another advantage of this criterion is that its use of the log-likelihood function makes it especially compatible with the supervised annexation criterion and the ML sample classifier.

As a final note, the distribution function $P(Q_j(\mathbf{Y}) > c|W_i)$ is chi-squared with mq degrees of freedom. This can be used to provide initial guidance in choosing c.

IV. EXPERIMENTAL RESULTS

Two aircraft and two LANDSAT-1 data sets, for which large amounts of training and test data are available, were classified by the following six methods:

1. Conventional ML No-Memory Classification [14]

2. Supervised Cell Selection only (t = 0); ML Sample Classification

3. "Optimized" MUV Unsupervised Partitioning; ML Sample Classification

4. Supervised Partitioning (t = 4); ML Sample Classification

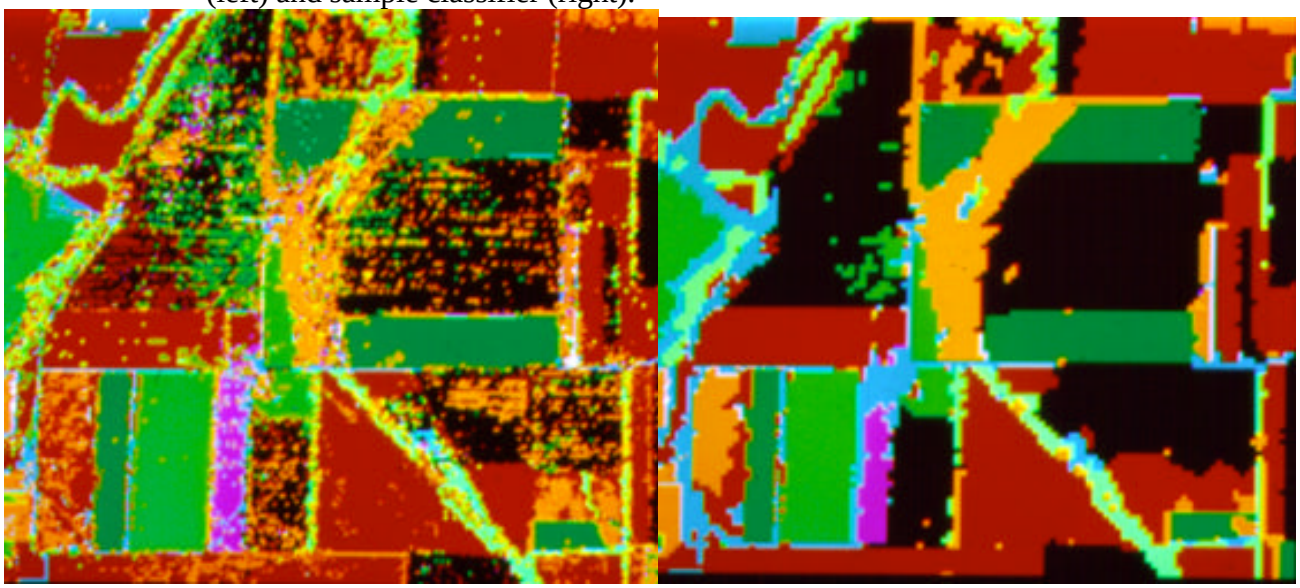
5. ML Sample Classification of Test Areas Only

6. MD (Bhattacharyya) Sample Classification of Test Areas Only [14].

The cell size for #2-#4 was fixed at 2 X 2 pixels, which is the minimum allowed in the unsupervised mode.

A qualitative assessment of the results is provided by Figures 2 and 3. Figure 2 (left side) shows a section of aircraft data that has been classified by method #1. Each class has been assigned a gray level, and each pixel has been displayed as the gray level assigned to its classification. A great deal of "classification noise" is readily apparent. In contrast to this, Figure 2 (right side) shows the same section as classified by method #4. The random errors have, for the most part, been eliminated. This map is much closer to the desired "type map" form of output that is generally desired.

Fig. 2. Gray-scale-coded classification maps produced by no-memory classifier (left) and sample classifier (right).



Pixel Classifier Result ECHO Classifier Result

Figure 3 shows the centers of these two maps in greater detail. Each class is represented by an assigned symbol and each symbol represents one pixel. The four rectangular areas are test areas designated as wooded pasture (displayed as a blank). The diversity of symbols in the test areas testifies to the inadequacy of the no-memory method for classifying this section, whereas most of the confusion is avoided by the ECHO technique.

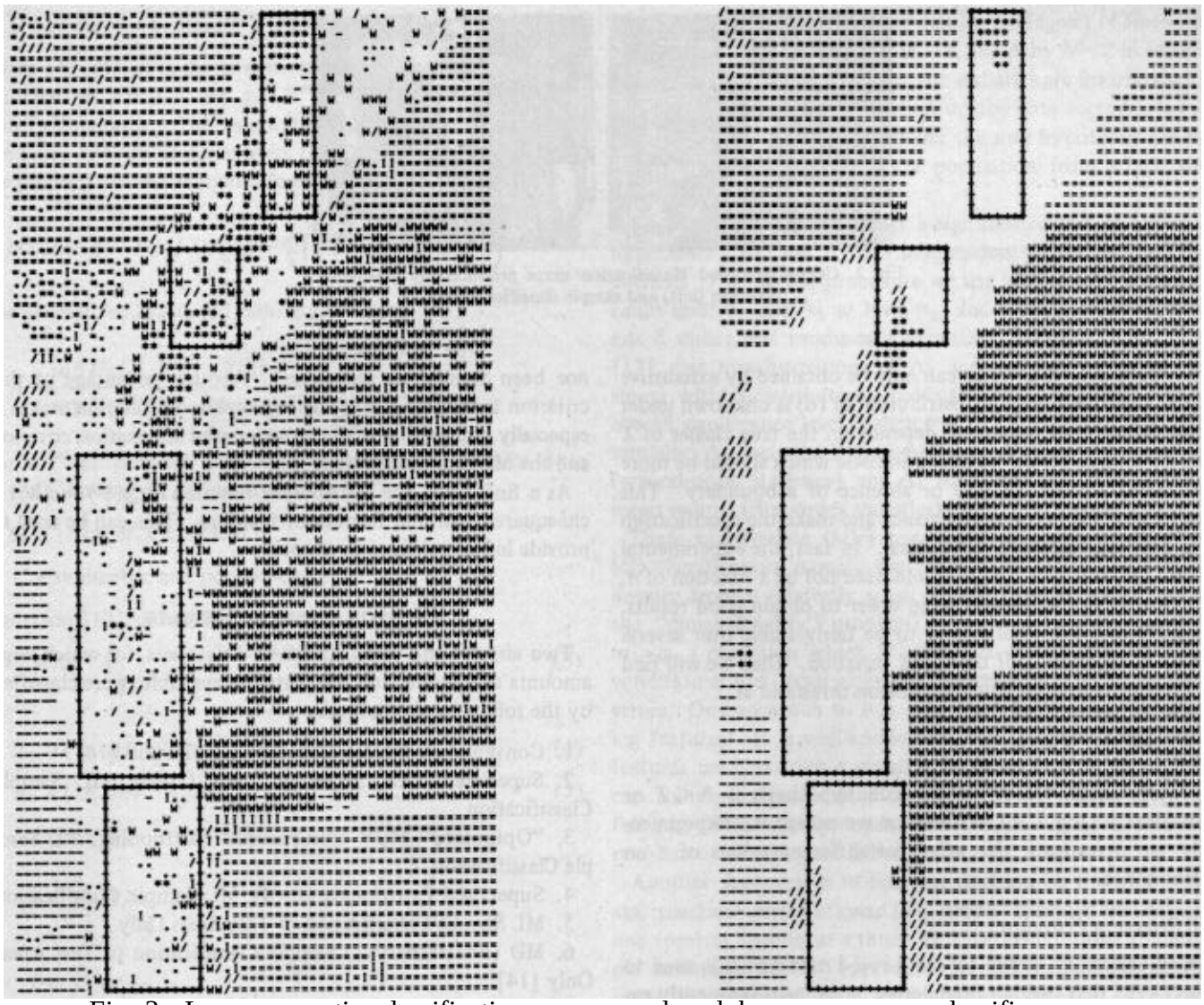


Fig. 3. Logogrammatic classification maps produced by no-memory classifier (left) and sample classifier (right).

The estimated probability of error for each method gives an important quantitative measure of performance. It is obtained as the ratio of the number of misclassified pixels in the test areas to the total number of pixels in the test areas. Figure 4 shows results obtained for each of the four data sets.[2] The results are about what one would expect. Method #1 consistently has the highest error rate because of its lack of use of spatial dependence. #2 uses some spatial information and consistently does somewhat better than #1. #3 uses more spatial information, which accounts for its improvement over cell selection alone, and #4 does consistently better than #3 because it uses more of the available information in the partitioning phase.

[2] Each data set contains different classes from the general categories: agriculture, forest, town, mining, and water. Refer to reference [15] for details.

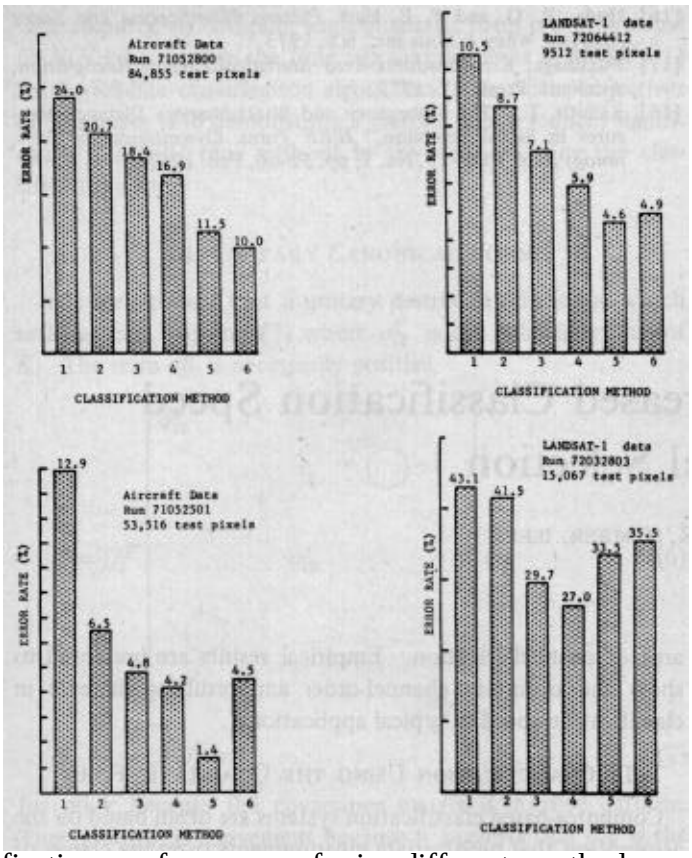


Fig. 4. Classification performance of six different methods applied to four different data sets.

#5 and #6 usually provide the best performance, because they are given more a *priori* information to begin with. One reason for including them here is to determine if either provides a distinct advantage over the other. On 3 of the 4 data sets, maximum likelihood sample classification achieved lower error rates than the minimum Bhattacharyya distance strategy. The differences are small, however. This justifies our use of the ML strategy in #2-#4. Another reason for including them is that the performance of #5 provides a "goal" (but not a bound) for the performance of #3 and #4; i.e., the nearness of the performance to this goal is an indication of the effectiveness of the partitioning process alone.

Although #3 appears to be fairly close to #4 in general, it must be pointed out that the "optimum" combination of α_1 and α_2 which achieves this performance is somewhat unpredictable at this time. All that we can say of a general nature is that α_1 tends to be effective at about .005 and α_2 at a smaller value such as .001 or 0.

The results for the supervised mode, however, are much more stable. Figure 5 shows only the results for t = 4, which are not always the optimum results, but they are within 1% of the optimum in all 4 cases. Figure 5 shows a typical example of the effect of t on classification error rate.

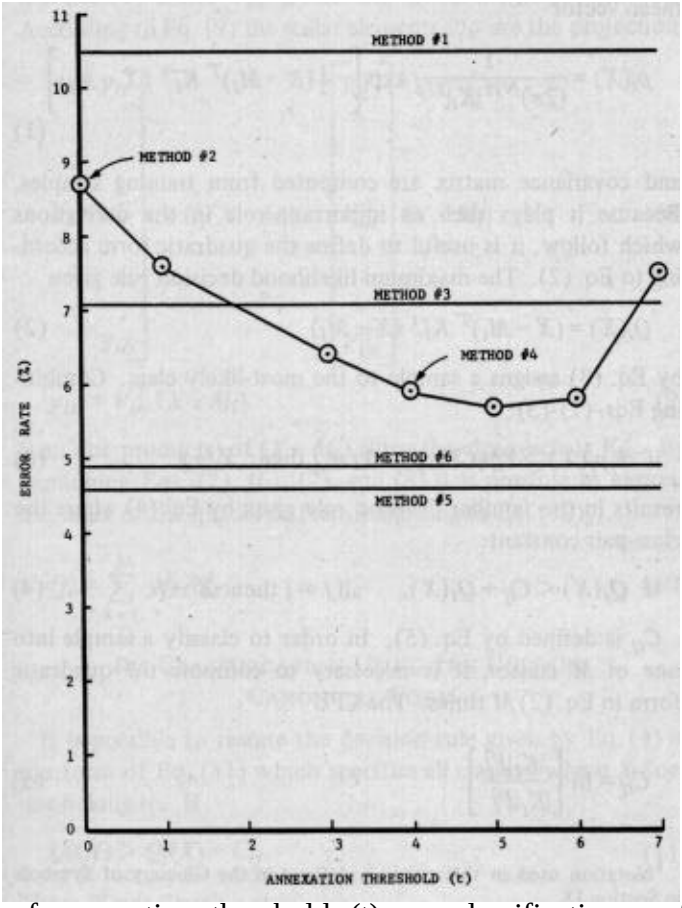


Fig. 5. Effect of annexation threshold (t) on classification performance—run 72064412.

The results are not a sensitive function of the Level-1 threshold, c. The values c = .25 (unsupervised mode) and c = 15q (supervised mode, $3 \leq q < 6$) usually provided the desired effect.

The main advantage of the unsupervised mode appears to be speed, when classification complexity is reasonably high. This is because the time saved by classifying pixels collectively can more than compensate for the time required to partition. For a LANDSAT-1 data set classified with 4 channels and 14 spectral classes, processor #3 required 22% less CPU time than #1, in spite of the fact that the classification subroutine in #1 is coded in assembler language for peak efficiency. (It has been estimated that this increases its efficiency by about 50%.) #3 and #4 are just developmental versions coded in Fortran. But for an aircraft data set with 6 channels and 17 spectral classes, #4 required 26% less time and #3 required 56% less time than #1.

V. CONCLUSION

We have successfully exploited the redundancy of states that is characteristic of sampled imagery of ground scenes to achieve better accuracy and reduce the number of actual classifications required. The only training used is the same as that required by a conventional maximum likelihood, no-memory classifier; i.e., estimates of the class- conditional, marginal densities for a single pixel. Thus we have not relied on specific spatial features, textural information (class-conditional spatial correlation), or on the contextual information associated with spatial relationships of objects.

REFERENCES

[1] Abend, K., "Compound Decision Procedures for Pattern Recogni*tion," Proc. NEC,* 22, pp. 777-780,1966.

[2] Robertson, T. V., K. S. Fu, P. H. Swain, "Multispectral Image Partitioning," Ph.D. Thesis #25970, School of Electrical Engineering, Purdue University, West Lafayette, IN, August, 1973. Also LARS Information Note #071373.

[3] Anuta, P. E., "Spatial Registration of Multispectral and Multitemporal Imagery Using Fast Fourier Transform Techniques," *IEEE Trans. Geoscience Electronics,* Vol. GE-8, No. 4, pp. 353-368, Oct. 1970.

[4] Muerle, J. L. and D. C. Allen, "Experimental Evaluation of Techniques for Automatic Segmentation of Objects in a Complex Scene," *Pictorial Pattern Recognition,* G. C. Cheng, R. S. Ledley D. K. Pollock, A. Rosenfeld (eds.), Thompson Book Co., Washington, DC, pp. 3-13,1968.

[5] Rodd, E. M., "Closed Boundary Field Selection in Multispectral Digital Images," IBM Publication No. 320.2420, Jan. 1972.

[6] Robertson, T. V., "Extraction and Classification of Objects in Multispectral Images," *Proceedings of the Conference on Machine Processing of Remotely Sensed Data, Purdue University,* West Lafayette, IN, Section 3B, pp. 27-34, Oct. 1973.

[7] Kettig, R. L. and D. A. Landgrebe, "Automatic Boundary Finding and Sample Classification of Remotely Sensed Multispectral Data," LARS Information Note 041773, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN, April 1973.

[8] Gupta, J. N. and P. A. Wintz, "Closed Boundary Finding, Feature Selection, and Classification Approach to Multi-Image Modeling," LARS Information Note 062773, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN, June 1973.

[9] Wacker, A. G. and D. A. Landgrebe, "The Minimum Distance Approach to Classification," Ph.D. Thesis #24733, School of Electrical Engineering, Purdue University, West Lafayette, IN, Jan. 1972. Also LARS Information Note #100771.

[10] Lehmann, E. L., *Testing Statistical Hypotheses,* Wiley & Sons Inc., NY, 1959.

[11] Van Trees, H. L., *Detection, Estimation, and Modulation Theory, Part 1,* Wiley & Sons Inc., NY, 1968.

[12] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis,* Wiley & Sons Inc., NY, 1958.

[13] Cooley, W. W. and P. R. Lohnes, *Multivariate Data Analysis,* Wiley & Sons Inc., NY, 1971.

[14] Phillips, T. L. (ed.), *LARSYS Version 3 User's Manual, Vol. 2,* Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN, June 1973.

[15] Kettig, R. L., "Computer Classification of Remotely Sensed Multispectral Image Data by Extraction and Classification of Homogeneous Objects," Ph.D. Thesis, School of Electrical Engineering, Purdue University, West Lafayette, IN, May 1975. Also LARS Information Note #050975.

[16] Duda, R. O. and P. E. Hart, *Pattern Classification and Scene Analysis,* Wiley & Sons Inc., NY, 1973.

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

[18] Kailath, T., "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Trans. Communication Technology,* Vol. COM-15, No. 1, pp. 52-60, Feb. 1967.