AVOIDING CLASS-CONDITIONAL INDEPENDENCE ASSUMPTIONS IN IMAGE CLASSIFICATION

Ian Poole*

Department of Computer Science
University College London
Gower Street
LONDON WC1E 6BT

Abstract

Much published work on contextual image classification is based on an assumption of class-conditional independence (CCI) of the measurement data - equivalent to assuming that an ideal classifier will recover the underlying true scene, with errors evenly and randomly distributed as white noise. This paper proposes a simple alternative model which, it is argued, is more realistic in many applications and upon which useful theory can still be built. The new model is then used to investigate the effect on the accuracy of an object classifier which makes the CCI assumption in a domain where it is not valid.

1 INTRODUCTION

Consider the task of identifying forest cover in satellite or aerial imagery, based on per-pixel spectral characteristics. Suppose that some form of maximum a posteriori (MAP) per-pixel classifier has been trained on samples from the classes "forest" and "non forest". What form of results might one realistically expect to obtain? Figure 1 (a) shows a true scene, with shading indicating forest. Would the reader agree that the classification shown in (c) is far more realistic than that in (b)? Image (b) is of the type beloved of workers in image restoration, the errors being conveniently distributed in a random, uncorrelated fashion. In (c) the errors show a large degree of spatial "clumping", analogous to burst errors over a communications link. It is not difficult to imagine how such clumping of errors occurs. "Non-forest" will consist of a large number of sub-classes, some of which— eg parks and perhaps large gardens—will be more similar to forests than others such as lakes and buildings. This will be less true for the class of "forest", but even here there may for example be large burnt areas which would be more difficult to identify. As these subclasses will themselves be spatially correlated (clumped), then so will the classifier show clumping in the accuracy of its results.

One might contend that this is simply an issue of definition—that burnt patches are not forest at all, but this would not be a map maker's view. Classification *is* definition. It is widely appreciated in remote sensing that when designing a land use study, care must be taken to avoid specifying classes which are in reality a mixture of many underlying sub classes[10]. This advice is usually proffered in connection with the unimodality assumption made by commonly used pixel classifiers (eg Gaussian maximum liklihood). In practice however, one may have no choice but to use compounded data — ground truth is costly to collect and one may be forced to use training data which was perhaps collected for another purpose. Indeed, providing an appropriate non-parametric classifier is used (a nearest neighbour method for example), then this issue need have *no* effect on the accuracy of the per-pixel classification. However, as this paper will demonstrate, the *spatial correlation* of the sub-classes becomes highly significant when one wishes to use some context exploiting scheme to improve upon the initial classification; at a purely intuitive level we can see that it would be of no use consulting the immediate neighbours around the point marked "X" in figure 1 (c) in order to correct the misclassification there!

This paper investigates the effect of correlated sub-classes on contextual classifiers which ignore the phenomenon through their assumption of CCI on the measurement data. After explicitly stating the image model upon which contextual classifiers traditionally rest, a new model will be formulated which acknowledges the existence of underlying sub classes and their correlation. This model will then be *imposed* upon an existing object classifying scheme to derive insight into the degradation of performance that the inappropriate CCI assumption engenders.

2 IMAGE MODELS

2.1 The existing model

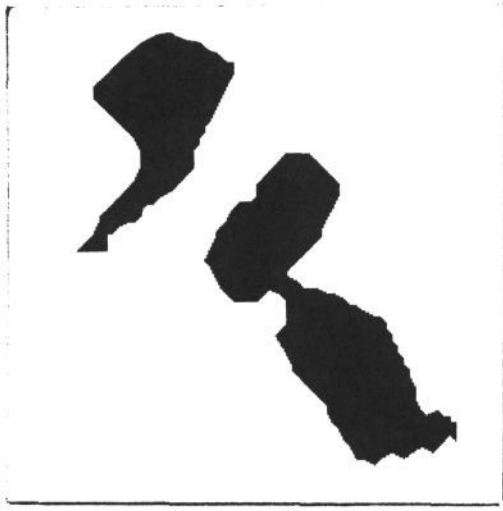
Many image classification schemes which aim to utilise contextual information [1,2,4,5,8,12,13,14] are implicitly or explicitly based on the following set of assumptions:

1. Associated with each pixel $i \in I$ is its class $Y_i \in \theta$, $\theta = \{1..K\}$

2. Pixel classes are locally correlated; it is often assumed that this correlation is positive and has no

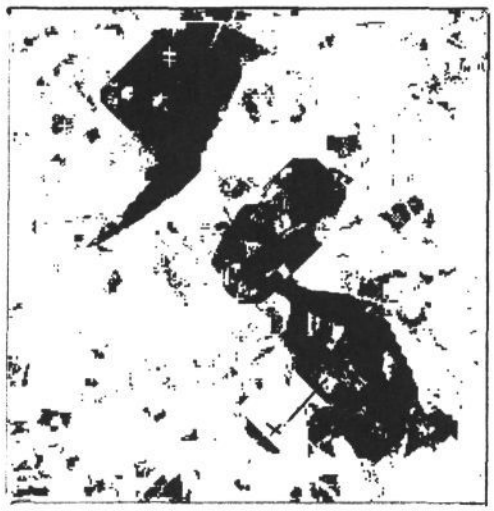
*Supported by an SERC CASE studentship in conjunction with the National Physical Laboratory, Teddington

AVC 1988 doi:10.5244/C.2.45



(a) True scene

(a) True scene with white noise



(a) True scene with correlated noise

Figure 1:

directional qualities, ie that

$$i \in \text{neighbours}(k) \Rightarrow P(Y^i = k \mid Y^k = j) \gg P(Y^i = k \mid Y^k \neq j)$$

3. Associated with each pixel i is some (probably multivariate) measurement data $X^i \in \Omega$.

4. Measurement data is class conditionally independent, ie

$$p(X^i \mid Y^i, X^k) = p(X^i \mid Y^i), \quad \forall i, k \in I$$

and thus

$$p(X^i, X^k \mid Y^i, Y^j) = p(X^i \mid Y^i) \cdot p(X^k \mid Y^j)$$

It is this latter assumption which leads to the prediction that the errors from a MAP classifier will be spatially uncorrelated, and with which this paper takes issue.

Many authors have recognized the shortcomings of this model (although Haralick[4] maintains that the CCI assumption holds true "in virtually all signal and image processing situations"). Various modifications have been proposed. Kalayeh and Landgrebe[6] recognize that correlation may be introduced by the target and model the effect with a causal Markov field on the observation data. Mohn et al[11] carry out a detailed and extensive simulation study of a number of context exploiting schemes using simulated scenes with data generated , by a model which assumes that the class-conditional correlation decreases exponentially with the distance between target points.

The approach in this paper however is simpler and perhaps more intuitive.

2.2 Formulation of the new model

We formally recognize the existence of subclasses and refer to them as *intrinsic classes*. The classes which are of actual interest to a given study we call *super classes*. Of course, it is only for the latter that training data/distributions will be directly available, though section 4.3 suggests a means of discovering these for the former also. Following is a formal statement of the proposed model.

1. Associated with each pixel $i \in I$ is its intrinsic class Z^i, $Z^i \in \Phi, \Phi = \{1..M\}$.

2. Associated with each pixel $i \in I$ is its super class Y^i, $Y^i \in \theta$, $\theta = \{1..K\}$.

3. There exists a transition matrix $\mathbf{T} = P(\mathbf{Z} \mid \mathbf{Y})$ which provides a probabilistic mapping from super class to intrinsic class.

4. Intrinsic classes are locally spatially correlated. It can often be assumed that this correlation is positive and has no directional qualities, thus

$$i \in \text{neighbours}(k) \Rightarrow P(Z^i = m \mid Z^k = m) \gg P(Z^i = m \mid Z^k \neq m)$$

In fact this will be taken to imply that intrinsic classes form into discrete regions.

5. Measurement vectors are *intrinsic* class conditionally independent

$$P(X^i \mid Z^i, X^k) = P(X^i \mid Z^i), \quad \forall i, k \in I$$

an thus

$$p(X^i, X^k \mid Z^i, Z^j) = p(X^i \mid Z^i).p(X^k \mid Z^k)$$

Some notes on notation. Superscripts will be reserved for indexing pixels and subscripts as a shorthand for indicating class (intrinsic or super). For example, $p(X \mid Y_j) \equiv p(X \mid Y = j)$. Super classes are indexed by j and intrinsic classes by m. Matrices are indicated bold, eg—

$$P(\mathbf{Y}) \equiv \begin{pmatrix} P(Y_1) \\ P(Y_2) \\ \vdots \\ P(Y_K) \end{pmatrix} \equiv \begin{pmatrix} P(Y=1) \\ P(Y=2) \\ \vdots \\ P(Y=K) \end{pmatrix}$$

ie, $P(\mathbf{Y})$ is a column matrix representing prior super class probabilities.

Note, also that $p(X \mid \mathbf{Y})$, and $p(X \mid \mathbf{Z})$ are column matrices of conditional distributions and that

$$p(X \mid \mathbf{Y}) = P(\mathbf{Z} \mid \mathbf{Y}) p(X \mid \mathbf{Z}) \tag{1}$$

or, in non matrix form

$$p(X \mid Y_j) = \sum_{m \in \Phi} P(Z_m \mid Y_j) \cdot p(X \mid Z_m) \tag{2}$$

as these will be useful later.

2.3 Special cases of T

In the case that $|\Phi| = |\theta|$ and the transition matrix $\mathbf{T} = P(\mathbf{Z} \mid \mathbf{Y})$ contains only 1s and 0s, then there is clearly a 1-to-1 deterministic mapping between intrinsic class and super class and the models of 2.1 and 2.2 are equivalent.

When $|\Phi| > |\theta|$ then there are more intrinsic classes than super classes (the most natural case). If each column of $\mathbf{T}$ contains only 1 non-zero element then we have the case of several intrinsic classes mapping deterministically to one super-class; if this is not the case then some of the super classes must be inherently indistinguishable, as even knowing the intrinsic class with certainty will not uniquely identify the super class. Clearly the rows of $\mathbf{T}$ must sum to 1.

3 HONESTY

It can be useful if a classifier is able to present its result in a probabilistic form, rather than as a categorical classification; that is, if the classifier generates the a posteriori probability (PP) vector $P(\mathbf{Y} \mid X)$ for each pixel. As well as the aid to interpretation that this may yield (when the vector is used to display a "probability image"), it will provide more information for any subsequent context exploiting phase (eg probabilistic relaxation[7]). Classifiers which have a statistical foundation will usually be able to provide this probabilistic assessment. However, what expectations should we have of the probabilities so produced? The vector should sum to one, certainly, but should we not also expect that the implied assessment of confidence be in some sense "honest"? We might say that a classifier is ***honest in the overall sense*** if an assessment of classification accuracy calculated by—

$$\int_{X \in \Omega} \max_j P(Y_j \mid X) \cdot P(X)\, dX \tag{3}$$

and using the *generated* PPs— closely matches the figures for *actual* accuracy achieved on test data. We could ask for more — that bands of probabilities are individually meaningful, so that, for example, if one were to select from a classified image, all pixels which were assigned a PP in the range 0.7-0.8, then one could expect that this selection of pixels would indeed achieve an accuracy of around 0.75.[1]

The author has observed that many probabilistic image classifiers are far from honest — particularly those which are context exploiting. For example, those based on probabilistic relaxation have a tendency to converge on a particular interpretation with near certainty when the actual results do not support this. If a classifier is correctly derived from statistical principals and all assumptions hold precisely true, then it will be, *a priori*, honest. The problem of course is that the real world rarely conforms to our idealized models and image classifiers at least often turn out optimistic (curiously, classifier pessimism appears to be rare).

4 IMPLICATIONS TO AN OBJECT CLASSIFIER

Landgrebe[9] describes an object classifier aimed at improving the accuracy of classification of remotely sensed data, using the assumptions stated in section 2.1 (though note Landgrebe's later work [6] which has already been mentioned above). This is the ECHO system — "Extraction and Classification of Homogeneous Objects". Following is a brief summary of this system.

Two distinct stages are involved. First the image is segmented into homogeneous regions; the details of the segmentation procedure used are not of interest here.

In the second stage, each region is classified as a whole. It is assumed that the segmentation process will have ensured that each region contains pixels from only one class. The joint distribution of the n pixels in each "object" is then treated as follows (denoting by $\widehat{X}$ the vector composed from all the measurements in the object).

$$p(\widehat{X} \mid Y_j) \equiv p(X^1, X^2 \ldots X^n \mid Y_j) \tag{4}$$

The assumption of (super) class conditional independence is then applied, permitting :-

$$p(X^1, X^2 \ldots X^n) \mid Y_j) = \prod_{i=1}^{n} p(X^i \mid Y_j) \tag{5}$$

but this violates the assumptions of the new model since CCI is assumed only at the *intrinsic* class level.

[1] This can be formulated more precisely, involving an integral over the prescribed range of probabilities

ECHO then uses a maximum likelihood (ML) decision rule to select the class for the whole object, however for the purpose of this discussion we will use a MAP or "Bayes minimum risk" decision rule since this gives the lowest theoretical misclassification. The ML rule is equivalent to the MAP rule with equal prior probabilities and a zero-one loss function [3].

4.1 Imposing the new model

We would like to answer the following questions:

1. What does the R.H.S of (5) equate to in terms of the intrinsic class model?
2. What is the 'correct' form of the joint conditional distribution on the L.H.S of (5) in terms of the intrinsic class model?
3. What is the theoretical classification accuracy for each of the two forms of the decision function for different sizes of object and given distributions under the intrinsic class model?
4. Will the a posteriori probabilities (PPs) predicted by each of the two decision functions be "honest", as discussed in section 3 above?

In order to answer these questions we must re-cast the decision function which was *actually* used in terms of the new model and then compare this with a "corrected" version which takes proper account of the model.

4.1.1 The decision functions

Using the MAP decision rule, both the ECHO original and the modified version will have the same basic form:

$$D(\widehat{X}) = j \text{ if } P(Y_j).p(\widehat{X} \mid Y_j) = \max_{k \in \theta} P(Y_k).p(\widehat{X} \mid Y_k)$$

The difference will lie in the way that $p(\widehat{X} \mid \mathbf{Y})$ is calculated. ECHO uses -

$$p(\widehat{X} \mid Y_j) = \prod_{i=1}^{n} p(X^i \mid Y_j) \tag{6}$$

which in terms of the new model translates using (2) to

$$\prod_{i=1}^{n} p(X^i \mid Y_j) \quad = \prod_{i=1}^{n} P(\mathbf{Z} \mid Y_j) p(X^i \mid \mathbf{Z}) \tag{7}$$

and expanding the matrix multiplication gives —

$$\prod_{i=1}^{n} \left(\sum_{m \in \Phi} P(Z_m \mid Y_j) \cdot p(X^i \mid Z_m) \right) \tag{8}$$

However, if we use the intrinsic class model correctly then we obtain:

$$p(\widehat{X} \mid Y_j) = P(\mathbf{Z} \mid Y_j) p(\widehat{X} \mid \mathbf{Z}) \tag{9}$$

or equivalently—

$$\sum_{m \in \Phi} P(Z_m \mid Y_j) \cdot p(\widehat{X} \mid Z_m) \tag{10}$$

We can now exploit intrinsic class conditional independence to expand $p(\widehat{X} \mid Z_m)$, ie

$$p(\widehat{X} \mid Z_m) = \prod_{i=1}^{n} p(X^i \mid Z_m) \tag{11}$$

And substituting (11) into (10) gives -

$$p(\widehat{X} \mid Y_j) = \sum_{m \in \Phi} \left(P(Z_m \mid Y_j) \cdot \prod_{i=1}^{n} p(X^i \mid Z_m) \right) \tag{12}$$

Clearly the two expressions for $p(\widehat{X} \mid Y_j)$ in equations (8) and (12) are not equivalent.

We will denote the decision function that uses equation (8) (ie the ECHO original) by $D_1(X)$, and the decision function which uses the modified version, equation (12) by $D_2(X)$. Of course $D_1(X)$ does not need the intrinsic class distributions $p(X \mid \mathbf{Z})$ since it could work directly from equation (6). We can now proceed to attempt an answer to question 3. posed above.

4.2 Effects on classification accuracy

The theoretical object misclassification rate of a decision rule $D_\alpha(\widehat{X})$, under the assumptions of the proposed model is (c.f [3, pg 21])

$$\sum_{j \in \theta} P(Y_j) \int_{\forall \widehat{X}} \lambda(D_\alpha(\widehat{X}) \mid j) \mathbf{p}(\widehat{X} \mid Y_j) \, d\widehat{X} \tag{13}$$

where $P(\widehat{X} \mid Y_j)$ is given by equation (12) and $\lambda(j_a \mid j_b)$ is loss on classifying class j_b as j_a; here we take the simple case of $\lambda(j_a \mid j_b) = 0$ if $j_a = j_a$, 1 otherwise. We can now use this equation to compare the classification accuracy of the decision rules $D_1(\widehat{X})$ and $D_2(\widehat{X})$ under various assumptions of intrinsic class conditional distributions $p(\widehat{X} \mid \mathbf{Z})$ and transition matrix $\mathbf{T} = P(\mathbf{Z} \mid \mathbf{Y})$. It may be possible to obtain results by comparing the rules' performance at the algebraic level, however no progress in this direction can yet be reported. For now then, we will have to be content with simply instantiating the required distributions and seeing what numbers come out.

4.2.1 A Numeric experiment

We will consider the simplest non-trivial case, with 2 super classes and 3 intrinsic classes, ie

$$\begin{aligned} K = 2, \quad \theta &= \{1, 2\} \\ M = 3, \quad \Phi &= \{1, 2, 3\} \end{aligned}$$

The measurement space Ω is limited and quantized as simply $\Omega = \{1, 2, 3, 4, 5\}$ (for reasons of computational efficiency). Integrals thus become summations.

Normal statistics are used for the intrinsic class distributions, although in view of coarse measurement space, these were tabulated and re-normalized to ensure they summed to exactly one. The normal distributions were—

$$\begin{aligned} p(X \mid Z_1) &\sim \mathrm{N}[\mu = 3, \sigma^2 = 4] \\ p(X \mid Z_2) &\sim \mathrm{N}[\mu = 3, \sigma^2 = .25] \\ p(X \mid Z_3) &\sim \mathrm{N}[\mu = 1, \sigma^2 = .25] \end{aligned}$$

These are illustrated in figure 2

The transition matrix $\mathbf{T}$ is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & .4 & .6 \end{pmatrix}$$

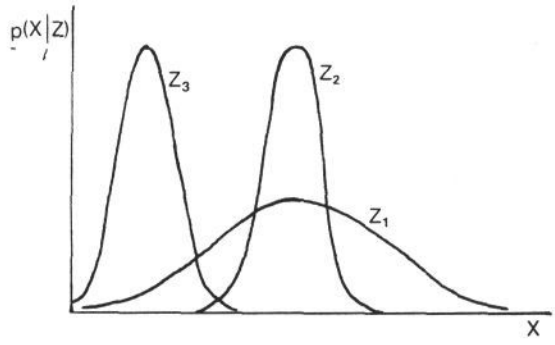


Figure 2: Intrinsic class distributions

Thus intrinsic class 1 maps uniquely onto super class 1 with intrinsic classes 2 and 3 contributing unevenly to super class 2.

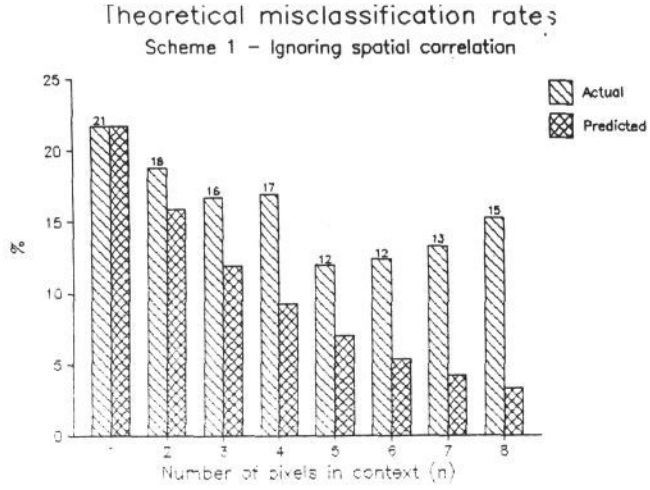
The prior probabilities used are

$$P(\mathbf{Y}) = \begin{pmatrix} .3 \\ .7 \end{pmatrix}$$

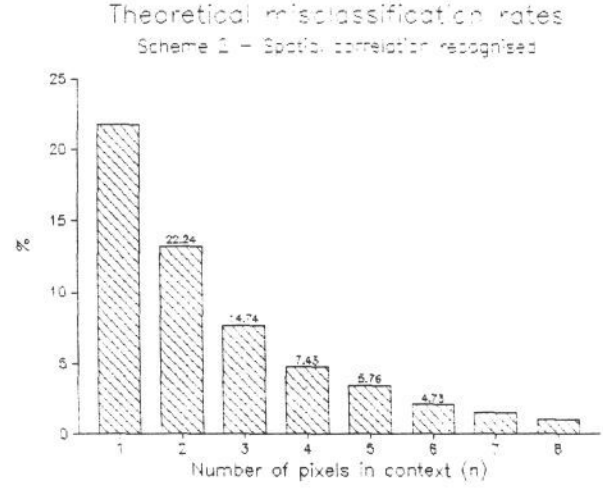
4.2.2 Results

The graphs in figure 3 show the misclassification rate — ie the value of eqn (13) as a percentage, for the two decision functions $D_1(\widehat{X})$ and $D_2(\widehat{X})$, against n, the number of pixels in the object (region). The actual performance of decision function $D_1(\widehat{X})$ (graph (a)), which ignores the correlation of intrinsic classes, is erratic, with the misclassification actually *increasing* after more than 4 pixels are involved in the decision (ie $n > 4$). In (b), which shows the performance for the 'corrected' decision rule, $D_2(\widehat{X})$, the performance consistently improves as more pixels are used in the decision.

Figure 3 (a) also shows the *predicted* misclassification rate for the D_1 decision function. These are also calculated from equation (13) but with $\mathbf{p}(\widehat{X} \mid Y_j)$ derived from formula (8). Notice the increasing disparity between the actual and predicted error rates as n is increased. The classifier becomes optimistic. Of course, the term "actual" here is somewhat false—it is the theoretically predicted "actual" rate *assuming* the proposed new model with the given distributions are precisely correct.



(a) D_1



(b) D_2

Figure 3: Misclassification rates

4.3 Estimation of $\mathbf{p}(X \mid \mathbf{Z})$ and $\mathbf{T}$

All of the above would be rather academic if there were no way of obtaining these distributions; we cannot contemplate gathering them explicitly, through more detailed training data. However, within the framework of an object classifier system, such a procedure can be suggested.

We assume that the super class training data is in the form of an image upon which randomly selected points have been labeled. Recall that the image is first segmented. The procedure is effectively a form of unsupervised learning:-

1. Apply the segmentation procedure to the training image. We assume that each segmented region may be regarded as representing an intrinsic class.

2. Gather distribution statistics for each of the arbitrarily numbered regions, ie intrinsic classes; it would seem reasonable to assume normal statistics. Where the statistics for two regions are not significantly different, their statistics can be pooled and the regions treated as belonging to one and the same intrinsic class. Thus Φ and $P(X \mid \mathbf{Z})$ can be obtained.

3. Construct the transition matrix $\mathbf{T}$ from the proportions of the super class training samples that fall into each intrinsic class. It may be that some intrinsic classes contain no, or too few labeled samples; in this case either that intrinsic class can be merged with its closest (in pattern space) neighbour, or perhaps more usefully, associated with an additional super class which implies "reject".

5 CONCLUSIONS

We have formulated a set of assumptions which aim to model the class conditional correlation encountered in image classification. These assumptions are based on the intuitive idea of there being spatially correlated "intrinsic classes" which are related to the 'super classes' (of interest in a particular study) in a probabilistic fashion via a transition matrix. When this model was imposed upon an object classifier which ignored the correlation, classification accuracy was considerably reduced and the predicted a posteriori probabilities became dishonestly optimistic as compared with the predictions under the new model.

The "results" presented here are theoretical— the motivation for the new model is entirely intuative at the present. Clearly further work is needed to construct an object classifier under this model in order to see if improved performance is *actually* achieved.

References

[1] J Besag, "On the statistical analysis of dirty pictures", *J. Royal statistical soc. B.,* vol. 48, pp. 259-302, 1986.

[2] C B Chittineni, "Utilization of spectral-spatial information in the classification of imagery data", *CGIP,* vol. 16, pp. 305-340, 1981.

[3] R O Duda and P E Hart, *Pattern classification and scene analysis,* Wiley, New York, 1973.

[4] R M Haralick, "An interpretation for probabilistic relaxation", *CVGIP,* vol. 22, pp. 388-395, 1983.

[5] R M Haralick and H Joo, "A Context Classifier", *IEEE Trans. on Geoscience and Remote Sensing,* vol. 24, no. 6, 1986.

[6] H M Kalayeh and D A Landgrebe, "Stochastic model utilizing spectral and spatial characteristics", *IEEE Trans. PAMI,* vol. 9, no. 3, pp. 457-461, 1987.

[7] J Kittler and J Illingworth, "Relaxation labelling algorithms - a review", *Image and Vision Computing,* vol. 3, no. 4, pp. 206-216, 1985.

[8] J Kittler and J Foglein, "On compatability and support functions in probabilistic relaxation", *CVGIP,* vol. 34, pp. 257-267, 1986.

[9] D A Landgrebe, "The development of a spectral-spatial classifier for Earth observational data", *Pattern Recognition,* vol. 12, pp. 165-175, 1980.

[10] T M Lillesand and R W Kiefer, *Remote sensing and image interpretation,* John Wiley & sons, 1987.

[11] E Mohn, N L Hjort and G O Storvik, "A simulation study of some contextual classification methods for remotely sensed data", *IEEE Trans. on Geoscience and Remote Sensing,* vol. 25, no. 6, pp. 796-804, 1987.

[12] S Peleg, "A new probabilistic relaxation scheme", *IEEE Trans. PAMI,* vol. 2, pp. 362-369, 1980.

[13] P H Swain, S B Vardeham and J C Tilton, "Contextuial classification of multispectral image data", *Pattern recognition,* vol. 13, no. 6, pp. 429-441, 1981.

[14] J C Tilton, S B Vardman and P H Swain, "Estimation of context for statistical classificatioon of multispectral image data", *IEEE Trans. on Geoscience and Remote Sensing,* vol. 20, no. 4, pp. 445-452, 1982.