1-1-1980

Context Distribution Estimation for Contextual Classification of Multispectral Image Data

James C. Tilton

Philip H. Swain

Stephen B. Vardeman

Reprinted from

Symposium on

Machine Processing of

Remotely Sensed Data

and

Soil Information Systems

and

Remote Sensing and Soil Survey

June 3-6, 1980

Proceedings

The Laboratory for Applications of Remote Sensing

Purdue University
West Lafayette
Indiana 47907 USA

IEEE Catalog No.
80CH1533-9 MPRSD

LARS Technical Report 040280

ERRATA

(P. 2) P. 172, Col. 2, Line 27:
Change northeast to northwest

(P. 3) P. 173, Col. 1, Line 4:
Change 25 to 26

(P. 3) P. 173, Col. 2, Line 57:
Change configuration to classification

(P. 5) Page 175, Figure 1, Line 5:
Insert after "no-context": classification

(P. 5) Page 175, Figure 2, Line 5:
Change distribution to classification

CONTEXT DISTRIBUTION ESTIMATION FOR CONTEXTUAL CLASSIFICATION OF MULTISPECTRAL IMAGE DATA

JAMES C. TILTON, PHILIP H. SWAIN,
AND STEPHEN B. VARDEMAN
Purdue University

ABSTRACT

A classification algorithm incorporating contextual information in a general, statistical manner is presented. Methods are investigated for obtaining adequate estimates of the context distribution (a statistical characterization of context) upon which the classification algorithm depends. Finally, a method of estimating optimal algorithm parameters prior to performing preliminary classifications is explored.

I. INTRODUCTION

The most widely used method for classifying remotely sensed data from such sources as multispectral scanners on aircraft or satellite platforms is a point-by-point classification technique in which data from each pixel in the scene are classified individually by a maximum likelihood classifier [1]. The information normally used by this classifier is only spectral or, in some cases, spectral and temporal. There generally is no provision for using contextual information.

In contrast, when scanner data are displayed in image form, a human analyst routinely uses context to help decide what is in the imagery. Using context, he may be able to easily pick out roads, delineate boundaries of agricultural fields, and differentiate between grass in an urban setting (lawns) and grass in an agricultural setting (pasture or forage crops) where a maximum likelihood point classifier would have much difficulty in doing so.

Recently we have developed a classification algorithm which incorporates contextual information in a general, statistical manner [2]. This algorithm exploits the tendency alluded to above of certain ground-cover classes to be more likely to occur in some contexts than in others.

An estimate of the "context distribution" (a statistical characterization of the context in the scene to be classified) must be made before this classification algorithm can be used. Methods are investigated here for obtaining sufficiently accurate estimates of the context distribution. The process of estimating the context distribution can involve a large number of preliminary classifications using the statistical context classifier. With the goal of limiting the number of preliminary classifications needed, a method of predicting the optimal algorithm parameters without performing classifications is explored.

II. THE CLASSIFICATION MODEL

Remote sensing imaging systems generally provide data in the form of a two-dimensional array of $N=N_1 \times N_2$ pixels of fixed but unknown classification. Let the observation at image coordinates (i,j) be X_{ij} and the true but unknown classification at that image point be $\theta_{ij} \in \{\omega_1, \omega_2, \ldots, \omega_m\}$ where m is the number of cover classes represented in the scene, and ω_k is the k^{th} cover class. Associated with each X_{ij} and θ_{ij} is a class-conditional density $p(X_{ij}|\theta_{ij})$. The maximum likelihood point classifier estimates each θ_{ij} in the following way: Decide $\theta_{ij}=\omega_k$ if and only if $g_k(X_{ij}) \geq g_\ell(X_{ij})$ for all $\ell=1,2,\ldots,m$ where $g_k(X_{ij})$ is the discriminant function

$$g_k(X_{ij}) = p(X_{ij}|\omega_k)p(\omega_k) \qquad (1)$$

and $p(\omega_k)$ is the prior probability of class

This research was funded in part by National Aeronautics and Space Administration Contract No. NAS9-15466 and National Science Foundation Grant MCS78-04366.

Reprinted from MACHINE PROCESSING OF REMOTELY SENSED DATA, June 1980

ω_k occurring in the scene. Usually a good estimate for $p(\omega_k)$ is not known (or even sought), and the approximation $p(\omega_k) = 1/m$ is used (uniform priors).

Contextual information can be incorporated into a decision rule of the same general type by modifying the discriminant function. Let the context at image point X_{ij} consist of observations spatially near, but not necessarily adjacent to, X_{ij}. Group these observations along with X_{ij} into a vector of observations $\underline{X}_{ij} = (X_1, X_2, \ldots, X_p)^T$ with $X_p = X_{ij}$ and the number of observations taken as context being p-1 (the ordering is fixed but arbitrary). Call the arrangement of pixels in X_{ij} the p-context array. Let the possible classes associated with $\underline{X}_{ij}$ be $\underline{\theta}^p = (\theta_1, \theta_2, \ldots, \theta_p)^T$ where $\theta_i \in \{\omega_1, \omega_2, \ldots, \omega_m\}$ and the ordering of the elements in $\underline{\theta}^p$ coincides with that in $\underline{X}_{ij}$. Assuming that the observations are class-conditionally independent gives a discriminant function incorporating context as

$$g_k(\underline{X}_{ij}) = \sum_{\ell_1=1}^{m} \cdots \sum_{\ell_{p-1}=1}^{m} \left[\left(\prod_{n=1}^{p} p(X_n | \theta_n) \right) G(\underline{\theta}^p) \right] \quad (2)$$

where θ_p is fixed as ω_k [2]. The context distribution, $G(\underline{\theta}^p)$, is the relative frequency of occurrence in the scene of the class configuration in the p-context array given by $\underline{\theta}^p$. The similarity of this discriminant function to the function used by the maximum likelihood point classifier becomes clearer by rewriting $g_k(\underline{X}_{ij})$ as

$$g_k(\underline{X}_{ij}) = p(X_{ij} | \omega_k) \cdot \sum_{\ell_1=1}^{m} \cdots \sum_{\ell_{p-1}=1}^{m} \left[\left(\prod_{n=1}^{p-1} p(X_n | \theta_n) \right) G(\underline{\theta}^p) \right]$$

where θ_p is again fixed as ω_k. The summation term carries the contextual information and can be thought of as an expanded context-carrying version of $p(\omega_k)$ from the point classifier case. This discriminant function is identical to the no-context discriminant function when p=1 since $G(\underline{\theta}^1) \equiv p(\omega_k)$.

III. ESTIMATING CONTEXT DISTRIBUTION--$G(\underline{\theta}^p)$

To evaluate $g_k(\underline{X}_{ij})$ we must know values for the $p(X_n | \theta_n)$ and $G(\underline{\theta}^p)$. Methods for estimating $p(X_n | \theta_n)$ are well established from considerable experience in using the no-context maximum likelihood decision rule (as in Eq. 1) for classification (see [1]). Optimal methods for estimating $G(\underline{\theta}^p)$ are not yet established. Preliminary work on finding practical methods for estimating $G(\underline{\theta}^p)$ is presented in [2].

The most successful method developed to date for estimating $G(\underline{\theta}^p)$ goes as follows:

1. Perform a no-context uniform-priors classification on the training set, restricting the classifier's decision rule to choosing among spectral classes in the correct information class.

2. Estimate the context distribution, $G(\underline{\theta}^p)$, from the resulting 100 percent accurate classification of the training set by counting the number of occurrences* of all possible class configurations given by $\underline{\theta}^p$.

This method was used on a 50-pixel square area from the northeast corner of the Large Area Crop Inventory Experiment (LACIE) Segment No. 1860 in Hodgman County, Kansas. The class-conditional densities were estimated for the 16 spectral classes from randomly located training fields scattered throughout the entire 117-by-194 pixel Landsat data frame. The coordinates of the training set fields were chosen by selecting pixel coordinates from a random number table and surrounding the selected pixel by the largest homogeneous rectangle (up to field size 20 by 20). The classifications were tested for accuracy over five information classes (pasture, idle, wheat, corn and alfalfa) from "wall-to-wall" pixel-by-pixel ground truth.

The restricted no-context classification was performed over the first 25 lines of the 50-pixel-square area and the context distribution was estimated over those 25 lines. The classification results were evaluated over the last 25 lines. The results show (Table 1) that this method produced an estimate of the context distribution, $G(\underline{\theta}^p)$, which in turn produced con-

* The estimate of the context distribution, $G(\underline{\theta}^p)$, does not need to be normalized so as to be an actual probability estimate. The normalization factor does not affect the classification decisions based on the discriminant function in Eq. 2.

Table 1

CLASSIFICATION CLASS RESULTS ON LACIE DATA

Classification	Accuracy, %** Lines 25-50 Overall	Average-by-Class
Uniform-priors no-context --unrestricted	78.0	75.6
4 nearest neighbors*	85.5	81.6
8 nearest neighbors*	87.1	81.9

* $G(\theta^P)$ estimated from restricted uniform-priors no-context classification over lines 1-25.

** Classification performance can be tabulated in two ways. Overall accuracy is simply the overall number of correct classifications divided by the total number attempted. Average-by-class accuracy is obtained by first computing the accuracy for each class and taking the arithmetic average of the class accuracies. The latter is significant when the classification results exhibit a tendency to discriminate in favor of or against a subset of the classes.

textual classifications with significant improvement in classification accuracy over the conventional uniform-priors no-context classification on this data set.

While this method can produce good estimates of the context distribution, it suffers the limitation that a sufficient number of blocks of ground truth of sufficient size are needed to make an accurate estimate of the context distribution. This method cannot be used at all when blocks of ground truth data are not available, while the conditional probabilities can be estimated from ground truth at random pixel locations.

Another possible method of estimating the context distribution would be to base the estimate on a uniform-priors no-context classification. Such an estimate might then be refined by basing a new estimate on the context classification made using the first context distribution estimate. The estimates might even be iterated until the estimate producing the most accurate classification over the training set is found. (The final result should then be evaluated on a test set disjoint from the training set.)

Results from a straightforward implementation of this iterative "bootstrap" method were reported earlier in [2]. Estimates of the context distribution were made from counting the number of occurrences of all possible class configurations in the appropriate classification. While this method produced excellent results when simulated data were used, results using real Landsat data were disappointing.

It is thought that the no-context uniform-priors classifications of real Landsat data simply did not produce an accurate enough classification for the "bootstrap" method to work. The classifiction of the simulated data was accurate enough because the class-conditional probabilities $p(X|\theta_n)$ were modeled exactly, whereas the class-conditional probabilities were not modeled exactly on the real data classifications. This resulted in estimates of the context distribution, $G(\theta^P)$, in the real data cases that contained more spurious class configuration counts than in the simulated case, which in turn gave poorer context classification results in the real data case.

There are several ways in which the context distribution estimates from real data no-context classifications could be "cleaned up." One could employ a threshold procedure which deletes all class configurations with counts below a certain number. Another approach would be to divide each class configuration count by a fixed number and take the integer part of the result as the new count, deleting all class configurations with counts that become zero.

Yet another method for reducing the effect of spurious class configuration counts is to raise each count to a power and use the result as the context distribution estimate. For powers greater than one, the class configurations with larger counts are favored even more heavily versus those with relatively small counts in the discriminant function in Eq. 2. Conversely, for powers less than one, the class configurations with large counts are less heavily favored. Going to the extreme of a power of zero results in all class configurations being equally favored as in a uniform-priors no-context ~~configuration~~ classification.

This power method was first tried on a simulated data set to investigate the method's characteristics undisturbed by unknown effects from inaccurate modeling in the real data sets. This simulated data

set [2] was generated from a very accurate no-context classification of Landsat-1 data from an urban area (Grand Rapids, Michigan). A 50-pixel-square segment was used in the tests. See Figure 1 for a summary of the results. The results seem to indicate that when the model is exact, as the power used is increased (to a certain point), the classification results tend towards the results obtained when the context distribution is estimated from ground truth. Also, as expected, as the power used is decreased below one, the results tend toward a uniform-priors no-context classification.

The power method was also used on a 50-pixel-square segment of Landsat data containing approximately equal amounts of urban and agricultural area located to the southeast of Bloomington, Indiana. Statistics for the spectral classes were estimated using the 100-pixel-square area centered on the 50-pixel-square segment. A very careful uniform-priors no-context classification using 14 spectral classes was performed to delineate agricultural, urban and forested areas. As there were too few forested pixels to delineate forest test areas reliably, the classification was tested only for accuracy in classifying the agricultural and urban classes. Out of the 2500 pixels in the segment, a total of 867 pixels were manually interpreted as agriculture and 450 pixels as urban. The identification was made by interpretation of color infrared photography taken by aircraft on the same day as the Landsat pass.

As mentioned earlier, a straightforward implementation of the iterative bootstrap method of estimating the context distribution for this data set produced disappointing results. Whereas the no-context uniform-priors classification had an overall accuracy of 83.1 percent and average-by-class accuracy of 82.7 percent, the best the bootstrap method could do in three iterations was 85.3 percent overall accuracy and 84.8 percent average-by-class accuracy. The fourth iteration produced no improvement.

Figure 2 summarizes the results using the power method on two-nearest-neighbors context (neighbors to the north and east) based on an estimate of $G(\underline{\theta}^p)$ from the no-context uniform-priors classification. Trading off overall accuracy against average-by-class accuracy, the best classification was produced using a power of 5, for which an overall accuracy of 87.0 percent and average-by-class accuracy of 86.1 percent was achieved. This nearly doubled the accuracy improvement over the no-context classification produced by the straight bootstrap method. Note also that the results in Figure 2 follow the general trend of the simulated data results in Figure 1.

A second iteration of estimating the context distribution, $G(\underline{\theta}^p)$, was then made based on the classifications listed in Figure 2. The second estimate of $G(\underline{\theta}^p)$ based on the classification using the first estimate raised to a power of 10 produced the best classification results with an overall accuracy of 88.5 percent and an average-by-class accuracy of 87.5 percent (using $G(\underline{\theta}^p)$ raised to a power of 5). See Table 2 and Figure 3 for a summary of results. This second estimate of $G(\underline{\theta}^p)$ gave a total 5.4 percent improvement in overall accuracy and 4.8 percent improvement in average-by-class accuracy over the no-context classification. Even though these improvements are not as large as in the results using simulated data, or using the more restrictive method on real data, these results are certainly encouraging.

Table 2

SECOND ITERATION POWER METHOD RESULTS

Best four nearest-neighbor classifications with $G(\underline{\theta}^p)$ based on the classification in Figure 2.

Power Used in Fig. 2	Power Used in This Classification	Accuracy, % Overall	Accuracy, % Average-by-Class
2	5	86.5	85.6
3	5	86.3	85.7
5	5	87.3	86.7
7	5	88.1	87.2
10	5	88.5	87.5
15	3	87.7	87.2

Prior to making the second iteration estimate of $G(\underline{\theta}^p)$ above, it was assumed that the more accurate a classification was, the more accurate the estimate of $G(\underline{\theta}^p)$ from it would be. The results quoted here show clearly that this is not always the case. Further study is required before it can be determined whether this type of behavior is typical, and before this behavior can be exploited optimally.

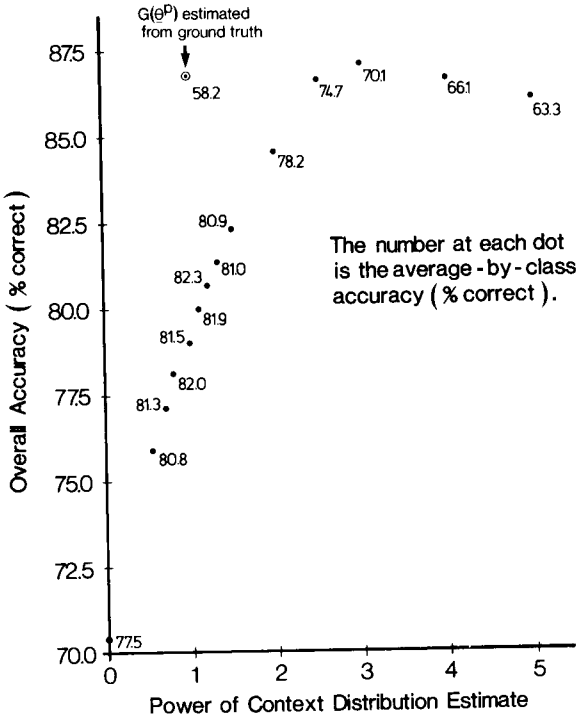

FIGURE 1. Power method results using as context one-nearest-neighbor (south) on the simulated data set. Context distribution, $G(\underline{\theta}^p)$, estimated from uniform-priors no-context, except where noted otherwise.

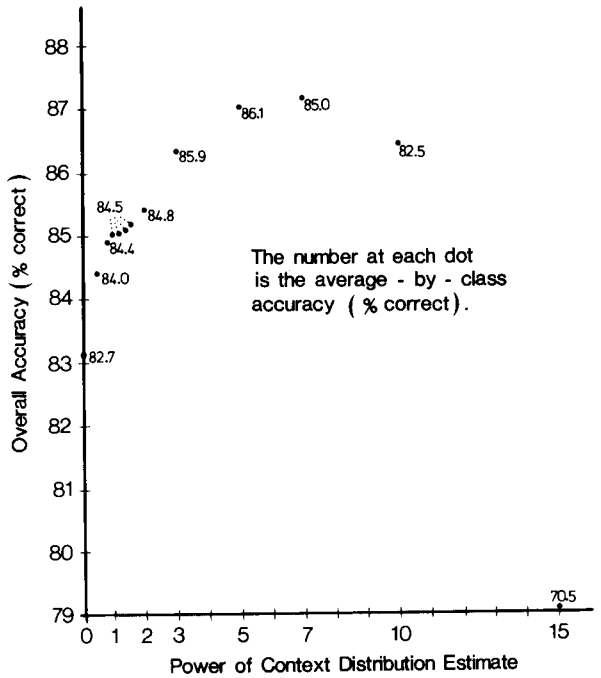

FIGURE 2. Power method results using two-nearest-neighbors (north and east) context on Bloomington, IN data set. Context distribution, $G(\underline{\theta}^p)$, estimated from uniform-priors no-context ~~distribution~~.

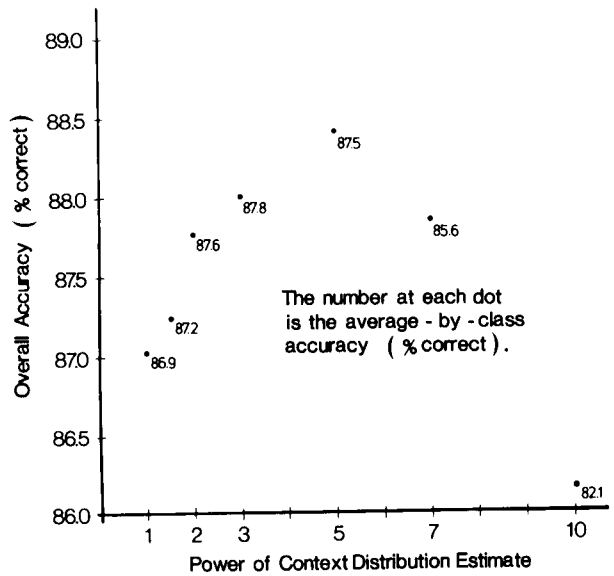

FIGURE 3. Power method results using four-nearest-neighbors context on Bloomington, IN data set. Context distribution, $G(\underline{\theta}^p)$, estimated from two-nearest neighbor (north and east) context classification with context distribution raised to power 10.

IV. PRACTICAL CONSIDERATIONS

The general approach to estimating the context distribution, as suggested by the results reported in the previous section, can involve a large number of context classifications before the best estimate is found. In addition to determining the best power of the context distribution to use at each iteration, the best p-context array (how many and which neighbor(s) to use) needs to be determined at each iteration.

The size and shape of the p-context array directly affect computation cost and classifiction accuracy. Generally, the larger the p-context array, the higher the computation cost. When the classification from which the context configuration is estimated is sufficiently accurate, larger p-context arrays yield higher classification accuracies. Less accurate template classifications can result in cases where a large p-context array will produce a classification that is less accurate than the no-context classification. Also, p-context arrays of given size may produce differing classification accuracies, depending on the shapes of the arrays. It would be desirable to be able to predict the optimal size and shape of the p-context array and the best power of the context distribution to use at each iteration before any actual classifications are performed.

V. ESTIMATION OF OPTIMAL P-CONTEXT ARRAY AND POWER

A theoretical measure of context has been developed from the perspective of applying this measure to predicting the optimal p-context array. This same measure may also be useful in estimating the best power to use of the context distribution.

Suppose that the relative frequency function $G(\underline{\theta}^p)$ is such that it can be written in factored form, i.e.,

$$G(\underline{\theta}^p) = G(\underline{\theta}_1^q)\ G(\underline{\theta}_2^{p-q}) \qquad (3)$$

1	2	3
4	5	6
7	8	9

Fig. 4. Pixel locations used in testing ΔG_q^p

where $\underline{\theta}_1^q$ and $\underline{\theta}_2^{p-q}$ are, respectively, q and p-q vectors of classes. The last element of $\underline{\theta}_2^{p-q}$ is the same as the last element of $\underline{\theta}^p$. If this factorization can indeed be realized, Eq. 2 can be rewritten as

$$g_k(\underline{X}_{ij}) = \left[\sum_{\ell_1=1}^{m} \cdots \sum_{\ell_q=1}^{m} \left(\prod_{n=1}^{q} p(X_n|\theta_n)\right) G(\underline{\theta}_1^q)\right] \cdot$$

$$\left[\sum_{\ell_{q+1}=1}^{m} \cdots \sum_{\ell_{p-1}=1}^{m} \left(\prod_{n=q+1}^{p} p(X_n|\theta_n)\right) G(\underline{\theta}_2^{p-q})\right] \qquad (4)$$

where $\ell_p=k$ and the last element of $\underline{\theta}_2^{p-q}$ is ω_k. Since the term in the first set of brackets is independent of k, it is just a constant term that can be ignored when classifying point (i,j). When such a factorization as in Eq. 3 can be made, we can reduce the size of the p-context array, reducing computation cost with no loss in classification accuracy.

If $G(\underline{\theta}^p)$ can be factored as in Eq. 3, it is clear that the distribution $G(\underline{\theta}^p)$ is one of independence for $\underline{\theta}_1^q$ and $\underline{\theta}_2^{p-q}$. This suggests that a measure of nonredundant contextual information from the pixel positions in $\underline{\theta}_1^q$ as compared to that from the pixel positions in $\underline{\theta}_2^{p-q}$ would be a measure of departure from independence for $\underline{\theta}_1^q$ and $\underline{\theta}_2^{p-q}$ in the distribution $G(\underline{\theta}^p)$. A possible measure of this departure would be

$$\Delta G_q^p = \sum_{\ell_1=1}^{m} \cdots \sum_{\ell_p=1}^{m} \left(G(\underline{\theta}_1^q)\ G(\underline{\theta}_2^{p-q}) - G(\underline{\theta}^p)\right)^2 \qquad (5)$$

where $G(\underline{\theta}_1^q)$ and $G(\underline{\theta}_2^{p-q})$ are now the marginals of $G(\underline{\theta}^p)$. Other distributions of independence with marginal $G(\underline{\theta}_2^{p-q})$ and other measures of departure from $G(\underline{\theta}^p)$ could be used. This particular form for ΔG_q^p is attractive because it is particularly easy to calculate.

The "context measure" ΔG_q^p can be used to estimate the optimal p-context array in the following way: Establish $\underline{\theta}_2^{p-q}$ as a fixed core (p-q)-context array. Calculate

the values of ΔG_q^p for various q-context arrays as $\underline{\theta}_1^q$, distinct from the core array. The best p-context array for $\underline{\theta}^p$ would be $\underline{\theta}_2^{p-q}$ combined with the $\underline{\theta}_1^q$ that produced the largest value for ΔG_q^p. This, of course, assumes that the contextual information contributed by $\underline{\theta}_1^q$ is not so erroneous that it would actually decrease classification accuracy. This may not be a reasonable assumption in all cases.

The first test of ΔG_q^p was made on the simulated data with p=2 and q=1 and the context distributions estimated from the ground truth. The context arrays $\underline{\theta}_1^1$ and $\underline{\theta}_2^1$ were defined with respect to the pixel locations defined in Figure 4. $\underline{\theta}_2^1$ was first fixed as pixel position 5 and $\underline{\theta}_1^1$ was varied over the remaining positions. $\underline{\theta}_2^1$ was also later fixed as pixel position 1 with $\underline{\theta}_1^1$ varied over the pixel positions relative to position 1 not covered previously (i.e., positions 3, 6, 7, 8 and 9).

Table 3

ΔG_q^p TESTED ON SIMULATED DATA WITH CONTEXT DISTRIBUTIONS ESTIMATED FROM GROUND TRUTH

$\underline{\theta}_1^1$ Pixel Location	$\underline{\theta}_2^1$ Pixel Location	$\Delta G_1^2 \times 10^4$	Accuracy, % Overall	Accuracy, % Average-by-class
8	5	5.09	92.7	74.0
2	5	4.99	91.6	73.5
4	5	4.90	91.7	71.8
6	5	4.90	91.7	73.9
7	5	3.42	90.8	71.2
3	5	3.31	90.4	69.8
9	5	3.26	90.6	70.6
1	5	3.19	90.6	70.1
7	1	2.58	90.3	68.6
3	1	2.27	90.2	70.3
8	1	1.98	89.4	67.9
6	1	1.87	90.4	70.2
9	1	1.53	89.9	69.5

Table 4

ΔG_q^p TESTED ON SIMULATED DATA WITH CONTEXT DISTRIBUTIONS ESTIMATED FROM UNIFORM-PRIORS NO-CONTEXT CLASSIFICATION

$\underline{\theta}_1^1$ Pixel Location	$\underline{\theta}_2^1$ Pixel Location	$\Delta G_1^2 \times 10^5$	Accuracy, % Overall	Accuracy, % Average-by-Class
8	5	7.56	79.8	81.7
2	5	7.30	79.1	81.9
4	5	6.13	78.8	80.6
6	5	6.11	79.0	81.4
7	5	4.71	78.8	80.9
3	5	4.53	78.6	80.6
9	5	4.28	78.4	80.6
1	5	4.22	78.3	79.7
7	1	3.77	78.5	80.9
8	1	2.73	78.0	80.0
3	1	2.65	78.0	80.9
6	1	2.31	78.0	80.8
9	1	2.17	78.0	80.1

As can be seen in Table 3, ΔG_q^p clearly predicted that the best neighbor to use for context would be any of the four nearest neighbors (pixel positions 2, 4, 6 or 8 relative to position 5). ΔG_q^p did not so clearly predict which nearest neighbor was best.

ΔG_q^p was again tested on the simulated data, but this time with the context distributions estimated from the uniform-priors no-context classification. As shown

in Table 4, in this case ΔG_q^p again tended to predict the best p-context array. This time ΔG_1^2 predicted pixel position 8 to be the best neithboring pixel to use as context while pixel position 2 came in as a close second. These predictions held up quite well when compared to the classification accuracies. These distinctions among the reamining pixels, however, weren't predicted as clearly.

A test of ΔG_q^p was also made using the Bloomington, Indiana Landsat data with the context distributions estimated from the uniform-priors no-context classification (see Table 5). Here ΔG_q^p did not predict the best p-context array as well as in the simulated data case. ΔG_q^p does correlate positively with the accuracy results, but the correlation is fairly weak. It seems that the context here is too erroneous for the predictor to function properly.

It was then checked to see if ΔG_q^p could be used to predict the power of the context distribution to use for a particular p-context array. $\underline{\theta}_2^1$ was set as position 5 and $\underline{\theta}_1^2$ was set as positions 2 and 6. The power used was varied as previously (see Figure 2). [NOTE: $G(\underline{\theta}^p)^\alpha$ was normalized for each value of α so as to remain a probability estimate.]

Table 5

ΔG_q^p TESTED ON BLOOMINGTON, IND. LANDSAT DATA SET. CONTEXT DISTRIBUTIONS ESTIMATED FROM UNIFORM-PRIORS NO-CONTEXT CLASSIFICATION

$\underline{\theta}_1^1$ Pixel Location	$\underline{\theta}_2^1$ Pixel Location	$\Delta G_1^2 \times 10^5$	Accuracy, % Overall	Accuracy, % Average-by-Class
4	5	7.69	84.2	83.8
6	5	7.68	84.6	84.1
2	5	5.40	85.2	84.8
8	5	5.31	83.8	83.4
3	5	3.79	84.2	83.8
7	5	3.61	84.0	83.5
1	5	3.04	84.4	84.1
9	5	2.96	83.7	83.2

Table 6

ΔG_q^p EVALUATED AS A PREDICTOR OF BEST TEST DISTRIBUTION POWER ON BLOOMINGTON, INDIANA, DATA TEXT

$\underline{\theta}_1^2$ = pixel locations 26

$\underline{\theta}_2^1$ = pixel location 5

Context distributions estimated from uniform-priors no-context distribution

Power	ΔG_2^3	Accuracy, % Overall	Accuracy, % Average-by-Class
.5	2.87×10^{-7}	84.4	84.0
.8	8.23×10^{-7}	84.9	84.4
1.0	2.05×10^{-6}	85.0	84.5
1.2	4.81×10^{-6}	85.0	84.5
1.4	9.27×10^{-6}	85.1	84.5
1.6	1.37×10^{-5}	85.2	84.5
2.0	1.34×10^{-5}	85.4	84.8
3.0	1.20×10^{-6}	86.3	85.9
5.0	4.04×10^{-9}	87.0	86.1
7.0	1.98×10^{-11}	87.2	85.0
10.0	underflow	86.4	82.5

In Table 6, ΔG_2^3 shows a distinct pattern of behavior as the power of the context distribution is varied. As the power is increased from one, ΔG_2^3 increases at first and then decreases. In this case, the power at which ΔG_2^3 falls to approximately its value in the power of one case corresponds closely to the power that yields the highest classification accuracies. As the power is increased further,

ΔG_1^2 decreases sharply. When the power is increased to the value that produces the classification that in turn produces the best context distribution estimate (in this case, a power of 10), ΔG_1^2 is so small that it can't be calculated in the precision used.

Further investigation with this and other data sets is needed to determine whether this is a universal pattern that can be exploited in estimating the power of the context distribution that yields the best classification results. These results make it seem unlikely, however, that ΔG_q^p could be used to predict the power which produces the best context distribution estimate.

CONCLUDING REMARKS

The multispectral maximum likelihood classifier has been extended to include contextual information from arbitrary points near, but not necessarily adjacent to, the point being classified. The successful application of this statistical context classifier depends, however, upon the successful estimation of the a priori context distribution, $G(\underline{\theta}^p)$. A method has been developed which can provide good estimates of the context distributions assuming that blocks of representative ground truth are available.

Attempts at developing a more general "bootstrap" method of estimating the context distribution have not yet been totally successful. Encouraging results have been obtained by using the power method described in this paper. Practical application of these bootstrap methods is clouded by the need to run several classifications to determine the best p-context array and the power of the context distribution to use at each iteration.

A theoretical basis for an estimator of the best p-context array has been developed. However, this estimator requires that the contextual information be reasonably accurate, an assumption that does not hold uniformly. Nevertheless, this same estimator may yet hold promise with respect to predicting the power of the context distribution which produces the most accurate classification results.

It is quite possible that no reliable estimation procedure simpler than actually performing a contextual classification can be found. If this is the case, the most effective way to "estimate" the best p-context array and context distribution power would be to perform contextual classifications on representative portions of the scene before the total scene is classified.

VII. REFERENCES

1. P. H. Swain and S. M. Davis, eds., Remote Sensing: The Quantitative Approach, McGraw-Hill, New York, 1978.

2. P. H. Swain, S. B. Vardeman, and J. C. Tilton, "Contextual Classification of Multispectral Image Data," Technical Report 011080, Laboratory for Applications of Remote Sensing (LARS), Purdue University, West Lafayette, Indiana 47907, Jan. 1980.

James C. Tilton is enrolled in the Ph.D. program in the School of Electrical Engineering at Purdue University and is a graduate research assistant at Purdue's Laboratory for Applications in Remote Sensing (LARS). B.A. cum laude, Rice University, 1976, in electrical engineering, environmental science and engineering, and anthropology; M.E.E., Rice University, 1976; M.S., optical sciences, University of Arizona, 1978. He came to Purdue in 1978 to pursue doctoral research in artificial intelligence and pattern recognition as applied to remote sensing. He is a member of Phi Beta Kappa and Tau Beta Pi honoraries.

Stephen Vardeman is an assistant professor of statistics at Purdue University. He received his B.S. and M.S. degrees in mathematics from Iowa State University, 1975. He is a member of the Institute of Mathematical Statistics and the American Statistical Association. His research interests include compound and empirical Bayes decision theory and applications of statistical techniques to pattern recognition problems.

Philip H. Swain is assistant professor of electrical engineering, Purdue University, and program leader for Data Processing and Analysis Research at the University's Laboratory for Applications of Remote Sensing (LARS); B.S.E.E., Lehigh University; M.S.E.E. and Ph.D., Purdue University. Prof. Swain has been affiliated with LARS since 1966 and has contributed extensively to the development of data processing methods for the management and analysis of remote sensing data. His areas of specialization include theoretical and applied pattern recognition and methods of artificial intelligence. He is co-editor and contributing author for the textbook Remote Sensing: The Quantitative Approach (McGraw-Hill, 1978).