Reprinted from

Seventh International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Range, Forest and Wetlands Assessment

June 23 - 26, 1981

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

CONTEXTUAL CLASSIFICATION OF MULTISPECTRAL IMAGE DATA: AN UNBIASED ESTIMATOR FOR THE CONTEXT DISTRIBUTION

JAMES C. TILTON AND PHILIP H. SWAIN
Purdue University

STEPHEN B. VARDEMAN
Iowa State University

ABSTRACT

Recent investigations have demonstrated the effectiveness of a contextual classifier that combines spatial and spectral information employing a general statistical approach.[1,2] This statistical classification algorithm exploits the tendency of certain ground-cover classes to occur more frequently in some spatial contexts than in others. Indeed, a key input to this algorithm is a statistical characterization of the context: the context distribution. Here we discuss an unbiased estimator of the context distribution which, besides having the advantage of statistical unbiasedness, has the additional advantage over other estimation techniques of being amenable to an adaptive implementation in which the context distribution estimate varies according to local contextual information. Results from applying the unbiased estimator to the contextual classification of three real Landsat data sets are presented and contrasted with results from non-contextual classifications and from contextual classifications utilizing other context distibution estimation techniques.

I. INTRODUCTION

The machine classification of multispectral image data collected by remote sensing devices aboard aircraft and spacecraft has usually been performed such that each pixel (picture element) is classified individually and independently.[3] The information used by this classifier is only spectral or, in some cases, spectral and temporal. There is no provision for using the spatial information inherent in the data. In contrast, when scanner data are displayed in image form, a human analyst routinely uses spatial information to establish a context for deciding what a particular pixel in the imagery might be. Using this context together with spectral information, the analyst may easily identify roads, delineate boundaries of agricultural fields, and differentiate between grass in an urban setting (e.g., lawns) and grass in an agricultural setting (e.g., pasture or forage crops) where a point-by-point classifier utilizing spectral information alone would have much difficulty in doing so.

The ECHO (Extraction and Classification of Homogeneous Objects) process is a variety of contextual classifier which has been found useful for classifying data sets which contain homogeneous objects that are large compared to the resolution of the imagery.[4] This classifier cannot be used effectively, however, if the data set does not contain a significant number of these large homogeneous objects.

In several recent papers,[1,2,5,6] we have described a general statistical classification method for exploiting both spatial and spectral information when classifying multispectral image data. This contextual classifier exploits the tendency alluded to earlier of certain ground-cover classes to occur more frequently in some contexts than in others. Unlike the ECHO process, this classifier can be used to advantage on any data set, even those data sets that do not have identifiable homogenous objects, such as is generally the case in forested, urban and other inhomogeneous areas.

We shall briefly review the statistical basis of the contextual decision rule and earlier methods for estimating a statistical characterization of context: the context distribution. We will then describe an unbiased estimator of the context distribution. Besides having the advantage of statistical unbiasedness, this estimator has the additional advantage over other estimation techniques of being amenable to an adaptive implementation in which the context distribution estimate varies according to local contextual information. Results from applying the unbiased estimator to the contextual classification of three real Landsat data sets are then presented and contrasted with results from non-contextual classifications and from contextual classifications utilizing other context distribution estimation techniques.

II. THEORETICAL BASIS OF THE CLASSIFIER

Consistent with the general characteristics of imaging systems for remote sensing, we assume a two-dimensional array of $N=N_1 \times N_2$ random observations

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

(pixels) X_{ij} having fixed but unknown classification ϑ_{ij}, as shown in Figure 1. The observation X_{ij} consists of n measurements (usually containing spectral and/or temporal information), while the classification ϑ_{ij} can be any one of m spectral or information classes* from the set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$.

$$\begin{matrix} \vartheta_{11} & \vartheta_{12} & \cdots & \vartheta_{1N_2} \\ \vartheta_{21} & \vartheta_{22} & \cdots & \vartheta_{2N_2} \\ \vdots & & & \\ \vartheta_{N_1 1} & & \cdots & \vartheta_{N_1 N_2} \end{matrix}$$

Figure 1. A two-dimensional array of $N=N_1 \times N_2$ pixels.

Let $\underline{X}$ denote a vector whose components are the random observations:

$$\underline{X} = [X_{ij} \mid i=1,2,\ldots,N_1; j=1,2,\ldots,N_2]^T.$$

Similarly, let $\underline{\vartheta}$ be the vector of states (true classifications) associated with the observations:

$$\underline{\vartheta} = [\vartheta_{ij} \mid i=1,2,\ldots,N_1; j=1,2,\ldots,N_2]^T.$$

The following notation will be useful. Let $\underline{\vartheta}^p \in \Omega^p$ and $\underline{X}^p \in (R^n)^p$ stand for p-vectors of classes and n-dimensional measurements, respectively; each component of $\underline{\vartheta}^p$ is a variable which can take on any classification value; each component of $\underline{X}^p$ is a n-dimensional random vector which can take on values in the observation space.

Let the action (classification) taken with respect to pixel (i,j) be denoted by $a_{ij} \in \Omega$. We restrict the action a_{ij} to be a function of a specified subset of observations in $\underline{X}$. This subset includes, along with X_{ij}, p-1 observations spatially near to, but not necessarily adjacent to, X_{ij}. These p-1 observations serve as the spatial context for X_{ij} and are taken from the same spatial positions relative to pixel position (i,j) for all i and j. Call this arrangement of pixels together with X_{ij} the p-context array, several examples of which are shown in Figure 2. Group the p observations in the p-context array into a vector of observations $\underline{X}_{ij} = (X_1, X_2, \ldots, X_p)^T$ and let $\underline{\vartheta}_{ij}$ be the vector of true but unknown classifications associated with the observations in $\underline{X}_{ij}$. Note that the $\underline{\vartheta}_{ij}$ and $\underline{X}_{ij}$ are the particular instance of $\underline{\vartheta}^p$ and $\underline{X}^p$ associated with pixel position (i,j). Correspondence of the components of $\underline{\vartheta}_{ij}$, $\underline{X}_{ij}$, $\underline{\vartheta}^p$ and $\underline{X}^p$ to the positions in the p-context array is fixed but arbitrary except that the p^{th} components will always correspond to the pixel to be classified.

Let the loss suffered by taking action a_{ij} be denoted by $\lambda(\vartheta_{ij}, a_{ij})$ for some fixed non-negative function $\lambda(\cdot,\cdot)$. The expected average loss (or risk) suffered

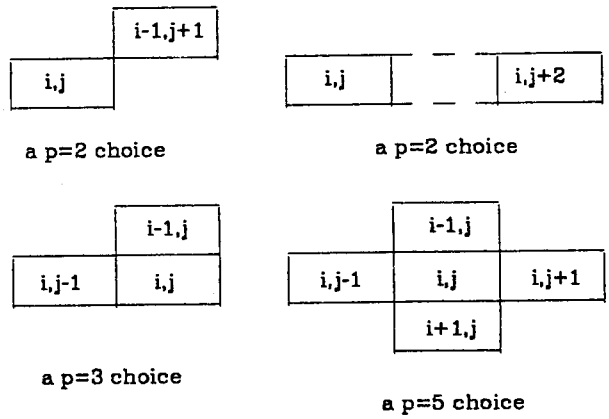

Figure 2. Examples of p-context arrays.

over the N classifications in the classification array is

$$R_{\underline{\vartheta}} = E\left[\frac{1}{N}\sum_{i,j}\lambda(\vartheta_{ij}, a_{ij}(\underline{X}_{ij}))\right] \qquad (1)$$

where the expectation is with respect to the distribution of $\underline{X}$.

Now consider finding a decision rule of the form

$$a_{ij}(\underline{X}_{ij}) = d(\underline{X}_{ij}) \qquad (2)$$

for a fixed function $d(\cdot)$ mapping p-vectors of observations to actions so that $R_{\underline{\vartheta}}$ is minimized. If we require that the distributions of the $\underline{X}_{ij}$ are spatially invariant, i.e. the value of the probability density for $\underline{X}_{ij}$ depends only on the measurement values in $\underline{X}_{ij}$ and the set of classifications in $\underline{\vartheta}_{ij}$ and not the location (i,j), the risk, $R_{\underline{\vartheta}}$, can be written as

$$\begin{aligned} R_{\underline{\vartheta}} &= \sum_{\underline{\vartheta}^p \in \Omega^p} C(\underline{\vartheta}^p) \int \lambda(\vartheta_p, d(\underline{X}^p)) f(\underline{X}^p \mid \underline{\vartheta}^p) d\underline{X}^p \\ &= \int \sum_{\underline{\vartheta}^p \in \Omega^p} C(\underline{\vartheta}^p) \lambda(\vartheta_p, d(\underline{X}^p)) f(\underline{X}^p \mid \underline{\vartheta}^p) d\underline{X}^p \end{aligned} \qquad (3)$$

where $C(\underline{\vartheta}^p)$, the *context distribution*, is the relative frequency with which $\underline{\vartheta}^p$ occurs in the array $\underline{\vartheta}$, and ϑ_p is the p^{th} element of $\underline{\vartheta}^p$. For any array $\underline{\vartheta}$, a decision rule $d(\underline{X}^p)$ minimizing $R_{\underline{\vartheta}}$ can be obtained by minimizing the integrand in (3) for each $\underline{X}^p$; thus for a specific $\underline{X}_{ij}$ (an instance of $\underline{X}^p$), an optimal action is:

$d(\underline{X}_{ij})$ = the action (classification) a which minimizes

$$\sum_{\underline{\vartheta}^p \in \Omega^p} C(\underline{\vartheta}^p) \lambda(\vartheta_p, a) f(\underline{X}_{ij} \mid \underline{\vartheta}^p). \qquad (4)$$

In practice, a "0-1 loss function" is usually assumed, i.e.,

$$\lambda(\vartheta, a) = \begin{cases} 0, & \text{if } \vartheta = a \\ 1, & \text{if } \vartheta \neq a \end{cases}.$$

* Spectral classes are spectrally differentiable subclasses of information classes (the classes of interest).

Then (4) simplifies and the decision rule becomes:

$$d(\underline{X}_{ij}) = \text{the action } a \text{ which } \textit{maximizes} \sum_{\substack{\underline{v}^p \in \Omega^p, \\ v_p = a}} C(\underline{v}^p) f(\underline{X}_{ij} \mid \underline{v}^p). \quad (5)$$

We now assume class-conditional independence for the observations. This assumption means that the joint class-conditional density over the p-context array can be written as

$$f(\underline{X}_{ij} \mid \underline{v}^p) = \prod_{k=1}^{p} f(X_k \mid v_k) \quad (6)$$

where X_k and v_k are the k^{th} elements of $\underline{X}_{ij}$ and $\underline{v}^p$, respectively. Evidence that this is a reasonable assumption may be found in Yamamoto.[7] With this assumption, the decision rule in (5) becomes:

$$d(\underline{X}_{ij}) = \text{the action } a \text{ which maximizes} \sum_{\substack{\underline{v}^p \in \Omega^p, \\ v_p = a}} C(\underline{v}^p) \prod_{k=1}^{p} f(X_k \mid v_k). \quad (7)$$

A more detailed derivation of this decision rule can be found in Swain, et al.[1]

The optimal choice of $d(\cdot)$ cannot be implemented in practice since it depends on $C(\underline{v}^p)$ and the $f(X_k \mid v_k)$ which are unknown. Methods for estimating the $f(X_k \mid v_k)$ are well established from considerable experience in using the conventional non-contextual maximum likelihood decision rule.[3] When the classification set Ω consists of spectral classes, the $f(X_k \mid v_k)$ are assumed to be multivariate normal densities. In the case where the classification set Ω consists of information classes, the $f(X_k \mid v_k)$ are assumed to be weighted sums of multivariate normal densities. We will next discuss methods for estimating the context distribution, $C(\underline{v}^p)$.

III. CONTEXT DISTRIBUTION ESTIMATION: EARLIER TECHNIQUES

Simulated data sets were utilized in the earliest experiments exploring the effectiveness of classifying multispectral remote sensing data using context classification as defined by the set of discriminant functions in (7). This was done to demonstrate the effectiveness of the classifier given that the underlying assumptions in the classification model are satisfied. At first, the context distribution was found by simple tabulation from the true classification used as a template for the data simulation. As reported in Swain, et al.[1] the classifier was very effective when the context distribution was determined in this way.

When dealing with real data, there is no direct way of determining the context distribution. We cannot tabulate the context distribution from the true classification since the true classification is not known. However, we do expect that, at least for large $N = N_1 \times N_2$, the decision rule in (5) where $C(\underline{v}^p)$ is replaced by an estimate $\hat{C}(\underline{v}^p)$ based on the data, $\underline{X}$, will have risk $\hat{R}_g$ approximating that of the optimal rule. Thus we should be able to base an adequate estimate of the context distribution on the data or, more practically, on representative sections from the data designated as a training set. The most straightforward way to develop an estimate of the context distribution from the training set would be to perform a conventional non-contextual classification of the training set and use the context distribution as tabulated from this classification as an estimate of the context distribution. One could then further refine this estimate of the context distribution by making another estimate from the contextual classification, and even iterate in this way until no further improvement in classification accuracy was obtained.

This iterative "classify-and-count" method was tested on one simulated data set and two real data sets. As reported in Swain, et al,[1] this method gave excellent results on the simulated data set, but disappointing results on the real data sets, stimulating a search for alternative methods for estimating the context distribution. One such method is the ground-truth-guided method. In this method, roughly equal subsets of the ground truth data are designated as a training set for estimating the context distribution and a test set for evaluating the classification results. The ground truth data are, of course, represented in terms of information classes. When the estimation is to be done in terms of spectral classes rather than information classes, the following method is used:

(1) Perform a conventional non-contextual classification of the training set using uniform prior probabilities, but allow the the classifier to choose only among spectral classes associated with the information class designated by the ground truth.
(2) Estimate the context distribution by tabulation from the resulting 100-percent accurate classification of the training set.
(3) Classify the entire scene with the contextual classifier and evaluate the results over a test set disjoint from the training set.

When the estimation is to be done in terms of information classes, the restricted spectral class classification in step (1) above must still be performed. In this case, however, this classification is used to provide (by tabulation) an estimate of the weights used in the weighted sum of class-conditional normal densities that make up the set of densities $f(X_k \mid v_k)$ in (7). Each weight is the relative frequency of occurrence in the training set of a particular spectral class for a given information class. The entire scene is then classified in terms of information classes using the contextual classifier, and evaluated over a test set disjoint from the training set, as in the spectral class case.

Both the spectral and information class formulations of the ground-truth-guided method were tested on two 50-pixel-square Landsat data sets. One data set was a LACIE data set from Hodgeman County, Kansas, containing pasture, wheat, corn and fallow fields. The other data set was from Tippecanoe County, Indiana, containing residential and commercial areas in northern Lafayette and West Lafayette, Indiana, as well as

areas of forest, agriculture and water (the Wabash River). For both data sets, the restricted spectral class classification was performed over the first 25 lines of the data set and the context distribution was estimated over those 25 lines. Contextual classifications of the scenes were performed and classification accuracies* were evaluated over the last 25 lines as well as over the entire data set.

Tables 1 and 2 present the results from contextual classifications using four-nearest-neighbor (4nn) estimates of the context distribution (the p=5 choice in Figure 2) for both the spectral and information class formulations of the ground-truth-guided method (gtgm). These results are also compared to the accuracies obtained from uniform-priors and estimated-priors non-contextual maximum likelihood classifications. The prior probabilities for the estimated-priors non-contextual classifications were estimated by tabulation from the uniform-priors non-contextual classification. These results show that contextual classifications using the ground-truth-guided method for estimating the context distribution give significantly better results than non-contextual classifications on these data sets. For these cases, the spectral class formulation of the ground-truth-guided method generally produces higher classification accuracies. However, since the spectral class estimate of the context distribution has substantially more non-zero elements than the information class estimate, contextual classifications using the spectral class formulation generally take over twice the computer time required for the information class formulation.

While this method can produce good estimates of the context distribution, it suffers the limitation that it requires large areas of spatially contiguous ground truth data. When such detailed ground truth data are not available, some other method is needed.

* Classification accuracy 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 then 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.

Table 1. Comparison of the contextual classifier using the ground-truth-guided method with non-contextual classifiers; Hodgeman County, Kansas, Landsat Data Set.

Classification	% Accuracy lines 26-50		lines 1-50	
	Overall	Average-by-Class	Overall	Average-by-Class
uniform priors	81.5	78.2	82.5	74.3
estimated priors	82.2	78.3	82.8	74.1
4nn gtgm, spectral	85.4	81.6	85.7	77.3
4nn gtgm, information	85.3	81.4	85.0	76.0

Table 2. Comparison of the contextual classifier using the ground-truth-guided method with non-contextual classifiers; Tippecanoe County, Indiana, Landsat Data Set.

Classification	% Accuracy lines 26-50		lines 1-50	
	Overall	Average-by-Class	Overall	Average-by-Class
uniform priors	82.7	81.7	81.8	83.4
estimated priors	84.2	82.0	83.7	83.7
4nn gtgm, spectral	88.7	91.1	89.3	90.7
4nn gtgm, information	88.2	87.3	88.2	86.2

The "Power Method" was the next method investigated as a generally applicable method of estimating the context distribution. To employ the method, one raises the relative frequency count for each class configuration to a power and uses the result as the context distribution estimate. This method is described in detail in Tilton, et al.[6] The context distribution estimates generated by the Power Method can produce classification accuracies of roughly the same high level as produced by the ground-truth-guided method. However, the method is very inconvenient to use.

With the the Power Method, an estimate of the context distribution is tabulated from a uniform-priors non-contextual clasification of the training set. Then contextual classifications of the training set and test set are performed using a power of the tabulated context distribution. To achieve the best possible results, a second iteration of this procedure must generally be performed, using a context distribution estimate tabulated from the training set of the first iteration classification. Unfortunately, no reliable predictor has been found for the optimal power to be used for the first or second iteration. It is not even the case that the most accurate first iteration classification will provide in general the best template for the second iteration. Further, on certain data sets, a spectral-class context-distribution estimate produces the best results, while on other data sets an information-class formulation works better. Despite the good results possible with the Power Method, these ambiguities make this method difficult to use, and not useful for practical applications. A search for a better generally applicable method for estimating the context distribution has led to the unbiased estimation technique described in the next section.

IV. CONTEXT DISTRIBUTION ESTIMATION: UNBIASED ESTIMATOR

One tactic for seeking an optimal estimate of the context distribution, $C(\underline{v}^p)$, is to look for an estimator function, $T_{\underline{X}^p}(\underline{X})$, which minimizes the mean-squared error given by

$$MSE = E\left[T_{\underline{X}^p}(\underline{X}) - C(\underline{v}^p)\right]^2. \tag{8}$$

Equation (8) can be rewritten as

$$MSE = Var[T_{\underline{X}^p}(\underline{X})] + b^2 \tag{9}$$

where $Var[T_{\underline{X}^p}(\underline{X})]$ is the variance of the estimate $T_{\underline{X}^p}(\underline{X})$ and b is the bias given by

$$b = E[T_{\underline{X}^p}(\underline{X})] - C(\underline{v}^p). \tag{10}$$

Finding the minimum mean-squared-error estimate is generally a difficult task, but since bias represents a systematic error, a reasonable approach would be to control bias before considering the variance. The best one can do in controlling bias is to seek an unbiased estimator, i. e., one for which $b = 0$.

As we saw in the previous section, the classify-and-count method performed poorly in tests on real Landsat data sets. One reason for this is that the estimate can be statistically biased. To prove this, consider the classification model as presented in section II. In addition to the symbol definitions given there, we make the following definitions. Let $\underline{\hat{v}}$ be the vector of classifications

$$\underline{\hat{v}} = [\hat{v}_{ij} \mid i=1,2,\ldots,N_1; j=1,2,\ldots,N_2]^T$$

where $\hat{v}_{ij}$ is the classification estimate from a non-contextual classification of the observation X_{ij}. Let $\underline{\hat{v}}_{ij}$ be a p-vector of classification estimates associated with the observations in the p-context array, $\underline{X}_{ij}$. Similarly, let $\underline{\hat{v}}^p$ be such an estimate associated with an arbitrary p-context array, $\underline{X}^p$. Let $\underline{\eta}^p \in \Omega^p$ represent an arbitrary p-vector of classes. The classify-and-count method can be described by the following estimator function for $C(\underline{v}^p)$:

$$T_{\underline{X}^p}(\underline{X}) \triangleq \hat{C}(\underline{v}^p) = \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} I(\underline{X}_{ij}, \underline{v}^p) \tag{11}$$

where

$$I(\underline{X}_{ij}, \underline{v}^p) = \begin{cases} 1, & \text{if } \underline{v}^p = \underline{\hat{v}}_{ij} \\ 0, & \text{otherwise.} \end{cases}$$

The expected value of $T_{\underline{X}^p}(\underline{X})$ is then

$$\begin{aligned} E[T_{\underline{X}^p}(\underline{X})] &\triangleq E\left[\frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} I(\underline{X}_{ij}, \underline{v}^p)\right] \\ &= \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} E\left[I(\underline{X}_{ij}, \underline{v}^p)\right] \\ &= \frac{1}{N}\sum_{\underline{\eta}^p \in \Omega^p}\ \sum_{\substack{i,j \text{ with} \\ \underline{\hat{v}}_{ij} = \underline{\eta}^p}} E\left[I(\underline{X}_{ij}, \underline{v}^p)\right] \\ &= \sum_{\underline{\eta}^p \in \Omega^p} C(\underline{\eta}^p) \int_{\substack{\underline{X}^p \in (R^n)^p \\ \text{with } \underline{\hat{v}}^p = \underline{\eta}^p}} f(\underline{X}^p \mid \underline{v}^p)\, d\underline{X}^p. \end{aligned} \tag{12}$$

Equations (10) and (12) show that the bias of the classify-and-count method is the difference between $C(\underline{v}^p)$ and a weighted sum of $C(\underline{\eta}^p)$. Note that this bias is independent of N, and cannot be reduced by increasing sample size. The bias can be non-zero or zero, depending of the values of $C(\underline{\eta}^p)$ and integrals in (12). To show this explicitly, let's consider the simple special case of a two-class problem (m=2) estimating non-contextual relative frequences of classes (p=1) for univariate random observations (n=1). Let the non-contextual classifier used to produce $\underline{\hat{v}}$ be the uniform-priors maximum-likelihood classifier with the decision rule:

$$d(X_{ij}) = \text{the action } a \text{ which maximizes } f(X_{ij} \mid a)$$

for all $a \in \{\omega_1, \omega_2\}$. The densities, $f(X_{ij} \mid a)$, are assumed to be normal with mean and variance $\mu_1 = -1$ and $\sigma_1^2 = 1$ for class ω_1 and mean and variance $\mu_2 = 1$ and $\sigma_2^2 = 1$ for class ω_2. For class ω_1 we have:

$$\begin{aligned} E[T_{\omega_1}(X)] &= \sum_{k=1}^{2} C(\omega_k) \int_{\substack{f(X \mid \omega_1) \\ \geq f(X \mid \omega_2)}} f(X \mid \omega_k)\, dX \\ &= \sum_{k=1}^{2} C(\omega_k) \int_{-\infty}^{0} f(X \mid \omega_k)\, dX \end{aligned}$$

$$= G(\omega_1)[\frac{1}{2} + \text{erf}\frac{0+1}{1}] + G(\omega_2)[\frac{1}{2} + \text{erf}\frac{0-1}{1}]$$
$$= .84G(\omega_1) + .16G(\omega_2). \quad (13)$$

The sum in (13) is equal to $G(\omega_1)$ only if $G(\omega_1) = G(\omega_2) = \frac{1}{2}$. For any other values of $G(\omega_1)$ and $G(\omega_2)$ the estimate is biased. Similar comments apply for class ω_2 where we have

$$E[T_{\omega_2}(X)] = .16G(\omega_1) + .84G(\omega_2). \quad (14)$$

We have shown, then, that the classify-and-count method does indeed generally produce biased estimates of the context distribution.

The unbiased estimator we have adopted can be most easily described by first considering the p=1 case and then generalizing to the arbitrary p-context array. For p=1, we examine the equation

$$\int h_k(X)\left[\sum_{l=1}^{m} f(X\,|\,\omega_l)G(\omega_l)\right]dX = \sum_{l=1}^{m}\left[\int h_k(X)f(X\,|\,\omega_l)dX\right]G(\omega_l) \quad (15)$$

where m is the number of classes; $f(X\,|\,\omega_l)$, $l=1,2,\ldots,m$, are the class-conditional densities described earlier; and the functions $h_k(X)$, $k=1,2,\ldots,m$, can be any set of m linearly independent functions. Equation (15) is valid provided all indicated sums and integrals are well defined, which will, for example, be the case when all of the functions in (15) are bounded. The functions $G(\omega_l)$ and $f(X\,|\,\omega_l)$ are always bounded because $G(\omega_l)$ is a relative frequency function and $f(X\,|\,\omega_l)$ is a multivariate normal density function. The functions $h_k(X)$ considered in the following development will also always be bounded.

The left-hand side of (15), which looks like the expected value of $h_k(X)$, can be estimated from the data $\underline{X}$ as follows:

$$\int h_k(X)\left[\sum_{l=1}^{m} f(X\,|\,\omega_l)G(\omega_l)\right]dX \cong \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} h_k(X_{ij}) \triangleq \overline{h}_k(\underline{X}) \quad (16)$$

where N, N_1 and N_2 are as defined in Figure 1, and $k \in \{1,2,\ldots,m\}$. Applying (15) and (16) m times, once for each class, we can write

$$\begin{bmatrix} \overline{h}_1(\underline{X}) \\ \overline{h}_2(\underline{X}) \\ \vdots \\ \overline{h}_m(\underline{X}) \end{bmatrix} = \begin{bmatrix} I_{11} & I_{12} & \cdots & I_{1m} \\ I_{21} & I_{22} & \cdots & I_{2m} \\ \vdots & \vdots & & \vdots \\ I_{m1} & I_{m2} & \cdots & I_{mm} \end{bmatrix} \begin{bmatrix} G(\omega_1) \\ G(\omega_2) \\ \vdots \\ G(\omega_m) \end{bmatrix} \quad (17a)$$

where

$$I_{kl} \triangleq \int h_k(X)f(X\,|\,\omega_l)\,dX\,. \quad (17b)$$

This can be more succinctly represented in vector-matrix notation as

$$\underline{h} \cong I\,\underline{G}\,. \quad (18)$$

Now $\underline{G}$ can be estimated by solving

$$\underline{G} \cong I^{-1}\,\underline{h} \triangleq \underline{T} \quad (19)$$

where $\underline{T} = (T_1(\underline{X}), T_2(\underline{X}), \ldots, T_m(\underline{X}))^T$ is the vector equivalent of $T(\underline{X})$ in (8), (9) and (10).

To show that $\underline{T}$ is indeed an unbiased estimator for $\underline{G}$, we note that

$$E(\underline{T}) = E(I^{-1}\,\underline{h}) = I^{-1}\,E(\underline{h})\,. \quad (20)$$

Looking at $E(\underline{h})$ element by element we have

$$E\left[\overline{h}_k(\underline{X})\right] \triangleq E\left[\frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} h_k(X_{ij})\right] \quad (21a)$$
$$= \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} E\left[h_k(X_{ij})\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\int h_k(X_{ij})\,f(X_{ij}\,|\,\omega_{ij})\,dX_{ij}$$
$$= \frac{1}{N}\sum_{l=1}^{m}\sum_{\substack{i,j \\ \text{with} \\ \omega_{ij}=\omega_l}}\int h_k(X_{ij})\,f(X_{ij}\,|\,\omega_{ij})\,dX_{ij}$$
$$= \sum_{l=1}^{m} G(\omega_l)\int h_k(X)\,f(X\,|\,\omega_l)\,dX \quad (21b)$$

Thus

$$E(\underline{h}) = I\,\underline{G}$$

and (20) becomes

$$E(\underline{T}) = I^{-1}\,E(\underline{h}) = I^{-1}\,I\,\underline{G} = \underline{G} \quad (22)$$

proving that $\underline{T}$ is an unbiased estimator for $\underline{G}$.

It is convenient to use a function of the class-conditional densities for the functions $h_k(X)$. More specifically, let $h_k(X) = (2\pi)^{\frac{n}{2}} f(X\,|\,\omega_k)$ and write (17b) as

$$I_{kl} = (2\pi)^{\frac{n}{2}}\int f(X\,|\,\omega_k)f(X\,|\,\omega_l)\,dX$$

where n is the dimensionality of X. Assuming the ω_k are normally distributed spectral classes with respective mean vectors μ_k and covariance matrices Σ_k $(k=1,2,\ldots,m)$, we find

$$I_{kl} = \left[\det(\Sigma_k + \Sigma_l)\right]^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mu_k-\mu_l)^T\,(\Sigma_k+\Sigma_l)^{-1}\,(\mu_k-\mu_l)\}\,. \quad (23)$$

When the ω_k are information classes, the I_{kl} are weighted sums of terms of the form given in (23).

When the estimate is made in terms of information classes, estimates must be made of the weights used to form the weighted sum of the class-conditional normal densities of the spectral subclasses. For each information class, the weights are estimated by using the unbiased estimator with p=1 for the spectral classes which make up the information class being considered.

The calculation of the estimate of $\underline{C}$ can proceed in one of two alternative ways. The vector $\overline{h}$ can be calculated for the entire image (as in (17a)), then multiplied by I^{-1} to give $\underline{T} \cong \underline{C}$; or as the $h_k(X_{ij})$ are calculated at each data point (pixel), the product with I^{-1} can be performed. The average of this product over the entire image is then $\underline{T} \cong \underline{C}$. The methods are completely equivalent; the difference between them amounts to a change in order of summation. However, the second method must be used when this unbiased estimator is extended to the arbitrary p-context array case, because the use of the first method for large values of p would require an impractical amount of storage. In calculating the estimate of $C(\underline{v}^p)$ at each image data point using the second method, individual unbiased estimates of the prior probabilities of each class are made for each position in the p-context array, and cross-products of these prior probabilities are taken to form the unbiased estimate of $C(\underline{v}^p)$ based on that image point. To save computer storage space, the cross-products having values below a specified threshold are ignored. The estimate of $C(\underline{v}^p)$ for the entire image is the average of the estimates of $C(\underline{v}^p)$ based on all the individual image points in the scene.

The unbiased estimator can be modified to provide an adaptive estimate of the context distribution. The local context distribution estimate for a particular $n_1 \times n_2$ block of image data is made from a $m_1 \times m_2$ block ($m_1 \geq n_1$ and $m_2 \geq n_2$). The $n_1 \times n_2$ block of image data is then classified using this local estimate of the context distribution. This process is repeated until the entire data set is classified. Better results have generally been obtained when $m_1 > n_1$ and $m_2 > n_2$. If $m_1 = n_2$ and $m_2 = n_2$, the context distribution estimate is not accurate for the pixels at the edges of the image data block being classified. Tests on three 50-pixel-square Landsat data sets have indicated good choices for n_1 and n_2 ranging from 10 up to 25 with the corresponding choices for m_1 and m_2 being 8 to 10 larger than the values chosen for n_1 and n_2.

V. CONTEXTUAL CLASSIFICATION RESULTS EMPLOYING THE UNBIASED ESTIMATOR

Table 3 presents the accuracies resulting from contextual classifications for three Landsat data sets using four-nearest-neighbor (4nn) estimates of the context distribution. The results using the spectral-class formulation are shown for the whole scene (non-adaptive) version and for an adaptive version employing local context distribution estimates for 25×25 pixel blocks made from the same 25×25 pixel block. The results using the information-class formulation are shown for an adaptive version employing estimates for various $n_1 \times n_2$ pixel blocks made from a $m_1 \times m_2$ pixel block centered on each $n_1 \times n_2$ pixel block. The uniform-priors non-contextual classification results are given for reference.

Figure 3 shows computer generated gray-scale maps of classifications of the Tippecanoe County, Indiana, Landsat data set. The contextual classification looks visually closer to the reference image than might be expected based on the accuracy improvement over the non-contextual classifications. This is due to the tendency of the contextual information to provide a smoothing effect, making classification maps that are not only more accurate, but also more pleasing to the eye.

The adaptive information-class formulation performs as well as or better than any other formulation shown. As noted earlier in the discussion of the ground-truth-guided method, the information-class formulation has the further advantage of having substantially fewer non-zero elements in the context distribution estimate, causing contextual classifications using an information-class formulation to require less than half the computer time required for contextual classifications using a corresponding spectral class formulation.

VI. CONCLUDING REMARKS

It had been shown earlier in this research[1,2,6] that the contextual classifier can provide improved classification performance, as compared to non-contextual classification, when accurate characterizations of the context distribution are available. The ground-truth-guided method has been shown to provide sufficiently accurate estimates of the context distribution, but suffers the disadvantage of requiring sizeable amounts of spatially contiguous ground truth. The unbiased estimator described herein overcomes this disadvantage, providing good estimates of the context distribution while requiring no more ground truth data than is required for a non-contextual classification. Furthermore, the unbiased estimator is amenable to an adaptive implementation so that the resulting context distribution estimate is more closely tailored to local conditions in the image data.

REFERENCES

1. P. H. Swain, S. B. Vardeman and J. C. Tilton (1980). *Contextual Classification of Multispectral Image Data.* Laboratory for Applications of Remote Sensing (LARS), Purdue University, West Lafayette, Indiana. LARS Technical Report 011080 (NASA Contract NAS9-15466). To appear in *Pattern Recognition.*

2. P. E. Anuta, D. A. Landgrebe, H. J. Siegel and P. H. Swain (1980). *Vol. III: Data Processing Research and Techniques Development.* Laboratory for Applications of Remote Sensing (LARS), Purdue University, West Lafayette, Indiana. LARS Contract Report 112880 (NASA Contract NAS9-15466).

3. P. H. Swain and S. M. Davis, eds. (1978). *Remote Sensing: The Quantitative Approach.* McGraw-Hill International Book Co., New York.

4. D. A. Landgrebe (1980). "The Development of a Spectral-Spatial Classifier for Earth Observation Data." *Pattern Recognition,* Vol. 12, No. 3, pp. 165-175, May-June 1980.

5. J. C. Tilton (1980). *Contextual Classification of Multispectral Image Data: Approximate Algorithm.* NASA, Johnson Space Center, Houston, Texas SR-PO-00491.

6. J. C. Tilton, P. H. Swain and S. B. Vardeman (1980). "Context Distribution Estimation for Contextual Classification of Multispectral Image Data." *Proceedings of the 1980 Machine Processing of Remotely Sensed Data Symposium* (IEEE Catalog No. 80 CH 1533-9 MPRSD), pp. 171-180, June 1980.

7. H. Yamamoto (1979). "A Method of Deriving Compatibility Coefficients for Relaxation Operators." *Computer Graphics and Image Processing,* Vol. 10, pp. 256-271.

8. J. Van Ryzin (1966). "The Compound Decision Problem With m × n Finite Loss Matrix." *Annals of Mathematical Statistics,* Vol. 37, pp. 412-424.

9. J. Hannan, D. Gilliland and S. B. Vardeman (in preparation). "Empirical Bayes and Compound Decision Theory: A Survey and Annotated Bibliography."

Table 3. Comparison of the contextual classifier using various unbiased estimator formulations and the uniform-priors non-contextual classifier.

Data Set	Classification	%Accuracy Overall	%Accuracy Average-by-Class
Hodgeman County, Kansas, 50-pixel-square Landsat (evaluated over lines and columns 6 through 50)	uniform-priors non-contextual	82.0	75.9
	4nn unbiased, spectral class whole image est. (nonadaptive)	83.1	75.8
	4nn unbiased, spectral class adaptive est., 25×25 from 25×25	84.0	77.8
	4nn unbiased, information class adaptive est., 25×25 from 35×35	84.0	78.0
Monroe County, Indiana, 50-pixel-square Landsat	uniform-priors non-contextual	83.1	82.7
	4nn unbiased, spectral class whole image est. (nonadaptive)	84.4	84.4
	4nn unbiased, spectral class adaptive est., 25×25 from 25×25	84.3	83.9
	4nn unbiased, information class adaptive est., 17×17 from 25×25	88.9	88.3
Tippecanoe County, Indiana, 50-pixel-square Landsat	uniform-priors non-contextual	81.8	83.4
	4nn unbiased, spectral class whole image est. (nonadaptive)	86.2	87.9
	4nn unbiased, spectral class adaptive est., 25×25 from 25×25	86.7	88.1
	4nn unbiased, information class adaptive est., 25×25 from 25×25	86.2	89.1
	4nn unbiased, information class adaptive est., 10×10 from 20×20	86.9	89.7

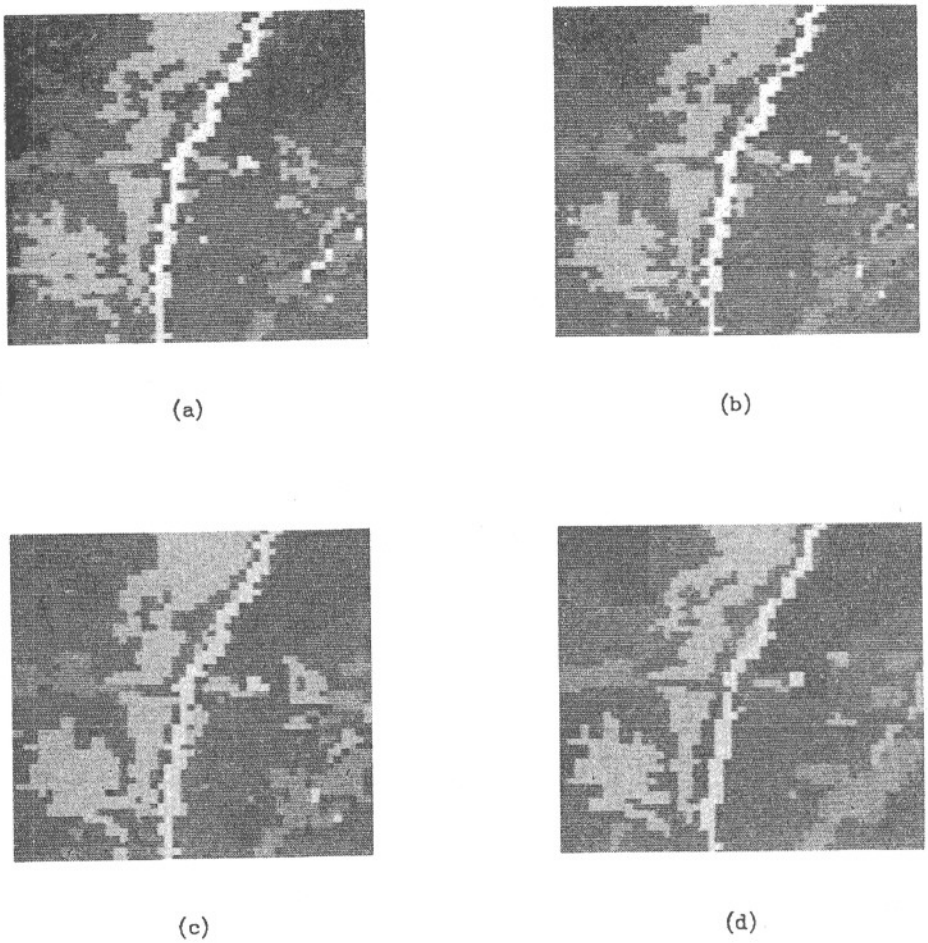

Figure 3. Visual comparison of classification results, Tippecanoe County, Indiana, Landsat data set. (a) uniform-priors no-context, (b) estimated-priors no-context, and (c) four-nearest-neighbor adaptive (17×17 from 27×27) unbiased estimator (d) reference image.

JAMES C. TILTON is candidate for the Ph.D. degree 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.

PHILIP H. SWAIN is associate 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).

STEPHEN B. VARDEMAN, formerly assistant professor of statistics at Purdue University, is assistant professor of statistics at Iowa State University. He received his B.S. and M.S. degrees in mathematics from Iowa State University in 1971 and 1973, respectively, and his Ph.D. in statistics from Michigan State University in 1975. He is a member of the Institute of Mathematical Statistics and the American Statistical Association. His research interests include decision theory and statistical pattern recognition.