# ADVANCED SPATIAL INFORMATION PROCESSES

# -MODELING AND APPLICATION

by

Mingchuan Zhang

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

APPROVED:

R. M. Haralick, Chairman

R. W. Ehrich

J. B. Campbell

K. B. Yu

J. W. Roach

October, 1985

Blacksburg, Virginia

# ADVANCED SPATIAL INFORMATION PROCESSES: MODELING AND APPLICATION

by

Mingchuan Zhang

R. M. Haralick, Chairman

Electrical Engineering

(ABSTRACT)

Making full use of spatial information is an important problem in information-processing and decision making. In this dissertation, two Bayesian decision theoretic frameworks for context classification are developed which make full use of spatial information. The first framework is a new multispectral image context classification technique which is based on a recursive algorithm for optimal estimation of the state of a two-dimensional discrete Markov Random Field (MRF). The implementation of the recursive algorithm is a form of dynamic programming. The second framework is based on a stochastic relaxation algorithm and Markov-Gibbs Random Fields. The relaxation algorithm constitutes an optimization using annealing. We also discuss how to estimate the Markov Random Field Model parameters, which is a key problem in using MRF in image processing and pattern recognition. The estimation of transition probabilities in a 2-D MRF is converted into two 1-D estimation problems. Then a Space-varying estimation method for transition probabilities is discussed.

## ACKNOWLEDGEMENTS

The author would like to thank Dr. Haralick, Dr. Ehrich, and Dr. Campbell for their encouragement, suggestions, criticisms, and bountiful patience in helping me for this dissertation.

The author also gives many thanks to his mother and wife for their continuous encouragement and moral support.

# CHAPTER I: INTRODUCTION

## (I-1) THE PROBLEM

An image encodes much information about the scene. The image data is contained in the brightness value at each point, and in the spatial configuration of brightness values.

What kind of knowledge about the intrinsic properties of objects can we get from the real world image ? This question is quite an important one for efficiently discriminating, analyzing, and identifying objects in spatial information processing.

From the image processing perspective, the world to be sensed is composed of units defined by the sensor. For digital imaging sensors, as a first approximation, the units can be thought of as small nonoverlapping areas, one such area for each picture element (pixel) in the image. The sensor makes an ordered set of measurements on each unit sensed. The ordered set of measurements is called a measurement vector or measurement pattern. Each value measured in this set is a number proportional to the energy received at a specific observation time. For example, remote sensing MSS data contains information about an object from measurements made at a distance from the object, i.e. , without actually coming into contact with it.

The quantity most frequently measured in present day remote sensing systems is the electromagnetic energy emanating from the object of interest.

When the pixel's information consists only of the sensor measurement pattern obtained from one observation time, the measurement pattern is called a multispectral feature vector. When spectral information from more than one observation time for the same object area are stacked in the same measurement pattern vector, this kind of pattern recognition is called multispectral multitemporal pattern recognition. When the measurement pattern for each pixel contains spectral information from its associated object's area as well as a neighboring objects area, or when the decision rule which makes category assignments uses the information from a pixel and some of its neighboring pixels, the pattern recognition is called spatial pattern recognition.

Multispectral scanner systems in remote sensing involve spectral data in quantitative format over a broad range of wavelengths. Computer-aided analysis techniques provide a powerful method for analyzing such quantitative spectral data. Spatial features, such as size, shape, texture or linear features, are also included in the scene from the image data. From an image-interpretation standpoint, the spatial features are often vitally important to achieve proper identification of the object. However, at present, the application of pattern-recognition algorithms to real image data

has not advanced operationally to the point where spatial features can be utilized as effectively as spectral features. It is a great challenge to effectively process and analyze spatial information in the real image. A typical representative of such a class of problems is pattern recognition of remotely sensed image data, which has been classically done by processing each pixel's information separately or independently.

There are two types of context information in most real world images. One is local or pixel-based context information; the other is high-level, or object-based context information. Some statistical classification methods have been developed for working on small neighborhoods using local contextual information. An alternative technique is to segment a scene into "objects" and then to classify the objects on the basis of their shapes and their interrelationships. But again, the information used for the classification has been strictly of a local nature. Although some attempts have been made to apply more global syntactic and semantic information to remote sensing data, these have so far met with only very limited success in specialized applications.

There is great potential in the effective use of high level contextual information (i.e. global contextual information) by means of "spatial reasoning". Contextual information extraction processes are significantly more powerful than the pixel-oriented processes. They are a very important tool

for spatial computing. Actually, a more appropriate term than "spatial reasoning" is "contextual reasoning", since what is actually referred to is reasoning based on any or all information available (local or global) about a given scene's location. Such processes would incorporate relatively high-level intelligence in the decision-making operations. Further study is needed in exploiting this area.

## (I-2) OVERVIEW

In this dissertation we will discuss some new advanced spatial information processing methods, such as a Dynamic programming approach for context classification under a Markov Random Field and Contextual Classification by Stochastic Relaxation. Chapter II contains the literature review for these areas. In the beginning of Chapter III, we create a statistical model for spatial information processes based on the Markov-Gibbs Random Field as a probabilistic model for digital image context information. Then we present three dynamic programming multispectral image context classifications algorithms, for optimal estimation of the state of a two-dimensional discrete Markov Random Field. The implementation of the recursive algorithm is a form of dynamic programming. In chapter III, three contextual decision rules for minimum error probability under the 2-D discrete Markov Random Field are discussed. The rules are characterized

by the range of their "look-ahead" capabilities. The forward-backward algorithm, which constantly takes a full look-ahead into account, always provides a result as good or better than the other algorithms using less look ahead. Finally, the computation requirement of this algorithm is discussed.

In recent years, several different contextual classification algorithms have been proposed. One limitation of these algorithms is the high computational cost as compared with context free classification. Thus, it is important to make contextual classification faster and more efficient.

The evolution of very-large-scale-integration(VLSI) technology has had a great impact on computer architecture. Many parallel algorithms in pattern recognition and image processing have been developed in recent years. With the realization that parallel algorithms might be implementable in special purpose VLSI hardware, many sophisticated methods which previously had appeared to be impractical began to appear to have practical implementations with special purpose VLSI hardware and parallel algorithms. Thus a highly parallel contextual classification algorithm, which is implemented by a neighborhood operator, is now practically possible.

In Chapter IV, a highly parallel context classification algorithm which is based on a stochastic relaxation algorithm and Markov-Gibbs Random Field, is discussed. First we motivate a Bayesian context decision rule, then introduce a Markov-Gibbs model for the original Landsat MSS image, and

then develop a new contextual classification algorithm, in which maximizing the posterior probability (MAP) is based on the stochastic relaxation and annealing method.

Here we make an analogy between image and statistical mechanic systems. Pixel gray levels and labels are viewed as states of atoms or molecules in a lattice-like physical system. The assignment of energy functions in a physical system determines its Gibbs distribution. Because of the Gibbs distribution-Markov Random Field (MRF) equivalence, this assignment also determines a MRF image model. The algorithm is highly parallel and exploits the equivalence between Gibbs distributions and Markov Random Fields (MRF).

In Chapter V, we present experimental results with both simulated and real multispectral remote sensing imagery, using the dynamic programming contextual classification method and the stochastic relaxation contextual classification method. We contrast these results with the context free classification to show how classification accuracy is greatly improved.

In chapter VI, we discuss how to estimate the Markov Random Field Model (MRF) parameters, which is a very important problem in using MRF in image processing and pattern recognition. The estimation of transition probabilities in a 2-D Markov Random Field can be converted into two 1-D problems, and then Maximum likelihood and Robust estimation methods of

transition probability are discussed. Furthermore, a confidence interval estimation technique for making sample size determinations for a desired parameter range at a specified confidence level is demonstrated.

## CHAPTER II: REVIEW OF PREVIOUS WORK

This Chapter contains a literature review of previous research on some spatial information processing techniques and their statistical models.

It is divided into 3 sections: Context Classification, Markov Random Field, and Gibbs distribution, and an optimization technique (dynamic programming and stochastic relaxation by annealing method).

### (II-1) CONTEXT CLASSIFICATION

Since the launch of the first Earth Resources and Technology Satellite (ERTS, later renamed Landsat) in July 1972, much work in Remote Sensing has been done by using pattern analysis and picture processing techniques for image classification. These classification methods, which have been widely used for applications such as classification of crops, are supported by well-defined statistical theoretical backgrounds. These methods, however, rely only on the multispectral characteristics of a point without considering spatial relations with its neighboring points. That is, these automatic classification algorithms for remote sensing data are done by processing each pixel's information separately or independently. A category assignment is made to each pixel purely on the basis of its own information. Most of the literature of pattern recognition deals with this simple (no context)

classification technique.

For multispectral context free pattern recognition problems, a very complete survey is given by Nagy (1972); Fu (1976), Fu and Rosenfeld (1976), Ho and Agrawala (1976), and Kanal (1972) also wrote survey papers. This type of multispectral classification can only exploit spectral or, in some cases, spectral and temporal information. There is no provision for using the coherent spatial information in the remote sensing MSS or TM data. As a matter of course, the simple methods can not satisfy the demand for high classification efficiency and accuracy in remote sensing applications. They have several intrinsic shortcomings : 1) They are very sensitive to random noise. 2) They can not handle areas with heavy texture. 3) For some aerial photographs, taken under different conditions, the reference spectral statistics of each category (obtained from one set of images), can not be directly applied to the classification of other pictures. 4) They cannot discriminate different objects with similar spectral properties because they rely only on multispecral characteristics. 5) The classification results can not give the information about the numbers and shapes of the objects in the scene (Nagy 1978).

In recent years, the effort to incorporate spatial information into the classification has become increasingly prevalent, and progress has been made. The use of context in pattern classification has been described by

many papers. The early contextual classification work has been done in the area of character and text recognition (Raviv 1967, Donaldson 1970, Edwards 1964). Toussaint (1978), in his paper " The Use of Context in Pattern Recognition ", presented a tutorial survey of techniques for using contextual information in pattern recognition; he emphasized the problem of text recognition.

Spatial information used for improving classification results is usually subdivided into two types -- textural and contextual. Texture refers to a description of the spatial variation within a contiguous group of pixels. The context of pixel ( or a group of pixels) refers to its spatial relationships with pixels in the remainder of the scene. Two main sources of context: 1) spatial pixel category dependences, and 2) two-dimensional correlation between pixels, are usually incorporated in the contextual classification approach.

Various measures of texture have been successfully used to texture features. The classification is made using these feature vectors in a method similar to that described for spectral data. N. Ahuja and A. Rosenfeld and R. M. Haralick (1980) used a simple way that combines neighboring pixel values or appropriate functions of them ( eg. mean variance) to the original vectors. An alternate method, in which new components can be derived from some texture descriptors, e. g., concurrence matrices or Fourier

coefficients , was proposed by R. M. Haralick, K. Shanmugam and Dinstein (1973), J. S. Weszka, C. R. Dyer and A. Rosenfeld (1976).

Problems with these methods are : excessive dimensionality of the augmented vector, slow processing, and poor performance at the object boundaries. Another disadvantage of textural measures is that there is an effective reduction in spatial resolution of the final classified image because an area has to be defined within which the measurements of texture are made. This is particularly disadvantageous when low resolution satellite data are used.

Contextual information can be taken into account at several different phases of the classification process. The first approach is to directly apply the contextual procedure into the raw multispectral image data. The classifier becomes a modification of the commonly used per-pixel statistical classifiers, and is based on the use of the spectral values of the immediate neighbors of a pixel (Welch and Salter 1971). Extension of this approach has been developed ( Swain et al. 1979, Swain et al. 1981 ) which permits consideration of the position of pixel in the local area for the purposes of land-cover classification. Where the relative locations of pixels in the local area are considered, it is possible to determine whether the pixel under consideration forms part of an edge or boundary between classes, or is part of a linear feature. Kettig and Landgrebe's ECHO classifier (1976) relies on a

two-stage procedure whereby homogeneous objects are recognized in local areas, and each object as a whole is spectrally classified using a non-contextual algorithm.

Contextual classifiers for linear feature detectors which may be used to identify roads and rivers have been widely discussed ( Bajcsy and Tavakoli 1976, VanderBrug 1976, Monotot 1977, and Gurney 1980). Some related procedures for recognizing edges can be used to interpret geologic features ( Goetz, al. 1975); to detect geological linearments ( Burdick and Speiver 1980); to assist regional delimitation (Strong and Rosenfeld 1973); and to aid in the registration of images from different dates (Nack 1977).

The second approach is to employ contextual post processing to the classified data. At least a proportion of the preclassification error may be corrected by reassigning classified pixels to another class. Bryant (1979) describes a contextual method for recognizing both homogeneous areas and boundary pixels followed by contextual reclassification. Also a very widely used method is to edit the classified pixels according to some rules. The postprocessing is governed by rules inferred from contextual knowledge. In the simplest case the postprocessing can be carried out by any type of majority filters (Rosenfeld and Kak 1982), special convolution or nonlinear filters (K. I. Itten and F. Fasler 1978, A. Rosenfeld and A. Kak 1982) or quite differently by syntactic rules (J. M. Brayer, P. H. Swain, K. S. Fu

1980).

Within the same category, but rather more sophisticated, is the approach advocated by Wharton (1982). After the classification of the picture, new feature vectors are composed from the class labels of the pixels in a given neighborhood. During a second pass these vectors are then classified. The contextual information is used in the second classification process. The common handicap of all of these methods is that they try to recover lost information and such attempts can be successful only to a limited extent.

The last approach combines the spectral and spatial (contextual) information and classifies pixels using both sources. This approach enables us to make the best use of the available information and, if necessary, additional pre- or post-processing operations can be subsequently performed. The contextual information is used by setting up a probabilistic model which incorporates contextual information. The model is then used for decision making. For example, to incorporate two-dimensional spatial correlations into a decision scheme the spatial stochastic modeling must first be established. Then the general decision rule is simplified under specific assumptions of practical significance. It is a sophisticated approach, and has been studied by several authors ( Tilton et al.; Yu and Fu 1983) in recent years.

Tilton, Vardeman and Swain (1982) use a P-context array which contains spatial information of (p-1) pixels surrounding and neighboring the

current pixel in the context pixel classification process. They derived the optimal decision rule for the context array and focused their attention on finding an unbiased estimate of the context function, which is a statistical characterization of the context to be used in the decision rule.

Yu and Fu (1983) also noted that the spectral information of the surrounding pixels is correlated with the center pixel being considered. They investigated the spatial correlation between pixels, developed a spatial stochastic recursive contextual analysis procedure based on the local frequency distribution of scene components, and developed a two-stage contextual classification.

## (II-2) MARKOV RANDOM FIELD MODEL

The use of the Markov Random Field (MRF) as a probabilistic model of digital images is pervasive in the literature. Under the Markov Random Field assumption, each pixel in the digital image is stochastic, stationary, and satisfies a conditional independence assumption. Observations of each pixel in the n band multispectral digital image is viewed as a n-tuple random vector, and a region is viewed as a finite sample of the two-dimensional random process describable by its statistical parameters. The Markov Random Field is a multidimensional generalization of the Markov chain. A time index of the the Markov chain is replaced by a space index in

the Markov Random Field.

The concept of a Markov Random Field originally came from attempts to put into a general probabilistic setting a very specific model named after the German physicist Enst Ising (1925). Ising discussed only the magnetic interpretation, and the same model has been found applicable to a number of other physical and biological system such as gases, binary alloyes, and cell strutures. A sociologically oriented application has been suggested by Weidlich.

The foundations of the theory of the Markov Random Fields may be found in Preston (1974) or Spitzer (1971). They argue that the Markov Random Field should enjoy the same wide variety of applications that Markov chains have.

Spitzer (1971) and S. Sherman (1973) showed that every Markov Random Field ( MRF) is a Gibbs Random Field (GRF). The proof that a Markov Random Field determines a Gibbs' measure is much more involved and may be found in Preston (1974), Grimmett (1973), Groffeath (1976), Kemmny, Snell (1976). The equivalence between Gibbs' measures and a Markov Random Field was established for lattices by Averintser (1970) and extended to graphs by Preston (1974), Grimmett (1973) and Groffeath (1973). Grimmett used the Mobius inversion theorem ( G. C. Rota 1964) to construct a natural expression for the potential function of a Markov field.

Markov Random Fields were first introduced into the pattern recognition community by Chow (1962) and Abend, Harley, and Kanal (1965). In the Abend, Harley and Kanal papers they investigated a causal Markovian dependence in imagery and based a classification method for binary random patterns on it.

Woods (1972) showed that when the distributions are Gaussian, the discrete Gauss Markov Random Field can be written as a equation, in which each pixel's value is a linear combination of the values in its neighborhood plus a correlated noise term.

Derin (1983) proposed a recursive Bayes smoothing procedure under a Markov Random Field Model. Jain and Angel (1974) used a nearest-neighbor system, white noise and a no blur assumption to achieve image restoration by a recursive filtering.

Hansen and Elliott (1980) proposed an algorithm based on a Markov Random Field Model for the segmentation of remotely sensed data with high levels of additive noise.

More recently, Elliott and Dervin (1983) has done MAP estimation, via dynamic programming, of very noisy but simple images. The major difference is the use of the Gibbs formulation and improvements in the algorithms.

Cooper and Sung (1983) discussed a boundary finding method using a Markov boundary model and deterministic relaxation scheme.

M. Hassner and J. Sklansky (1980) defined a MRF on a finite square lattice from an independent identically distributed (i.i.d) array of random variables, and then outlined an MRF parameter estimation method. Based on them a texture classification procedure was developed.

## (II-3) OPTIMIZATION TECHNIQUE : DYNAMIC PROGRAMMING AND STOCHASTIC RELAXATION BY ANNEALING METHOD:

In the Bayes framework for decision making, the goal is to find a best decision rule, which minimizes the expected loss. Usually, the contextual information is used by setting up a probabilistic model which incorporates contextual information. Then the general decision rule is simplified under the specific assumptions of the statistical model. The contextual classification procedure optimally handles the prior and contextual information by using the simplified decision scheme. Each such decision scheme has some combinatorial computational aspect to it.

The aim of combinatorial optimization is to find minimum or maximum values of a function of many independent variables. This function, usually called the cost function or objective function, represents a quantita-

tive measure of the "goodness" of a complex system. The cost function depends on the detailed configuration of the many parts of that system.

A difficult problem in iterative combinatorial optimization occurs when the system is started near some local minimum. The desired behavior is to find the global minimum and not to fall into a local minimum. For constraint satisfaction tasks, the system must try to escape from local minima in order to find the configuration that is the global minimum given the current input.

Metropolis et al. (1953) introduced a simple way to get out of the local minima which was to occasionally allow jumps to configurations of higher "energy". Kirkpatrick et al. (1982) proposed simulated annealing for optimization, in which at high "temperatures," the system ignores small energy difference and approaches equilibrium ( optimal or near optimal status ) rapidly.

In recent years, the application of stochastic relaxation has raised many issues. Kirkpatrick first used the simulated annealing technique into the physical design of a computer. Geman and Geman (1983) discussed the relationship of a simulated annealing technique to Markov Random Fields. They introduced these concepts into image restoration, and gave a few simple results using synthetic images. Hinton and Sejonwski (1984) used it on neural modeling of inference and learning.

One effective way to find a decision rule, minimizing the expected loss function under a Bayesian Model and Markov Random Field assumption, is the Viterbi algorithm (1967). The Viterbi algorithm (VA) is a recursive optimal solution to the problem of estimating the parameters of a discrete finite state Markov process. It has already had a significant impact on understanding of certain problems, notably in the theories of convolutional codes (G. D. Forney,1972) and text recognition (D. L. Neuhoff 1975; J. Raviv 1967), intersymbol interference (G. D. Forney 1972). Haralick (1983) argued that the Viterbi algorithm can be used to determine the Bayes labeling under a Markov Random Field assumption.

# CHAPTER III: DYNAMIC PROGRAMMING APPROACH FOR CONTEXT CLASSIFICATION USING THE MARKOV RANDOM FIELD

In this chapter, we develop set of multispectral image context classification techniques which are based on a recursive algorithm for optimal estimation of the state of a two-dimensional discrete Markov Random Field. The first task of this approach is to define the two-dimensional discrete Markov Random Field Model. Three contextual decision rules for achieving minimum error probability under the 2-D discrete Markov Random Field are discussed in this chapter. The rules are characterized by the range of the their look-ahead capabilities. The values of their parameters are obtained from local measurements under a model assuming noise independence of the Markov Random Field Model. The three approaches differ only in the amount of context they handle.

The three recursive algorithms are forms of dynamic programming. Because the estimation equations of the recursive algorithm are quite simple, the computation complexity of the approach is low. It is shown that recursive contextual classifications can improve classification performance, as compared to noncontextual classification. In addition, this algorithm has the advantage over other techniques in that it handles multispectral data naturally and simultaneously.

## (III-1) Statement of problem :

Consistent with the two-dimensional (2-D) discrete Markov Random Field for multispectral image processing applications we assume a random observation vector $D_{ij}$, each of whose components takes one gray tone value from the set $D = \{ 0, \ldots, s \}$, where s is a final integer value, and the pixel position ( i,j ) is defined on the 2-D dimensional finite set of I x J.

Unlike 1-D discrete time series, where the existence of a preferred direction is inherently assumed, no such preferred ordering of the neighborhood is appropriate. In other words, the notion of "past" and "future" (as understood in unilateral 1-D Markov processes) is restrictive in 2-D as it implies a particular ordering in which the observations are scanned top down and left to right. It is quite possible that an observation at a pixel p may be dependent on surrounding observations in all four directions: north, south, east and west.

The Markov Random Field models may be defined as below :

Let $\{ d(s), s \in I \times J \}$ be an observed pixel value from an image, where I designates the row index set , and J designates the column index set. The Markov Random Field Model characterizes the statistical dependency among pixels by requiring that

$$P(d(s) \mid \text{all } d(r),\ r \neq s) = P(\ d(s) \mid \text{all } d(s+r),\ r \in N\ ),$$

where N is a symmetric neighbor set. For instance N ={ (0,1), (0,-1), (-1,0), (1,0) } corresponds to the simplest Markov model where the dependence is only on the nearest neighbors. By including more neighbors we can construct a higher order Markov model. Since the model is defined only for symmetric neighbor sets, often N is equivalently described by means of an asymmetrical neighbor set $N_s$ ; i.e. if $r \in N_s$ then - r does not belong to $N_s$ and $N = (\ r \mid r \in N_s \text{ or } -r \in N_s\ )$.

Only the nearest neighbors will be used in this chapter for N. Let the observation vector $d_{ij}$ at pixel( i,j ) have fixed but unknown classification $C_{ij}$, as shown in Figure 3.1. The assignment of the best category label $C_{ij}$ to the pixel is based on the conditional probability $P(\ C_{ij} \mid d_{ij}\ )$ when using pixel independent processing.

In contrast, context dependent processing uses more than the locally observed d in order to assign the best category label to each pixel. Full contextual processing uses all of the observed data from entire image. Contextual processing uses a substantial sized context neighboring the unit. As in the one-dimensional case (Devijver 1984), context dependent processing based on the Markov Random Field assumption can be achieved with three kinds of algorithms which can be characterized by their look-ahead capabilities.

Before presenting these algorithms, we must first give the notation and assumptions for the two-dimensional Markov Random Field under which the algorithm can be derived.

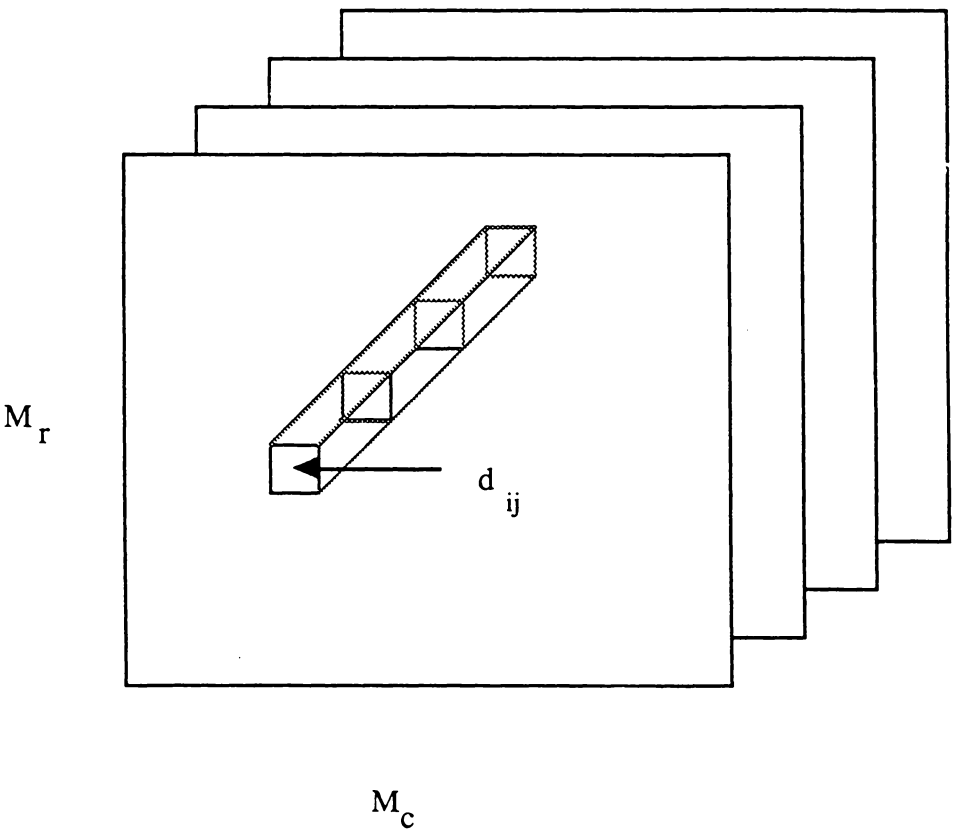


Figure 3.1 Two dimensional discrete Markov Random Field for multispectral image processing application. $M_r$ : row size of image, $M_c$ : column size of image; I : row index set of image, J: column index set of image. $d_{ij}$ : random obervation vector, each of whose components takes one gray tone value from the pixel position (i,j) which is an element of the 2-D finite set I x J. The Markov Random Filed is characterized by : $P(d_{ij} \mid \text{all } (l,k); (i,j) \neq (l,k)) = P(d_{ij} \mid \text{all } d_{lk}, l,k \in N)$; where N is symmetric neighbor set.

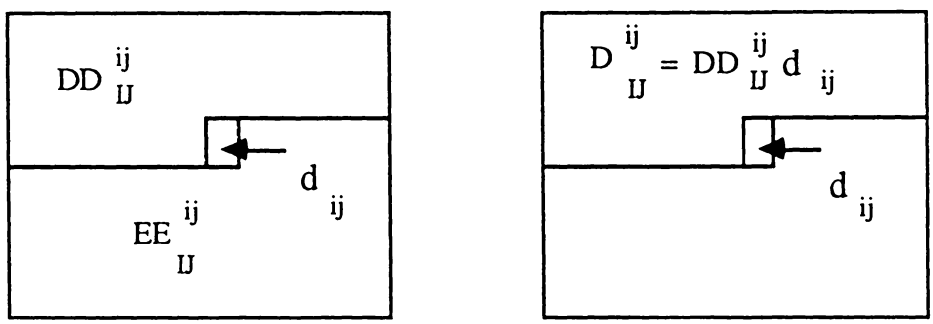


Figure 3.2 Notation for a two pass recursive algorithm. $d_{ij}$ : an observed measurement vector from pixel ( i,j ); $D_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left or above (i,j), including (i,j). $E_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right or below (i,j), including (i,j). $DD_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left or above (i,j), excluding (i,j). $EE_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right or below (i,j), excluding (i,j).

**(III-2.1) Notation :**

(1) $M_r$ : designates the number of rows in an image.

(2) $M_c$ : designates the number of columns in an image.

(3) I : designates the row index set of an image

$I = \{ 1,.....,M_r \}$

(4) J : designates the column index set of an image

$J = \{ 1,.....,M_c \}$

(5) ( i,j ) : designates a two-dimensional image position.

(6) R : the set of possible measurement vectors.

(7) $\Omega$ : the set of possible categories.

(8) $d_{ij}$ : an observed measurement vector from pixel ( i,j );

$d_{ij} \in R$.

(9) $C_{ij}$ : an assigned category label C to pixel ( i,j );

$C_{ij} \in \Omega$.

(10) $D_{IJ}$ : the collection of all observed measurements from the set of I x J pixels.

(11) $D_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left or above (i,j), including (i,j).

(12) $E_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right or below (i,j), including (i,j).

(13) $DD_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left or above (i,j), excluding (i,j).

(14) $EE_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right or below (i,j), excluding (i,j).

see figures 3.1 and 3.2 for the geometric picture of the above sets.

## (III-2.2) Three recursive algorithms :

The first context algorithm uses only past context with no look-ahead context (Fig 3.3). The assignment of the best category label for pixel ( i,j ) depends only on using the set of all observed data values $d_{lk}$, for ( l,k ) to the left or above ( i,j ), including ( i,j ) (ie. $D_{IJ}^{ij}$). The decision rule is:

assign pixel (i,j) to category $C_{ij}$ if

$$P( C_{ij} \mid D_{IJ}^{ij} ) \geq P( Z_{ij} \mid D_{IJ}^{ij} ) \qquad \text{for all } Z_{ij} \in \Omega$$

The second algorithm uses some fixed look-ahead. With n-step look-ahead, the context dependent processing makes the assignment of the best category label using $D_{IJ}^{ij}$ and the (n+1) x (n+1) neighborhood whose center pixel is ( i,j ). Figure 3.4 shows the one-step look-ahead algorithm, its decision rule is that

assign pixel ( i,j ) to category $C_{ij}$ if

$$P( C_{ij} \mid D_{IJ}^{ij}, d_{ij+1}, d_{i+1j}, d_{i+1j+1} ) \geq P( Z_{ij} \mid D_{IJ}^{ij}, d_{ij+1}, d_{i+1j}, d_{i+1j+1} )$$

for all $Z_{ij} \in \Omega$

The last algorithm uses the largest possible look-ahead. The context dependent processing is based upon all the observed data from all pixels to decide the best label (figure 3.2). The decision rule is :

assign pixel ( i,j ) to category $C_{ij}$ if

$$P( C_{ij} \mid D_{IJ}^{ij} ) \geq P( Z_{ij} \mid D_{IJ}^{ij} ) \qquad \text{for all } Z_{ij} \in \Omega$$

This algorithm is sometimes referred to as the "foward-backward look-ahead algorithm".

In a word, the above context dependent algorithms make the assignment of the best category label based on the conditional probabilities $P( C_{ij} \mid D_{IJ}^{ij} )$, $P( C_{ij} \mid D_{IJ}, d_{1k} ; (1,k) \in (n+1) \times (n+1)$ neighborhood whose

center pixel is ( i,j ), and $P(C_{ij} \mid D_{IJ})$, respectively.

In section (III-3) we will show how to compute these conditional probabilities using recursive neighborhood operators under the assumption of a Markov Random Field.

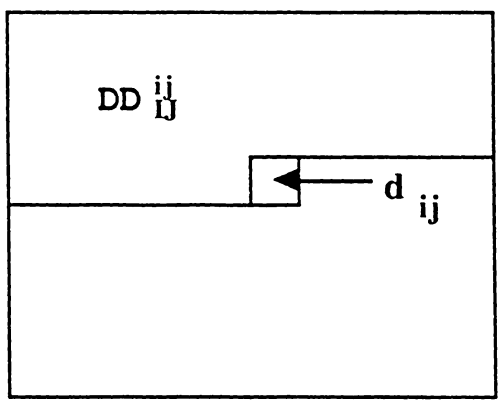


Figure 3.3 Notation for no look-ahead recursive context classification algorithm. $DD_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left or above (i,j), excluding (i,j). $d_{ij}$ : an observed measurement vector from pixel ( i,j ).

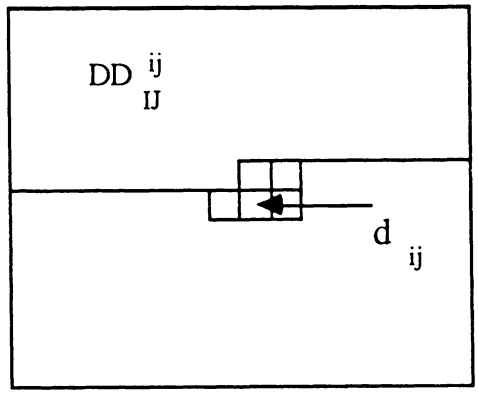


Figure 3.4 Notation for fixed step look-ahead recursive context classification algorithm. $DD_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left or above (i,j), excluding (i,j). $d_{ij}$ : an observed measurement vector from pixel ( i,j ).

**(III-2.3) assumptions :**

a) We assume that an n-tuple of measurements is determined by some local measuring process. Given the true category of a pixel, all measurements of all pixels are independent of each other. In effect this says that given the true state of affairs, the observed measurement variations are independent. Hence,

$$P(D_{IJ} \mid C_{ij}) = P(DD_{IJ}^{ij} \mid C_{ij})P(EE_{IJ}^{ij} \mid C_{ij})P(d_{ij} \mid C_{ij})$$

$$P(D_{IJ}^{ij} \mid C_{ij}) = P(DD_{IJ}^{ij} \mid C_{ij})P(d_{ij} \mid C_{ij})$$

$$P(E_{IJ}^{ij} \mid C_{ij}) = P(EE_{IJ}^{ij} \mid C_{ij})P(d_{ij} \mid C_{ij})$$

b) the n-tuple measurement of pixel(i,j) depends only upon the true interpretation associated with pixel(i,j) and does not depend upon any relationship pixel(i,j) may have with other units or upon the interpretation associated with any other unit.

$$P(d_{ij} \mid C_{ij}, D_{IJ}^{ij-1}) = P(d_{ij} \mid C_{ij})$$

$$P(d_{ij-1} \mid C_{ij-1}, C_{i-1j}, D_{IJ}^{ij-1}) = P(d_{ij-1} \mid C_{ij-1})$$

$$P(d_{ij} \mid C_{ij}, E_{IJ}^{ij+1}) = P(d_{ij} \mid C_{ij})$$

$$P(d_{ij+1} \mid C_{ij+1}, C_{i+1j}, E_{IJ}^{ij+1}) = P(d_{ij+1} \mid C_{ij+1})$$

c) given the true categories of some of neighboring pixels, the observed measurement data for all pixels before the current pixel, in raster scan order, tells nothing more about the current pixel's category.

$$P( C_{ij} \mid C_{ij-1} , C_{i-1j} , D_{IJ}^{ij-1} ) = P( C_{ij} \mid C_{ij-1} , C_{i-1j} )$$

$$P( C_{ij} \mid C_{ij+1}, C_{i+1j}, E_{IJ}^{ij+1} ) = P( C_{ij} \mid C_{ij+1} , C_{i+1j} )$$

d) given the measurements of all pixels up to and including the current pixel, the true category of its diagonally adjacent neighbor tells nothing more.

$$P( C_{ij-1} \mid C_{i-1j} , D_{IJ}^{ij-1} ) = P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$P( C_{ij+1} \mid C_{i+1j} , E_{IJ}^{ij+1} ) = P( C_{ij+1} \mid D_{IJ}^{ij+1} )$$

e) approximation: (conditioning on a set which omits a row of data values should not make much of a difference)

$$P( C_{i-1j} \mid D_{IJ}^{ij-1} ) = P( C_{i-1j} \mid D_{IJ}^{i-1j} )$$

$$P( C_{i+1j} \mid E_{IJ}^{ij+1} ) = P( C_{i+1j} \mid E_{IJ}^{i+1j} )$$

f) A weaker assumption than the independence of neighboring categories is the Markov Random Field assumption, which is used when the true interpretation of any unit given the true interpretations of all the surrounding depends only upon the interpretation of the nearest neighboring

pixel.

In our four neighbor system, given the true interpretation of pixel (i,j), the categories of diagonal pixels are independent of each other.

$$P( C_{ij-1} , C_{i-1j} \mid C_{ij} ) = P( C_{ij-1} \mid C_{ij} ) P( C_{i-1j} \mid C_{ij} )$$

$$P( C_{ij+1} , C_{i+1j} \mid C_{ij} ) = P( C_{ij+1} \mid C_{ij} ) P( C_{i+1j} \mid C_{ij} )$$

## (III-3) Implementation of the recursive algorithm :

The recursive algorithms which we present in this chapter determine the Bayes labeling under the Markov field assumption by determining the labeling $\{ C_{ij} : i = 1,...,I ; j = 1,...,J \}$ maximizing $P( C_{ij} \mid D_{IJ}^{ij} )$ ( past contextual no look-ahead ), $P( C_{ij} \mid D_{IJ}^{ij}, d_{lk} ; (l,k) \in ( n+1 ) \times ( n+1 )$ neighborhood whose center pixel is (i,j)) (fixed n-step look-ahead), and $P( C_{ij} \mid D_{IJ} )$ (forward-backward).

The first two algorithms can be regarded as special cases of the forward-backward algorithm.

In this section we first derive the forward-backward algorithm in detail, and then give brief discussions about the implementation of past contextual no look-ahead and fixed look-ahead algorithms.

In chapter V, we will give comparative results of all three algorithms.

### (III-3.1) Two pass forward-backward algorithm :

To compute the conditional probability $P( C_{ij} \mid D_{IJ} )$, we can argue as follows: from the definition of conditional probability:

$$P( C_{ij} \mid D_{IJ} ) = \frac{P( D_{IJ} \mid C_{ij} )P( C_{ij} )}{P( D_{IJ} )} \qquad (3.1)$$

Then

$$P( C_{ij} \mid D_{IJ} ) = \frac{P( DD_{IJ}^{ij} , EE_{IJ}^{ij} , d_{ij} \mid C_{ij} )P( C_{ij} )}{P( D_{IJ} )}$$

$$= \frac{P( DD_{IJ}^{ij} \mid C_{ij} )P( EE_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )P( C_{ij} )}{P( D_{IJ} )}$$

$$= \frac{P( DD_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )P( EE_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )P( C_{ij} )}{P( d_{ij} \mid C_{ij} )P( D_{IJ} )}$$

$$= \frac{P( D_{IJ}^{ij} \mid C_{ij} )P( E_{IJ}^{ij} \mid C_{ij} )P( C_{ij} )}{P( D_{IJ} )P( d_{ij} \mid C_{ij} )}$$

$$= \frac{P(C_{ij} \mid D_{IJ}^{ij} )P( D_{IJ}^{ij} )P( C_{ij} \mid E_{IJ}^{ij} )P( E_{IJ}^{ij} )P( C_{ij} )}{P( C_{ij} )^2 P( d_{ij} \mid C_{ij} )P( D_{IJ} )}$$

$$= \frac{\dfrac{P( C_{ij} \mid D_{IJ}^{ij} )P( C_{ij} \mid E_{IJ}^{ij} )}{( P( C_{ij} )P( d_{ij} \mid C_{ij} ) )}}{\dfrac{P( D_{IJ} )}{( P( D_{IJ}^{ij} )P( E_{IJ}^{ij} ) )}} \qquad (3.2)$$

In equation (3.2) probability distributions $P( C_{ij} )$ and $P( d_{ij} \mid C_{ij} )$ are either known or estimated in the supervised remote sensing classification situation. The denominator is some constant not depending on the true category. Then to determine the maximizing $C_{ij}$ , the problem becomes one of calculating: $P( C_{ij} \mid D_{IJ}^{ij} )$ and $P( C_{ij} \mid E_{IJ}^{ij} )$.

Now, we consider $P( C_{ij} \mid D_{IJ}^{ij} )$ :

$$P( C_{ij} \mid D_{IJ}^{ij} ) = P( C_{ij} \mid D_{IJ}^{ij-1}, d_{ij} )$$

$$= \frac{P(C_{ij}, d_{ij} \mid D_{IJ}^{ij-1} )P( D_{IJ}^{ij-1} )}{P( D_{IJ}^{ij} )}$$

$$= \frac{P( d_{ij} \mid C_{ij} , D_{IJ}^{ij-1} )P( C_{ij} , D_{IJ}^{ij-1} )}{P( D_{IJ}^{ij} )}$$

using assumption b :

$$= \frac{P( d_{ij} \mid C_{ij} )P( C_{ij} , D_{IJ}^{ij-1} )}{P( D_{IJ}^{ij} )}$$

$$= \frac{( \sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij-1} , C_{i-1j} , C_{ij} , D_{IJ}^{ij-1} ) )P( d_{ij} \mid C_{ij} )}{P( D_{IJ}^{ij} )}$$

$$= \sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} , D_{IJ}^{ij-1} ) P( C_{ij-1} , C_{i-1j} , D_{IJ}^{ij-1} )$$

$$* \frac{P( d_{ij} \mid C_{ij} )}{P( D_{IJ}^{ij} )}$$

using assumption c :

$$= [ \sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} ) P( C_{ij-1} \mid C_{i-1j} , D_{IJ}^{ij-1} )$$

$$* \frac{P( C_{i-1j} , D_{IJ}^{ij-1} ] * P( d_{ij} \mid C_{ij} )}{P( D_{IJ}^{ij} )}$$

using assumption d :

$$= [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} ) P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \frac{P( C_{i-1j} \mid D_{IJ}^{ij-1}) ] * P( d_{ij} \mid C_{ij} )P( D_{IJ}^{ij-1} )}{P( D_{IJ}^{ij} )}$$

$$= [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} ) P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \frac{P(C_{i-1j} \mid D_{IJ}^{i-1j} , d_{i-1j+1} ,.., d_{i-1J} , d_{i1} ,.., d_{ij-1} ) ] P(d_{ij} \mid C_{ij} )P(D_{IJ}^{ij-1})}{P(D_{IJ}^{ij} )}$$

using approximation (e) :

$$= [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} ) P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \frac{P( C_{i-1j} \mid D_{IJ}^{i-1j} )]*P( d_{ij} \mid C_{ij} )P( D_{IJ}^{ij-1})}{P( D_{IJ}^{ij} )}$$

so finally we have the formula (3.3)

$$P(C_{ij} \mid D_{IJ}^{ij} ) = [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} ) P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \frac{P( C_{i-1j} \mid D_{IJ}^{i-1j} )]*P( d_{ij} \mid C_{ij} )}{\Delta_D} \qquad (3.3)$$

where

$$\Delta_D = \sum_{k_{ij}} P( d_{ij} \mid k_{ij} ) \sum_{k_{ij-1}} \sum_{k_{i-1j}} P( k_{ij} \mid k_{i-1j} , k_{ij-1} )$$

$$* P( k_{ij-1} \mid D_{IJ}^{ij-1} ) P( k_{i-1j} \mid D_{IJ}^{i-1j} )$$

which is normalizing constant, not dependent on the optimizing category label.

The recursive formula for $P( C_{ij} \mid D_{IJ} )$ (3.3) requires a left right top to bottom recursive scan of the image. Similarly $P( C_{ij} \mid E_{IJ} )$ requires a recursive scan right left bottom to top as shown in equation (3.4).

$$P(C_{ij} \mid E_{IJ}^{ij} ) = [\sum_{C_{ij+1}} \sum_{C_{i+1j}} P( C_{ij} \mid C_{i+1j} , C_{ij+1} ) P( C_{ij+1} \mid E_{IJ}^{ij+1} )$$

$$* \frac{P( C_{i+1j} \mid E_{IJ}^{i+1j} )]*P( d_{ij} \mid C_{ij} )}{\Delta_E}$$

where

$$\Delta_E = \sum_{k_{ij}} P( d_{ij} \mid k_{ij} ) \sum_{k_{ij+1}} \sum_{k_{i+1j}} P( k_{ij} \mid k_{i+1j} , k_{ij+1} )$$

$$* P( k_{ij+1} \mid E_{IJ}^{ij+1} ) P( k_{i+1j} \mid E_{IJ}^{i+1j} )$$

In (3.3) and (3.4) $P( d_{ij} \mid C_{ij} )$ is either known or estimated in the supervised remote sensing classification situation, and the denominator $\Delta$ is some constant not dependent on the true category. In fact, it is a normalizing constant, which makes $\sum_{C_{ij}} P( C_{ij} \mid E_{IJ}^{ij} ) = 1$ The transition probabilities

$P( C_{ij} \mid C_{ij-1} , C_{i-1j} )$, $P( C_{ij} \mid C_{i+1j} , C_{ij+1} )$, in (3.3) and (4.4) can be estimated from the conventional non-contextual preclassification results.

When i =1 or j =1, $D_{IJ}^{ij}$ and $E_{IJ}^{ij}$ are out of the image. For these cases we assume

$$\sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} ) P( C_{ij-1} \mid D_{IJ}^{ij-1} ) P( C_{i-1j} \mid D_{IJ}^{i-1j} )$$

$$\sum_{C_{ij+1}} \sum_{C_{i+1j}} P( C_{ij} \mid C_{i+1j} , C_{ij+1} ) P( C_{ij+1} \mid E_{IJ}^{ij+1} ) P( C_{i+1j} \mid E_{IJ}^{i+1j} )$$

are constant, that is, it is not dependent on the true category. So, the initial values are evaluated from $P( d_{ij} \mid C_{ij} )$.

Finally, the application of the Bayes decision rule requires calculation of the following formulas for finding the class $C_{ij}$ which produces the maximum probability $P(C_{ij} \mid D_{IJ})$.

$$P(C_{ij} \mid D_{IJ} ) = \frac{P( C_{ij} \mid D_{IJ}^{ij} ) P( C_{ij} \mid E_{IJ}^{ij} )}{\Delta}$$

$$\Delta = \frac{P( d_{ij} \mid C_{ij} ) P( D_{IJ} )}{P( D_{IJ}^{ij} ) P( E_{IJ}^{ij} )}$$

$P( C_{ij} \mid D_{IJ}^{ij} )$ is given by :

$$\begin{cases} P(\, d_{ij} \mid C_{ij}\,) & \text{where} \quad (i=1 \text{ or } j=1) \\ P(d_{ij} \mid C_{ij}) \sum_{C_{i-1j}} \sum_{C_{ij-1}} P(C_{ij} \mid C_{i-1j}, C_{ij-1}) P(C_{i-1j} \mid D_{IJ}^{ij}) P(C_{i-1j} \mid D_{IJ}^{ij-1}) & \text{otherwise} \end{cases}$$

$P(\, C_{ij} \mid E_{IJ}^{ij}\,)$ is given by :

$$\begin{cases} P(d_{ij} \mid C_{ij}) & \text{where} \quad (i=1 \text{ or } j=1) \\ P(d_{ij} \mid C_{ij}) \sum_{C_{i+1j}} \sum_{C_{ij+1}} P(C_{ij} \mid C_{i+1j}, C_{ij+1}) P(C_{ij+1} \mid E_{IJ}^{ij+1}) P(C_{ij+1} \mid E_{IJ}^{i+1j}) & \text{otherwise} \end{cases}$$

(3.5)

**(III-3.2) Four pass algorithm :**

A more symmetric and accurate model to solve the conditional probability $P( C_{ij} \mid D_{IJ} )$ is the four blocks model (i.e., we consider propagation of conditional probability in four directions rather than in two directions only.) The notation for the recursive algorithm in the four blocks case is shown in Figure 3.5. Note that some of the symbols we use in this section have a different meaning than in the previous section.

(1) $D_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left of (i,j) or to the left and above (i,j), including (i,j).

(2) $E_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right of (i,j) or to the right and above (i,j), including (i,j).

(3) $F_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left of (i,j) or to the left and below (i,j), including (i,j).

(4) $G_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right of (i,j) or to the right and below (i,j), including (i,j).

(5) $DD_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left of (i,j) or to the left and above (i,j), including (i,j).

(6) $EE_{IJ}^{ij}$ : set of all observed data values $d_{ij}$ for (l,k) to the right of

(i,j) or to the right and above (i,j), including (i,j).

(7) $FF_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right of (i,j) or to the right and below (i,j), including (i,j).

(8) $GG_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right of (i,j) or to the right and below (i,j), including (i,j).

See figure 3.5 for the geometric picture of the above sets.

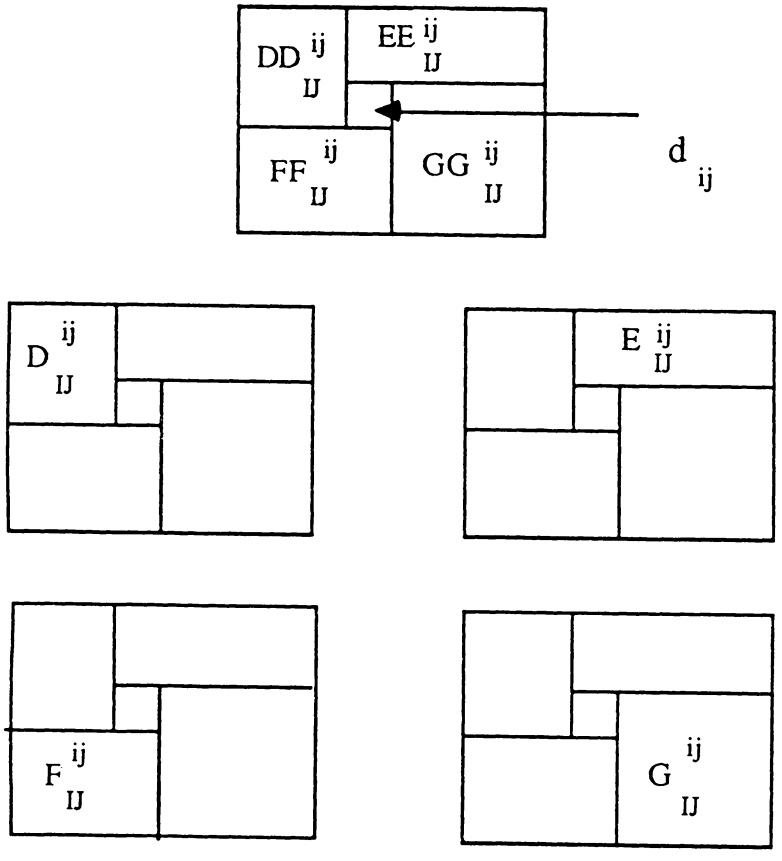


Figure 3.5 Notation for a four pass recursive algorithm. $d_{ij}$ : an observed measurement vector from pixel ( i,j ); $D_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left of (i,j) or to the left and above (i,j), including (i,j). $E_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right of (i,j) or to the right and above (i,j), including (i,j). $F_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the left of (i,j) or to the left and below (i,j), including (i,j). $G_{IJ}^{ij}$ : set of all observed data values $d_{lk}$ for (l,k) to the right of (i,j) or to the right and below (i,j), including (i,j).

**Assumptions for four pass algorith:**

a) We assume that an n-tuple of measurements is determined by some local measuring process. Given the true category of a pixel, all measurements of all pixels are independent of each other. In effect this says that given the true state of affairs, the observed measurement variations are independent. Hence,

$$P( D_{IJ} \mid C_{ij} ) = P( DD_{IJ}^{ij} \mid C_{ij} )P( EE_{IJ}^{ij} \mid C_{ij} )P( FF_{IJ}^{ij} \mid C_{ij} )P( GG_{IJ}^{ij} \mid C_{ij} )$$

$$P( D_{IJ}^{ij} \mid C_{ij} ) = P( DD_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )$$

$$P( E_{IJ}^{ij} \mid C_{ij} ) = P( EE_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )$$

$$P( F_{IJ}^{ij} \mid C_{ij} ) = P( FF_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )$$

$$P( G_{IJ}^{ij} \mid C_{ij} ) = P( GG_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )$$

b) the n-tuple measurement of pixel ( i,j ) depends only upon the true interpretation associated with pixel ( i,j ) and does not depend upon any relationship pixel ( i,j ) may have with other units or upon the interpretation associated with any other unit.

$$P( d_{ij} \mid C_{ij} , D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1} ) = P( d_{ij} \mid C_{ij} )$$

$$P( d_{ij} \mid C_{ij} , E_{IJ}^{i-1j} , d_{i-1j+1} , \ldots , d_{i-1N} ) = P( d_{ij} \mid C_{ij} )$$

$$P( d_{ij} \mid C_{ij} , F_{IJ}^{i+1j} , d_{i+1,1} , \ldots , d_{i+1j-1} ) = P( d_{ij} \mid C_{ij} )$$

$$P( d_{ij} \mid C_{ij} , G_{IJ}^{ij+1} , d_{i+1j+1} , \ldots , d_{Nj+1} ) = P( d_{ij} \mid C_{ij} )$$

c) given the true categories of two adjacent neighboring pixels, the observed measurement data for all pixels before the current pixel, in raster scan order, tells nothing more about the current pixel's category.

$$P( C_{ij} \mid C_{ij-1}, C_{i-1j} , D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1} ) = P( C_{ij} \mid C_{ij-1} , C_{i-1j} )$$

$$P( C_{ij} \mid C_{ij+1}, C_{i-1j} , E_{IJ}^{i-1j} , d_{i-1j+1} , \ldots , d_{i-1N} ) = P( C_{ij} \mid C_{ij+1} , C_{i-1j} )$$

$$P( C_{ij} \mid C_{i+1j}, C_{ij-1} , F_{IJ}^{i+1j} , d_{i+1,1} , \ldots , d_{i+1j-1} ) = P( C_{ij} \mid C_{i+1j} , C_{ij-1} )$$

$$P( C_{ij} \mid C_{i+1j}, C_{i+1j} , G_{IJ}^{ij+1} , d_{i+1j+1} , \ldots , d_{Ij+1} ) = P( C_{ij} \mid C_{i+1j} , C_{ij+1} )$$

d) given the measurements of all pixels up to and including the current pixel, the true category of diagonally adjacent neighbors tells nothing more.

$$P( C_{ij-1} \mid C_{i-1j} , D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1} )$$

$$= P( C_{ij-1} \mid D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1} )$$

$$P( C_{i-1j} \mid C_{ij+1} , E_{IJ}^{i-1j} , d_{i-1j+1} , \ldots , d_{i-1N} )$$

$$= P( C_{i-1j} \mid E_{IJ}^{i-1j} , d_{i-1j+1} , \ldots , d_{i-1N} )$$

$$P( C_{i+1j} \mid C_{i,j-1} , F_{IJ}^{i+1j} , d_{i+1,1} , \ldots , d_{i+1j-1} )$$

$$= P( C_{i+1j} \mid F_{IJ}^{i+1,j} , d_{i+1,1} , \ldots , d_{i+1j-1} )$$

$$P( C_{ij+1} \mid C_{i+1j} , G_{IJ}^{ij+1} , d_{i+1j+1} , \ldots , d_{Ij+1} )$$

$$= P( C_{ij+1} \mid G_{IJ}^{ij+1} , d_{i+1j+1} , \ldots , d_{Ij+1} )$$

e) approximation: (omitting a column of data values should not make a significant a difference)

$$P( C_{ij-1} \mid D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1} ) = P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$P( C_{i-1j} \mid E_{IJ}^{i-1j} , d_{i-1j+1} , \ldots , d_{i-1N} ) = P( C_{i-1j} \mid E_{IJ}^{i-1j} )$$

$$P( C_{i+1j} \mid F_{IJ}^{i+1j} , d_{i+1,1} , \ldots , d_{i+1j-1} ) = P( C_{i+1j} \mid F_{IJ}^{i+1,1} )$$

$$P( C_{ij+1} \mid G_{IJ}^{ij+1} , d_{i+1j+1} , \ldots , d_{Ij+1} ) = P( C_{ij+1} \mid G_{IJ}^{ij+1} )$$

f) A weaker assumption than the independence of neighboring categories is the Markov Random Field assumption, which is used when the true interpretation of any unit given the true interpretations of all the surrounding depends only upon the interpretation of the nearest neighboring pixel.

In our four neighbor system, given the true interpretation of pixel ( i,j ), the categories of diagonal pixels, which are not nearest neighboring pixels of pixel (i,j), are independent of each other.

$$P( C_{ij-1} , C_{i-1j} \mid C_{ij} ) = P( C_{ij-1} \mid C_{ij} )P( C_{i-1j} \mid C_{ij} )$$

$$P( C_{ij+1} , C_{i+1j} \mid C_{ij} ) = P( C_{ij+1} \mid C_{ij} )P( C_{i+1j} \mid C_{ij} )$$

g) given the true category of a diagonally adjacent pixel in the nearest neighborhood, the observed measurement data for all pixels before the current pixel, in raster scan order, tells nothing more about the current pixel's category.

$$P( C_{i-1j} \mid C_{ij-1} , D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1} ) = P( C_{i-1j} \mid C_{ij-1} )$$

$$P( C_{ij+1} \mid C_{i-1j} , E_{IJ}^{i-1j} , d_{i-1j+1} , \ldots , d_{i-1N} ) = P( C_{ij+1} \mid C_{i-1j} )$$

$$P( C_{i+1j} \mid C_{ij-1} , F_{IJ}^{i+1j} , d_{i+1,1} , \ldots , d_{i+1j-1} ) = P( C_{i+1j} \mid C_{ij-1} )$$

$$P( C_{ij+1} \mid C_{i+1j} , G_{IJ}^{ij+1} , d_{i+1j+1} , \ldots , d_{Ij+1} ) = P( C_{ij+1} \mid C_{i+1j} )$$

To compute the conditional probability $P( C_{ij} \mid D_{ij} )$, we can argue as follows:

$$P( C_{ij} \mid D_{IJ} )$$

$$= \frac{P( DD_{IJ}^{ij} , EE_{IJ}^{ij} , FF_{IJ}^{ij} , GG_{IJ}^{ij} , d_{ij} \mid C_{ij} )P( C_{ij} )}{P( D_{IJ})}$$

$$= P( DD_{IJ}^{ij} \mid C_{ij} )P( EE_{IJ}^{ij} \mid C_{ij} )P( FF_{IJ}^{ij} \mid C_{ij} )P( GG_{IJ}^{ij} \mid C_{ij} )$$

$$* \frac{P( d_{ij} \mid C_{ij} )P( C_{ij} )}{P(D_{IJ})}$$

$$= [ P( DD_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )][ P( EE_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )]$$

$$*\frac{[ P( FF_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )][ P( GG_{IJ}^{ij} \mid C_{ij} )P( d_{ij} \mid C_{ij} )]P( C_{ij} )}{P( D_{IJ} ) P( d_{ij} \mid C_{ij})^3}$$

$$= \frac{P( D_{IJ}^{ij} \mid C_{ij} )P( E_{IJ}^{ij} \mid C_{ij} )P( F_{IJ}^{ij} \mid C_{ij} )P( G_{IJ}^{ij} \mid C_{ij} )P( C_{ij} )}{P( D_{IJ})P( d_{ij} \mid C_{ij} )^3}$$

$$= \frac{\frac{P( C_{ij} \mid D_{IJ}^{ij} )P( D_{IJ}^{ij} )}{P( C_{ij} )} * \frac{P( C_{ij} \mid E_{IJ}^{ij} )P( E_{IJ}^{ij} )}{P( C_{ij} )} * \frac{P( C_{ij} \mid F_{IJ}^{ij} )P( F_{IJ}^{ij} )}{P( C_{ij} )}}{P( D_{IJ})}$$

$$*\frac{P( C_{ij} \mid G_{IJ}^{ij} )P( G_{IJ}^{ij} )}{P( C_{ij} )}* \frac{P( C_{ij} )}{P(D_{IJ} )*P( d_{ij} \mid C_{ij} )^3}$$

$$= \frac{P( C_{ij} \mid D_{IJ}^{ij} )P( C_{ij} \mid E_{IJ}^{ij} )P( C_{ij} \mid F_{IJ}^{ij} )P( C_{ij} \mid G_{IJ}^{ij} )}{\frac{P( D_{IJ} )P( d_{ij} \mid C_{ij} )^3 P( C_{ij} )^3}{P( D_{IJ}^{ij} )P( E_{IJ}^{ij} )P( F_{IJ}^{ij} )P( G_{IJ}^{ij} )}}$$

$$= \frac{P( C_{ij} \mid D_{IJ}^{ij} )P( C_{ij} \mid E_{IJ}^{ij} )P( C_{ij} \mid F_{IJ}^{ij} )P( C_{ij} \mid G_{IJ}^{ij} )}{\Delta * P( d_{ij} \mid C_{ij} )^3 P( C_{ij} )^3} \qquad (3.6)$$

and $\Delta$ is a normalizing constant which makes the sum over all $C_{ij}$ be unity.

In (3.6) $P( C_{ij} )$ and $P( d_{ij} \mid C_{ij} )$ are known or estimated in the supervised remote sensing classification situation, and $\Delta$ does not depend on the category label for $C_{ij}$. Then to solve (3.6) the problem becomes to calculate

$P( C_{ij} \mid D_{IJ}^{ij} )$, $P( C_{ij} \mid E_{IJ}^{ij} )$, $P( C_{ij} \mid F_{IJ}^{ij} )$ and $P( C_{ij} \mid G_{IJ}^{ij} )$.

We will give the derivation for the recursive calculation of $P( C_{ij} \mid D_{IJ}^{ij} )$. The other three are same as that one, except for the change in the scan directions.

For conditional probability $P( C_{ij} \mid D_{IJ}^{ij} )$, we have

$$P( C_{ij} \mid D_{IJ}^{ij} ) = \frac{P( C_{ij} , D_{IJ}^{ij-1}, d_{ij} , d_{i-1j-1} , \ldots , d_{1j-1} )}{P( D_{IJ}^{ij} )}$$

$$= \frac{P( d_{ij} \mid C_{ij} , D_{IJ}^{ij-1}, d_{i-1j-1} , \ldots , d_{1j-1} ) P( C_{ij} , D_{IJ}^{ij-1}, d_{i-1j-1} , \ldots , d_{1j-1})}{P( D_{IJ}^{ij} )}$$

using assumption (b):

$$= \frac{P( d_{ij} \mid C_{ij} ) P( C_{ij} , D_{IJ}^{ij-1}, d_{i-1j-1} , \ldots , d_{1j-1} )}{P( D_{IJ}^{ij} )}$$

$$= \frac{[ \sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} , C_{i-1j} , C_{ij-1} , D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1} )] P( d_{ij} \mid C_{ij} )}{P( D_{IJ}^{ij} )}$$

$$= [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P(\, C_{ij} \mid C_{i-1j}\,, C_{ij-1}\,, D_{IJ}^{ij-1}\,, d_{i-1j-1}\,, \ldots,\, d_{1j-1})$$

$$* \frac{P(\, C_{ij-1}\,, C_{i-1j}\,, D_{IJ}^{ij-1}\,, d_{i-1j-1}\,, \ldots,\, d_{1j-1})]P(\, d_{ij} \mid C_{ij}\,)}{P(\, D_{IJ}^{ij}\,)}$$

using assumption (c)

$$= [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P(\, C_{ij} \mid C_{i-1j}\,, C_{ij-1}\,)P(\, C_{ij-1} \mid C_{i-1j}\,, D_{IJ}^{ij-1}\,, d_{i-1j-1}\,, .., d_{1j-1})$$

$$* \frac{P(\, C_{i-1j}\,, D_{IJ}^{ij-1}\,, d_{i-1j-1}\,, \ldots,\, d_{1j-1})]P(\, d_{ij} \mid C_{ij}\,)}{P(\, D_{IJ}^{ij}\,)}$$

using assumption (d) and (e) :

$$= [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P(\, C_{ij} \mid C_{i-1j}\,, C_{ij-1}\,)P(\, C_{ij-1} \mid D_{IJ}^{ij-1}\,)$$

$$* \frac{\sum_{K_{ij-1}} P(\, K_{ij-1}\,, C_{i-1j}\,, D_{IJ}^{ij-1}\,, d_{i-1j-1}\,, \ldots,\, d_{1j-1})]P(\, d_{ij} \mid C_{ij}\,)}{P(\, D_{IJ}^{ij}\,)}$$

$$= [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P(\, C_{ij} \mid C_{i-1j}\,, C_{ij-1}\,)P(\, C_{ij-1} \mid D_{IJ}^{ij-1}\,)$$

$$* \sum_{K_{ij-1}} P(\, C_{i-1j} \mid K_{ij-1}\,, D_{IJ}^{ij-1}\,, d_{i-1j-1}\,, \ldots,\, d_{1j-1})$$

$$* \frac{P(\, K_{ij-1}\,, D_{IJ}^{ij-1}\,, d_{i-1j-1}\,, \ldots,\, d_{1j-1})P(\, d_{ij} \mid C_{ij}\,)}{P(\, D_{IJ}^{ij}\,)}$$

using assumption (g) :

$$= \sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} )P( C_{ij-1} \mid D_{IJ}^{ij-1} ) \sum_{K_{ij-1}} P( C_{i-1j} \mid K_{ij-1} )$$

$$* P( K_{ij-1} \mid D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1})P( D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1})$$

$$* \frac{P( d_{ij} \mid C_{ij} )}{P( D_{IJ}^{ij} )}$$

approximation (e) :

$$= \sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} )P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \sum_{K_{ij-1}} P( C_{i-1j} \mid K_{ij-1} )P( K_{ij-1} \mid D_{IJ}^{ij-1} )P( D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1})$$

$$* \frac{P( d_{ij} \mid C_{ij} )}{P( D_{IJ}^{ij} )}$$

$$= [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} )P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \frac{\sum_{K_{ij-1}} P( C_{i-1j} \mid K_{ij-1} )P( K_{ij-1} \mid D_{IJ}^{ij-1} )]P( d_{ij} \mid C_{ij} )}{\frac{P( D_{IJ}^{ij} )}{P( D_{IJ}^{ij-1} , d_{i-1j-1} , \ldots , d_{1j-1})}}$$

finally we have

$$P( C_{ij} \mid D_{IJ}^{ij} )$$

$$= \sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} )P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \frac{\sum_{K_{ij-1}} P( C_{i-1j} \mid K_{ij-1} )P( K_{ij-1} \mid D_{IJ}^{ij-1} )P( d_{ij} \mid C_{ij} )}{\Delta_D}$$

where

$$\Delta_D = \sum_{k_{ij-1}} \sum_{k_{i-1j}} P( k_{ij} \mid k_{i-1j} , k_{ij-1} )P( k_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \sum_{k'_{ij-1}} P( k_{i-1j} \mid k'_{ij-1} )P( k'_{ij-1} \mid D_{IJ}^{ij-1} )P( d_{ij} \mid k_{ij} ) \qquad (3.7)$$

which is normalizing constant, not dependent on true category.

Using a similar derivation the resulting recursive formulas for $P( C_{ij} \mid E_{IJ}^{ij} )$, $P( C_{ij} \mid F_{IJ}^{ij} )$ and $P( C_{ij} \mid G_{IJ}^{ij} )$ are

$$P( C_{ij} \mid E_{IJ}^{ij} )$$

$$= \sum_{C_{ij+1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij+1} )P( C_{i-1j} \mid E_{IJ}^{i-1j} )$$

$$* \frac{\sum_{K_{i-1j}} P( C_{ij+1} \mid K_{i-1j} )P( K_{i-1j} \mid E_{IJ}^{i-1j} )P( d_{ij} \mid C_{ij} )}{\Delta_E}$$

where

$$\Delta_E = \sum_{k_{ij+1}} \sum_{k_{i-1j}} P( k_{ij} \mid k_{i-1j} , k_{ij+1} )P( k_{i-1j} \mid E_{IJ}^{i-1j} )$$

$$* \sum_{k'_{i-1j}} P( k_{ij+1} \mid k'_{i-1j} )P( k'_{i-1j} \mid E_{IJ}^{i-1j} )P( d_{ij} \mid k_{ij} ) \qquad (3.8)$$

$$P( C_{ij} \mid F_{IJ}^{ij} )$$

$$= \sum_{C_{ij-1}} \sum_{C_{i+1j}} P( C_{ij} \mid C_{i+1j} , C_{ij-1} )P( C_{i+1j} \mid F_{IJ}^{i+1j} )$$

$$* \frac{\sum_{K_{ij+1}} P( C_{ij-1} \mid K_{ij+1} )P( K_{ij+1} \mid F_{IJ}^{ij+1} )P( d_{ij} \mid C_{ij} )}{\Delta_F}$$

where

$$\Delta_F = \sum_{k_{ij-1}} \sum_{k_{i+1j}} P( k_{ij} \mid k_{i+1j} , k_{ij-1} )P( k_{i+1j} \mid F_{IJ}^{i+1j} )$$

$$* \sum_{k'_{ij+1}} P( k_{ij-1} \mid k'_{ij+1} )P( k'_{ij+1} \mid F_{IJ}^{i+1j} )P( d_{ij} \mid k_{ij} ) \qquad (3.9)$$

$$P( C_{ij} \mid G_{IJ}^{ij} )$$

$$= \sum_{C_{ij+1}} \sum_{C_{i+1j}} P( C_{ij} \mid C_{i+1j} , C_{ij+1} )P( C_{ij+1} \mid G_{IJ}^{ij+1} )$$

$$* \frac{\sum_{K_{ij+1}} P( C_{i+1j} \mid K_{ij+1} )P( K_{ij+1} \mid G_{IJ}^{ij+1} )P( d_{ij} \mid C_{ij} )}{\Delta_G}$$

where

$$\Delta_G = \sum_{k_{ij+1}} \sum_{k_{i+1j}} P( k_{ij} \mid k_{i+1j} , k_{ij+1} )P( k_{i+1j} \mid G_{IJ}^{ij+1} )$$

$$* \sum_{k'_{ij+1}} P( k_{i+1j} \mid k'_{ij+1} )P( k'_{ij+1} \mid G_{IJ}^{ij+1} )P( d_{ij} \mid k_{ij} ) \qquad (3.10)$$

The recursive formulas (3.7), (3.8), (3.9), and (3.10) use a left-right top-bottom scan, right-left top-bottom scan, left-right bottom-top scan and right-left bottom-top scan , respectively.

**(III-3.3) No look-ahead algorithm:**

For this algorithm the context dependent processing makes the assignment of the best category label only using the set of all observed data values $d_{lk}$, for ( l,k ) to the left of or above ( i,j ), including ( i,j ) (i.e. $D_{IJ}^{ij}$ ) (see figure 3.4). To determine the Bayes labeling under the Markov Field assumption we determine $C_{ij}$ which maximizes the conditional probability $P( C_{ij} \mid D_{IJ}^{ij} )$.

For conditional probability $P( C_{ij} \mid D_{IJ}^{ij} )$, we have

$$P( C_{ij} \mid D_{IJ}^{ij} ) = P( C_{ij} \mid D_{IJ}^{ij-1}, d_{ij} )$$

$$= \frac{P(C_{ij}, d_{ij} \mid D_{IJ}^{ij-1} )P( D_{IJ}^{ij-1} )}{P( D_{IJ}^{ij} )}$$

$$= \frac{P( d_{ij} \mid C_{ij} , D_{IJ}^{ij-1} )P( C_{ij} , D_{IJ}^{ij-1} )}{P( D_{IJ}^{ij} )}$$

using assumption b :

$$= \frac{P( d_{ij} \mid C_{ij} )P( C_{ij} , D_{IJ}^{ij-1} )}{P( D_{IJ}^{ij} )}$$

and using similar derivation as the section (III-3.4.a) we can get:

$$P(C_{ij} \mid D_{IJ}^{ij} ) = [\sum_{C_{ij-1}} \sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} ) P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \frac{P( C_{i-1j} \mid D_{IJ}^{i-1j} )]*P( d_{ij} \mid C_{ij} )}{\Delta} \qquad (3.11)$$

where

$$\Delta = \sum_{k_{ij}} P( d_{ij} \mid k_{ij} ) \sum_{k_{ij-1}} \sum_{k_{i-1j}} P( k_{ij} \mid k_{i-1j} , k_{ij-1} )$$

$$* P( k_{ij-1} \mid D_{IJ}^{ij-1} ) P( k_{i-1j} \mid D_{IJ}^{i-1j} ) \qquad (3.12)$$

The probability distribution $P(C_{ij})$ and $P(d_{ij} \mid C_{ij})$ are either known a priori, or estimated in the case of supervised remote sensing classification, and the denominator is some constant not dependent upon the true category. Obviously, this is special case of equation ( 3.1 ). There is only one recursive item $P( C_{ij} \mid D_{IJ}^{ij} )$ instead of two recursive items $P( C_{ij} \mid D_{IJ}^{ij} )$ $P( C_{ij} \mid E_{IJ}^{ij} )$ The figure 3.4 shows the algorithm, scanning from top-left to bottom-right only. The recursive formula for $P( C_{ij} \mid D_{IJ}^{ij} )$ is given by (3.11).

Finally, application of the Bayes decision rule requires calculation of the formulas (3.11) and determination of the class $C_{ij}$ which produces the maximum probability $P( C_{ij} \mid D_{IJ}^{ij} )$. In effect, this is a one pass scanning algorithm.

**(III-3.4) Fixed step look-ahead algorithm:**

This algorithm uses some fixed n-step look-ahead. The context dependent processing makes the assignment of the best category label using $D_{IJ}^{Ij}$

and the (n+1) x (n+1) neighborhood of measurements centered at pixel ( i,j ). In this section, we specifically adopt a simple one-step look-ahead case.

For this algorithm the context dependent processing makes the assignment of the best category label only using $D_{IJ}^{ij}$ and the nearest neighboring pixels $d_{ij+1}$ ,$d_{i+1j-1}$ ,$d_{i+1j}$ ,$d_{i+1j+1}$ ... . To determine the Bayes labeling under the Markov Field assumption we determine that $C_{ij}$ maximizing the conditional probability $P( C_{ij} \mid D_{IJ}^{ij}, d_{ij+1}, d_{i+1j-1}, d_{i+1j}, d_{i+1j+1} )$

The implementation of the decision rule requires the calculation of the probability.

$$P( C_{ij} \mid D_{IJ}^{ij} , d_{ij+1} , d_{i+1j-1} , d_{i+1j} , d_{i+1j+1} )$$

$$= \frac{P( D_{IJ}^{ij} , d_{ij+1} , d_{i+1j-1} , d_{i+1j} , d_{i+1j+1} \mid C_{ij} )P( C_{ij} )}{P( D_{IJ} , d_{ij+1} , d_{i+1j-1} , d_{i+1j} , d_{i+1j+1} )}$$

using assumption (a) :

$$= \frac{P( D_{IJ}^{ij} \mid C_{ij} )P( d_{ij+1} , d_{i+1j-1} , d_{i+1j} , d_{i+1j+1} \mid C_{ij} )P(C_{ij})}{P( D_{IJ} , d_{ij+1} , d_{i+1j-1} , d_{i+1j} , d_{i+1j+1} )} \quad (3.13)$$

From assumption (a) we have :

$$P( d_{ij+1} , d_{i+1j-1} , d_{i+1j} , d_{i+1j+1} \mid C_{ij} ) = \prod_{l} \prod_{k} P( d_{lk} \mid C_{ij} ) \quad (3.14)$$

where (l,k) belongs to the set of nearest neighboring pixels following pixel

(i,j) in the raster scan order. Finally, (3.13) can be expressed as follows:

$$= \frac{P( D_{IJ}^{ij} \mid C_{ij} )P( C_{ij} )\prod_{l}\prod_{k} P( d_{lk} \mid C_{ij} )}{P( D_{IJ} , d_{ij+1} , d_{i+1j-1} , d_{i+1j} , d_{i+1j+1} )}$$

$$= \frac{\frac{P( C_{ij} \mid D_{IJ}^{ij} )P(D_{IJ}^{ij} )}{P( C_{ij} )}P( C_{ij} )\prod_{l}\prod_{k} P( d_{lk} \mid C_{ij} )}{P( D_{IJ} , d_{ij+1} , d_{i+1j-1} , d_{i+1j} , d_{i+1j+1} )}$$

$$= \frac{P( C_{ij} \mid D_{IJ}^{ij} )\prod_{l}\prod_{k} P( d_{lk} \mid C_{ij} )}{\frac{P( D_{IJ} , d_{ij+1} , d_{i+1j-1} , d_{i+1j} , d_{i+1j+1} )}{P( D_{IJ}^{ij} )}} \qquad (3.15)$$

In equation (3.15) the denominator is some constant not dependant on the optimizing category label. For determining the maximizing $C_{ij}$ the problem becomes one calculating: $P( C_{ij} \mid D_{IJ}^{ij} )$ and $\prod_{l}\prod_{k} P(d_{lk} \mid C_{ij} )$. For the conditional probability $P( C_{ij} \mid D_{IJ}^{ij} )$ we have from equation (3.3):

$$P(C_{ij} \mid D_{IJ}^{ij} ) = [\sum_{C_{ij-1}}\sum_{C_{i-1j}} P( C_{ij} \mid C_{i-1j} , C_{ij-1} ) P( C_{ij-1} \mid D_{IJ}^{ij-1} )$$

$$* \frac{P( C_{i-1j} \mid D_{IJ}^{i-1j} )]*P( d_{ij} \mid C_{ij} )}{\Delta}$$

where

$$\Delta = \sum_{k_{ij}} P( d_{ij} \mid k_{ij} ) \sum_{k_{ij-1}}\sum_{k_{i-1j}} P( k_{ij} \mid k_{i-1j} , k_{ij-1} )$$

$$* P( k_{ij-1} \mid D_{IJ}^{ij-1} )P( k_{i-1j} \mid D_{IJ}^{i-1j} )$$

Now we give the derivation of $P( d_{ij+1} \mid C_{ij} )$; the other three are similar.

$$P( d_{ij+1} \mid C_{ij} ) = \sum_{C_{ij+1}} P( d_{ij+1} , C_{ij+1} \mid C_{ij} )$$

$$= \frac{\sum_{C_{ij+1}} P( C_{ij} \mid d_{ij+1} , C_{ij+1} )P( d_{ij+1} , C_{ij+1} )}{P( C_{ij} )}$$

using assumption (c)

$$= \frac{\sum_{C_{ij+1}} P( C_{ij} \mid C_{ij+1} )P( d_{ij+1} \mid C_{ij+1} )P( C_{ij} )}{P( C_{ij} )}$$

The probability of each class $P( C_{ij} )$ or $P( C_{ij+1} )$ can be directly estimated from the preclassification results, and the transition probabilities $P(C_{ij} \mid C_{ij+1} )$, $P( C_{ij} \mid C_{i+1j+1} )$ can also be estimated from preclassification results. The conditional probability $P(d_{ij+1} \mid C_{ij+1} )$ is either known or estimated in the supervised remote sensing classification situation.

### (III-3.5) Computation requirement :

Because the estimation equations of the recursive algorithm are quite simple, the computation required increases linearly with the number of pixels. Let K be the number of categories and N the number of pixels. In the two pass algorithm, the evaluation of $P( C_{ij} \mid D_{IJ}^{ij} )$ and $P( C_{ij} \mid E_{IJ}^{ij} )$ requires $2K^2N$ multiplications, and $P( C_{ij} \mid D_{IJ} )$ requires KN multiplications. So, the multiplications total (4K+1)KN. Similarly, in the four pass algorithm, the computation complexity is (8K+1)KN. When we implement the recursive equation for each pixel, it requires only adjacent pixels previously processed, and the measurement dependency is entirely included in $P( d_{ij} \mid C_{ij} )$. So in this algorithm the internal memory required grows linearly with the number of columns in the image. Because of the above two facts, the method is particularly suitable for large sized images.

## (III-4) Summary of the recursive contextual classification procedure

Before presenting some experimental results with the proposed classification procedure, we shall first summarize the steps necessary to perform such a classification where the conditional distributions are assumed normal.

(1) Evaluate training statistics. This includes the mean vector and covariance matrices of the Gaussian class conditional distributions.

(2) Preclassify the image using a pixel independent or context free Bayes classification technique.

(3) Estimate the transition probabilities $P( C_{ij} \mid C_{ij-1}, C_{i-1j} )$ from the preclassification results.

(4) For the full look-ahead algorithm use recursive formula 3.5 to compute $P( C_{ij} \mid D_{IJ}^{ij} )$ and $P( C_{ij} \mid E_{IJ}^{ij} )$ (the two pass algorithm) or use 3.7-3.10 to compute $P( C_{ij} \mid D_{IJ}^{ij} )$, $P( C_{ij} \mid E_{IJ}^{ij} )$, $P( C_{ij} \mid F_{IJ}^{ij} )$ and $P( C_{ij} \mid G_{IJ}^{ij} )$ (for the four pass algorithm).

Similarly for the no look-ahead algorithm and the fixed look-ahead algorithm, we use equation 3.12 and 3.15, respectively.

(5) For the full lookahead algorithm, we use equation 3.1 to evaluate $P( C_{ij} \mid D_{IJ} )$ for each category (the two pass algorithm), or use equation 3.6 to compute it (the four pass algorithm). Similarly, for the no look-ahead algorithm and the fixed look-ahead algorithm, we use equation 3.11 and 3.15, respectively.

(6) Use the classification rule to select that label $C_{ij}$ which produces the maximum posterior probability of the category label given the context the algorithm uses.

**(III-5) Related literature:**

The approaches to the context classification which we review here were given by Tilton et al. (1982) and Yu (1983). Tilton in his paper " Estimation of Context for Statistical Classification of Multispectral Image Data " uses an estimation method of the context function, which provides a 2-6% improvement in classification accuracy ; in another paper " Context classification of multispectral image data", he uses an eight-nearest-neighbor contextual classification which is a generalization of the familiar maximum likelihood classifier. It again provides a 6.5% improvement in average- by-class accuracy. Yu uses a coding technique, by which the spatial correlation parameter can be estimated. His contextual classification provides about 5% improvement in the first-stage and 2% additional improvement in the second-stage.

Tilton's decision rule assigns $C_{ij}$ to pixel (i,j) if $C_{ij}$ maximizes

$$[\sum_{C^p} G(C^p) \prod_{(l,k) \in N_p} f(d_{lk} \mid C_{lk})] f(d_{ij} \mid C_{ij})$$

where $N_p$ is p-context array contains spatial information of (p-1) pixels surrounding and neighboring the pixel (i,j), and $C^p$ is the p-vectors of spectral classes; $G(C^p)$, the 'context function', is the relative frequency with which $C^p$ occurs in the scene being analyzed. Tilton mentions that the computational complexity of his contextual classification method is proportional to $k^p$, where k is the number of classes and the context array (including the pixel to be classified ) has p cells. The computation is expensive,

therfore only 50 X 50 size images are tested in his paper.

The classification rule in Yu's paper (1983) is to select that class $C_k$ which, after normalizing pixel k, will produce the maximum joint probability $P( D_k , D_{k1} , D_{k2} , D_{k3} , D_{k4} \mid C_k )$, where $D_k , D_{k1} , D_{k2} , D_{k3} , D_{k4}$ are four-adjacent pixels of pixel $D_k$. In this method he uses a coding technique to estimate the spatial correlation for each pair; for each iteration it has to evaluate the joint normal density function and use a maximum likelihood decision rule for each pixel once, so the computation cost is relatively high.

Compared with the above two papers, the main advantage of our approach is its simplicity and low cost in computation. The computational complexity of our approach is proportional to $4K^2N$ instead of $NK^p$ in Tilton's method, in that N is the number of pixels, K is the number of categories, and p is the size of context array (ie. p = 2,3,4,5... ). So the method proposed in the chapter is particularly suitable for larger sized images. In addition, this algorithm has the advantage over other techniques in that it handles multiple observation data naturally and simultaneously. The experimental results are shown in chapter V, which indicate the classification accuracy improvement of our approach to be about the same in comparison with the above two methods.

# CHAPTER IV: CONTEXTUAL CLASSIFICATION BY STOCHASTIC RELAXATION

There appears to be two main approaches to the specification of spatial stochastic processes. Whittle (1963) proposed a random field model which arises from the joint probability distribution of the variables in a neighborhood. Whittle's definition requires that the joint probability distribution of the variables in a given neighborhood be of the product form

$$P( D_{ij} ) = \prod_{ij} Q_{ij} ( D_{ij} ) \tag{4.1}$$

where $D_{ij}$ is a set of random variables within the neighborhood of the pixel ( i,j ) and $Q_{ij}$ is a nonnegative function.

On the other hand, Bartlett (1955, 1967, 1968) proposed a model which arises from the conditional probability distribution of $D_{ij}$ . His definition requires that the conditional probability distribution of $D_{ij}$ depends only upon the values at the neighbors of ( i,j ).

The contextual classification method we discussed in Chapter III is based on the model which arises from the conditional probability distribution of $D_{ij}$ . Most of the existing methods (T. S. Yu and K. S. Fu 1983 , P. H. Swain and S. B. Vardeman 1981) also belong to the later one.

Besay (1974) found that constraints on the conditional probability

structure are so severe that they actually dictate particular models.

Under the stochastic model for spatially oriented pixels image correlations exist between any pair of pixels. Fu (1980) noted that the best way to incorporate these correlations statistically is to consider the joint probability density function of all the site-variables involved. For example, a five dimensional joint probability density function will account for all the correlation between any pair of sites in a 4-neighbor system. Similarly a nine dimensional joint density function can be used for the eight-neighbor system. For practical reasons most of the estimated context decision rules deal with maximizing the conditional probability distribution of pixel ( i,j ), given all values within the nearest neighborhood of pixel ( i,j ).

Besay (1974) argues that the conditional probability model has a number of disadvantages. First, there is no obvious method of deducing the joint probability structure associated with a conditional probability model. Secondly, the conditional probability structure itself is subject to some unapparent and highly restrictive consistency conditions. Third, it has been remarked by Whittle (1963) that the natural specification of an equilibrium process in statistical mechanics is in terms of the joint distribution rather than the conditional distribution of their variables. Similarly, the most natural and important quantity in evaluation of the discriminate function in contextual classification is the joint probability density function (Fu

1980). The conditional probability approach, however, has served as the basis for a commonly used class of models - the Markov image models.

In this chapter we develop an alternative context classification approach, which is based on a stochastic relaxation algorithm and Gibbs distribution. The favorable features of this approach are that its random model arises from the joint probability distribution of the variates in a neighborhood, and that the algorithm is highly parallel. The parallel algorithm, which is performed by a neighborhood operator, might be implementable in special purpose VLSI hardware.

The stochastic relaxation methods are not new concepts, as they have been used in statistical physics for many years. There, the problem of analyzing the macroscopic properties of a physical system is translated into one of analyzing the global properties of random fields with a given local structure. However, only Geman (1985) has introduced these concepts into image restoration, and he has given a few simple results using synthetic images. In this chapter we use them in context classification, and we make an analogy between image and statistical mechanics systems. Pixel gray levels and category labels are viewed as states of atoms or molecules in a lattice-like physical system. The assignment of energy functions in a physical system determines its Gibbs distribution. Because of the Gibbs distribution-Markov Random Field (MRF) equivalence, this assignment also

determines a MRF image model.

In this chapter we first motivate a Bayesian context decision rule, and then we use a Markov-Gibbs model to develop a new contextual classification algorithm, in which maximizing the posteriori probability (MAP) is based on stochastic relaxation, an annealing optimization method. Finally, we present experimental results with both simulated and real multispectral remote sensing data to show how classification accuracy is greatly improved.

## (IV-1) MOTIVATION AND PROPOSED APPROACH

The contextual information, which we would like to study, is a form of correlation existing among the successive pattern classes in the two-dimensional image. Every pixel in the image can be considered as having one random variable associated with a 2-D Markov random field. Two pixels in spatial proximity to one another are unconditionally correlated, with the degree of correlation decreasing as the distance between them increases. All the spatial correlations among "site-variables" on a lattice of image can be extracted by the specified spatial process. As mentioned earlier, the most important quantity in the contextual Bayes' decision problem is the joint density function of all the site-variables within the specified contextual neighborhood. So the best way to incorporate these correlations statistically

is to estimate the joint probability density function of all the site-variables involved.

For practical reasons, most of previous studies deal with some specific cases (in which the 4-neighborhood assumption is invariably utilized in the context algorithms) and the contextual information is incorporated by considering the conditional probabilities of pixel ( i,j ), given its neighbors.

These approaches are certainly based on a realistic premise but it is computationally feasible only for first order neighborhoods. The new contextual decision rule introduced in this section improves this by considering the joint probability of the pixels in the neighborhood of the pixel ( i,j ) and by using a larger context.

Before presenting this rule, we must first give some notational conventions, and assume that each pixel of the multiband image considered in the chapter has a N-tuple of finite gray tone values.

**(IV-1.1) Notation:**

In order to have a precise framework within which we can describe the stochastic relaxation algorithm, we need some notational conventions.

(1) $N_{ij}$ : designates a neighborhood of pixel ( i,j ).

(2) $D_{ij}$ : the collection of all measurement vectors in the neighborhood $N_{ij}$ of pixel ( i,j ).

(3) C : assigned category labels in the neighborhood.

(4) $C^o$ : assigned category labels in the neighborhood, excluding the central pixel of the neighborhood.

(5) $2N_r + 1$ : designates the number of rows in a neighborhood.

(6) $2N_c + 1$ : designates the number of columns in a neighborhood.

(7) $L_r$ : designates the local row index set of the neighborhood.

$$L_r = \{ -N_r , ....., N_r \}$$

(8) $K_c$ : designates the local column index set of the neighborhood.

$$K_c = \{ -N_c, ....., N_c \}$$

(9) (l,k) : designates a local position in neighborhood.

$$(l,k) \in L_r \times K_c.$$

(10) $d_{lk}$ : an observed measurement vector from pixel (l,k).

(11) $C_{lk}$ : assigned category label of pixel (l,k) in the neighborhood.

(12) $\Omega$ : set of all possible categories.

(13) $\theta$ : designates a pattern configuration of assigned labels in neighborhood.

(14) Q : set of all possible pattern configurations of assigned labels in neighborhood.

$Q = \{ \theta_1 , \theta_2 , \cdots \theta_m \}$ ,

m is total number of pattern configuration of assigned lables in neighborhood.

(15) N : designates a neighborhood system on the two dimensional finite set of I x J.

(16) L : designates a clique in the neighborhood of pixel ( i,j ).

(17) W : designates family of cliques in the neighborhood of pixel ( i,j ).

$W = \{L_1, L_2 .... L_k \}$ , k is total number of cliques in $N_{ij}$.

(18) $V_L$ : designates a potential associated with clique L.

$V_L : R^{|L|} \rightarrow [ 0, \infty )$ $\qquad L \in W$

(19) $U(D_{ij})$ : designates an energy function associated with Gibbs distribution $P(D_{ij})$.

U(C) : designates an energy function associated with Gibbs distribution P(C).

$U(D_{ij} , C )$ : designates an energy function associated with

Gibbs distribution $P(D_{ij},C)$.

## (IV-1.2) Bayesian context classification model

From the Bayesian Model (Haralick,1983), the context classification problem can be stated as follows: to assign labels C to the pixels in the neighborhood of pixel ( i,j ) which minimizes the expected loss

$$\sum_{C^*} Los( C, C^* )P( C^* | D_{ij},Q ) \qquad (4.2)$$

where $P(C^* | D_{ij},Q )$ is the probability that true labeling of the pixels is $C^*$ given i) the measurements $D_{ij}$ of the pixels in the neighborhood of pixel ( i,j ) ii) the prior information Q we have about pixel dependencies. And where $Los(C,C^* )$ is the loss incurred for the assignment of interpretation C to the pixels in the neighborhood of pixel ( i,j ), when the true interpretation is $C^*$ .

We use the most common zero-one loss function for our study problem. There is no loss for a correct joint assignment and unit loss for any incorrect joint assignment. Here, correct assignment means that each pixel in the neighborhood is assigned correctly. Thus, there is no distinction in loss between an incorrect joint assignment in which only one pixel is incorrectly assigned or an incorrect assignment in which all pixels are incorrectly

assigned.

Such a loss function is defined by

$$Los(\, C\, ,\, C^{*}\, ) \;=\; \begin{cases} 0 & \text{where } C = C^{*} \\ 1 & \text{otherwise} \end{cases} \qquad (4.3)$$

There are two assumptions about the world and pixel measurement process which can simplify the expected loss expression (4.2).

The first assumption states that the description process is local. When the pixel ( i,j ) is being examined, no characteristics from any other pixel but pixel ( i,j ) affect the description obtained from pixel ( i,j ). Hence,

$$P(\, D_{ij} \mid C^{*}\, ,\, Q\, ) \;=\; \prod_{(l,k)\, \epsilon\, Nij} P(\, d_{lk} \mid C^{*}\, ,\, Q\, ) \qquad (4.4)$$

The second assumption states that the n-tuple measurement of pixel ( i,j ) depends only upon the true interpretation $C^{*}$ associated pixel ( i,j ) and does not depend upon any relationships pixel ( i,j ) may have with other units or upon the interpretation associated with any other pixel. Hence

$$P(\, d_{ij} \mid C^{*}\, ,\, Q\, ) \;=\; P(\, d_{ij} \mid c^{*}_{ij}\, ) \qquad (i,j)\, \epsilon\, \Omega \qquad (4.5)$$

Under these assumptions, the optimal decision rule determines interpretation C for the pixels in the neighborhood which minimize:

$$\sum_{C^{*}} Los(\, C, C^{*}\, ) \prod_{(l,k)\, \epsilon\, N_{ij}} P(\, d_{lk} \mid c^{*}_{lk}\, )\, P(\, C^{*}\, ) \,/\, P(\, D_{ij}\, ) \qquad (4.6)$$

With the loss function defined by (4.3), the best decision procedure chooses interpretation C which satisfies the maximality condition

$$\prod_{(l,k) \in N_{ij}} P( d_{lk} \mid C_{lk}) P( C ) \geq \prod_{(l,k) \in N_{ij}} P( d_{lk} \mid Z_{lk}) P( Z ) \qquad (4.7)$$

for all $Z \in \Omega$

The choice of C satisfying this maximality condition cannot be independently done pixel by pixel.

## (IV-2) MARKOV RANDOM FIELDS WITH NEAREST NEIGHBOR ASSUMPTION

It is clear that any efficient computer algorithm for image analysis, classification, and processing can only be done using the framework of a proper image model. The Markov Random field and Gibbs model, which is pervasive in the image processing literature, constitute a promising natural way to capture context assumptions in classification.

Consistent with the two-dimensional (2-D) discrete Markov Random Field for multispectral image processing applications, we assume a random observation vector $d_{ij}$ , and the pixel position ( i,j ) is defined on the two dimensional finite integer set of size I x J.

The Markov Random Field model may be defined as below :

Let $\{d_{ij} \mid (i,j) \in I \times J\}$ be an observed image, and I, J and (i,j) are seen the Section 4.1. Let $N_{ij}$ be the appropriate symmetric neighbor set of the pixel ( i,j ). It is postulated that this is generated by an appropriate 2-D ( non causal) Markov Random Field model. The model characterizes the statistical dependency among pixels by requiring that

$$P(d_{ij} \mid d_{mn}: (m,n) \in I \times J, (m,n) \neq (i,j)) = P(d_{ij} \mid d_{mn}: (m,n) \in N_{ij}) \quad (4.8)$$

If $N_{ij} = \{(0,1),(0,-1),(-1,0),(1,0)\}$ it corresponds to taking the simplest Markov model. By including more neighbors we can construct a higher order Markov model.

Unlike the 1-D discrete time series, where the existence of a preferred direction is inherently assumed, no such preferred ordering of the discrete neighborhood is appropriate for 2D. In other words, the notion of "past" and "future" as understood in a unilateral 1-D Markov process is restrictive in 2-D, as it implies a particular ordering in which the observations are scanned top down and left to right. It is quite possible that an observation at a pixel p may be dependent on surrounding observations in all directions.

An alternative representation of random fields is by the Gibbs distribution. We say that a random field has a Gibbs distribution if its density function is given by

$$P(D_{ij}) = \frac{1}{Z} e^{\frac{-U(D_{ij})}{KT}} \tag{4.9}$$

where K is a Boltzmann's constant, T is 'temperature', and $U(D_{ij})$ is called energy. Z is the normalizing constant, which is shown below:

$$Z = \sum_{D_{ij}} e^{\frac{-U(D_{ij})}{KT}} \tag{4.10}$$

The energy $U(D_{ij})$ is the sum of local potentials $V_L(D_{ij})$ such that:

$$U(D_{ij}) = -\sum_{L \in W} V_L(D_{ij}) \tag{4.11}$$

The $V_L(D_{ij})$ are called local potentials, which are evaluated over each clique on the neighborhood of pixel ( i,j ), where a clique is a such subset in which each pixel is a neighbor of all pixels in the subset L. $W_{ij}$ designates the family of cliques in the neighborhood of pixel ( i,j ).

## (IV-2.1) Markov Properties

Before describing the relaxation algorithm, it is first necessary to define neighborhood system and Markov Random Field, and then to discuss the factorizability property characteristic to Gibbs states with nearest neighbor potentials.

### i) Definition 4.1 : (H. Derin, 1986)

A collection of subsets of I x J described as $N_{ij} \in I \times J$ is a neighborhood system over the two-dimensional finite set of I x J, if and only if $N_{ij}$ , the neighborhood of pixel ( i,j ) does such that : i) ( i,j ) does not belong to $N_{ij}$ and ii) if $(k,l) \in N_{ij}$ , then $( i,j ) \in N_{kl}$ , for any $(i,j) \in I \times J$.

We can now formally define a MRF with respect to the neighborhood system N defined over the two-dimensional finite set of I x J.

**ii) Definition 4.2 :**

Suppose N is a neighborhood system defined over the two-dimensional finite set of I x J. A random field $\{ d_{ij} : ( i,j ) \in I \times J \}$ is a Markov random field with respect to the neighborhood system N if and only if

$$P( d_{ij} \mid d_{lk} ,(l,k) \in I \times J, (k,l) \neq ( i,j ) ) = P( d_{ij} \mid d_{lk}, (k,l) \in N_{ij} )$$
$$\text{for all } ( i,j ) \in I \times J.$$

**iii) Factorizability property :**

A set of pixels L is a clique if every pair of pixels in L are neighbors. We include the empty set ( no pixel ) as a clique. If the Gibbs model employs only the empty set ( no pixel ), single pixel, and cliques consisting of pairs of pixels, it is called a first order model. If we also consider cliques consisting of triples of pixels, it is called a second-order model.

The factorizability property is the theoretical base of decomposing the potential functions by cliques (see formula 4.11). It is very useful for finding the canonical potential form in our problem.

The factorization property can be stated as below:

Suppose that the rectangular lattice G, defined as G = I x J has several rectangular sublattices ( connected components ) $G_m$, m = 1,2..., defined as $G_m \in G$. If $D = \{ d_{ij}:(i,j) \in G\}$ is Markov random field on the rectangular lattice G, and each $D_m = \{ d_{ij}:(i,j) \in G_m \}$ is also Markov Random Field defined on the rectangular sublattice $G_m$, Then the probability P over the Markov Random Field G is the product measure

$$P = \prod_m P_m \qquad (m = 1,2,....)$$

where $P_m$ is probability measurement of Markov Random Fields $G_m$.

From the above, we see that the factorization property guarantees the decomposition of the complex potential function (see the equation 4.11) into a set of simple potential functions of each clique over the neighborhood.

**(IV-2.2) Markov-Gibbs equivalence :**

We call two state representations "equivalent", if one of them deter-

mines another and vice versa.

Preston (1974) proved that the following are "equivalent" for a state $\pi$ in a two-dimensional discrete random field:

i) $\pi$ is an equilibrium state.

ii) $\pi$ is a state of Markov Random Field.

iii) $\pi$ is Gibbs state with nearest neighbor potential.

The above equivalence, called the Markov-Gibbs equivalence, implies that a purely probabilistic notion of a Markov random field can be equated to the physically based Gibbs distribution. The Gibbs model describes the interaction of a macroscopic system in thermal equilibrium in the same way the spatial Markov models describe local dependence. For a Markov Random Field, the conditional probabilities are expressed in terms of nearest neighborhoods , while for a Gibbs distribution the energy E is the sum of potentials V measured over the same neighborhood .

From the above we see that the MRF-Gibbs equivalence provides an explicit formula for the joint probability distribution in terms of an energy function, and it supplies a powerful mechanism for modeling spatial features.

Kinderman (1973) proved that a nearest neighbor Gibbs distribution determines a Markov random field, and provided theoretical formulas to

determine the canonical potential from the local characteristics.

The proof that the Markov Random Field determines a nearest neighbor Gibbs distribution is given by Preston (1974), and Grimmett (1973). They note that there is not a unique potential function. However, there is a unique canonical potential which is singled out in the following manner: the states are renumbered 0,1,2,...,r with 0 playing the rule of a preferred states. The potential is then said to be a canonical potential if $V_c(\omega) = 0$ when $\omega$ assigns the value 0 to at least one site in C. It is then proved that there is a unique canonical potential for a given Markov Random Field .

From the previous discussion we have seen that we are able to determine a Markov field if we have a nearest neighbor potentials. If we have a Markov random field on a finite two-dimensional integer set I x J, then the local characteristics do uniquely determine this measure. In fact, the canonical potential can be determined from these local characteristics by a formula, which we will discuss in section (IV-3), and then the Gibbs distribution is determined. We cannot merely choose any set of local characteristics, because we have to satisfy the measure and be consistent with these characteristics. The greatest difficulty is in choosing appropriate local characteristics, so the key step in the method becomes to determine the proper potential function. We will discuss this in the section (IV-3).

**(iv-2.3) maximizing entropy :**

In this section we note that the Gibbs distribution also can be derived by maximizing entropy.

For any probability measure $P(\omega)$ on the two-dimensional integer set I x J. the entropy S(P) is defined by

$$S(P) = -\sum_{\theta_{ij} \in Q} P(\theta_{ij}) \log P(\theta_{ij}) \qquad (4.12)$$

where Q is set of all possible pattern configurations of assigned lables in neighborhood.

We can view the entropy of a measure as the amount of uncertainty in the outcome. For example, in the two-dimensional finite integer set of size I x J the measure with greatest entropy is the measure which assigns all outcomes with equal probability.

In the typical application of the Ising model (Ising,1925), which is the earliest and best-known lattice system, one attempts assign a probability measure to a sample space N which represents outcomes which cannot be observed. In practice, only very broad properties of the system can be observed in statistical mechanics. Assume that we could estimate at least the expected value of energy $U(D_{ij}) = a$ . Then among all measures having expected energy value of a the Gibbs measure defined by

$$P( D_{ij} ) = \frac{1}{Z} \; e^{\frac{- U( D_{ij} )}{KT}} \qquad (4.13)$$

is the measure which maximizes entropy among all measures which make the expected energy agree with our estimated value a. Thus we have chosen the measure which has the greatest uncertainty, as measured by entropy, among all possible measures with given expected energy.

## (IV-3) The Markov-Gibbs model for Bayes' context classification:

In this section, we will show how the Markov-Gibbs model is incorporated with the Bayes' context classification, and how the optimal decision rule determines an interpretation C for the pixels in the neighborhood of pixel ( i,j ) which satisfies the maximality condition (4.7).

From (4.7), we know that with the zero-one loss function, the best decision procedure chooses a labeling C which satisfies the maximality condition

$$\prod_{(l,k)} P( d_{lk} \mid C_{lk}) \, P( C ) \geqslant \prod_{(l,k)} P( d_{lk} \mid Z_{lk}) \, P( Z ) \quad \text{for all } Z \in \Omega$$

From the previous section, we can see that

$$P( D_{ij} , C ) = \prod_{(ij) \in N} P( d_{ij} \mid C_{ij}) \, P(C) \qquad (4.14)$$

is a Gibbs distribution, since we can rewrite $P(D_{ij} ,C)$ as

$$P( D_{ij} , C ) = \frac{1}{Z} \; e^{\frac{- U (D_{ij}, C)}{KT}} \tag{4.15}$$

where $U(D_{ij} , C )$ is an energy function associated with Gibbs distrubution $P(D_{ij} , C )$, which has the form:

$$U( D_{ij} , C ) = - \sum \log P(d_{lk} \mid c_{lk} ) + U(C) \tag{4.16}$$

and U(C) satisfies

$$P (C) = \frac{1}{Z} e^{\frac{- U(C)}{KT}} \tag{4.17}$$

Z and K are constants and U(C) is the energy function associated with Gibbs distrubution P(C), which has the form

$$U(C) = \sum_{L \in W_{ij}} V_L(C) \tag{4.18}$$

$W_{ij}$ denotes the family of cliques in the the neighborhood $N_{ij}$ of pixel ( i,j ). Each $V_L$ is a function on $N_{ij}$ with the property that $V_L(C)$ depends only on those coordinates and assigned labels of pixel ( i,j ), which are located in the clique L. Such a family $\{ V_L , L \in W \}$ is called a potential . The $V_L$ functions in (4.18) represent contributions to the energy U(C) from external fields (singleton cliques), pair interactions ( doubletons), etc.

Z is the normalizing constant given by

$$Z = \sum_{C} e^{\frac{- U(C)}{KT}} \tag{4.19}$$

T, stands for "temperature". For our purposes, T controls the degree of "peak' in the "density". Choosing T "small" makes it easier to find the

minimal energy configurations by sampling; this is the principle of annealing, and will be applied to our procedure in section IV-4.

The assigned category, in the sense of Bayesian inference, is determined by maximizing (4.15). This is a maximum posterior estimate. The probability is maximized when energy is minimized. This is analogous to the situation of thermal equilibrium in statistical physics, in which the most probable molecular configurations occur at the lowest energies. For the case of Bayes context classification, the most probable labeling occurs when the negative exponent is minimized. Using conventional gradient techniques, maximizing posterior probability is virtually impossible for all but the first order Markov Random Field models, because of the existence of many local extrema. However, stochastic relaxation , which is a new multivariate or combinatorial optimization technique (finding the minimum of a given function depending on many parameters), developed by Kirkpatrick et al (1983), offers a practical solution .

After creating the Gibbs models for Bayes' context classification, the problem now is to find $U(D_{ij},C)$.

A general form of $U(D_{ij},C)$ is that

$$U( D_{ij} , C) = \sum_{l,k} \log P( d_{lk} | C_{lk} ) + U(C)$$

$$U(C) = \sum_{(1,k) \in N_{ij}} V_{(1,k)}(C_{1k}) + \sum_{(1,k) \in N_{ij}} V_{(1,k),(1+1,k)}(C_{1k}, C_{1+1,k})$$

$$+ \sum_{(1,k) \in N_{ij}} V_{(1,k),(1,k+1)}(C_{1k}, C_{1,k+1}) \qquad (4.20)$$

where the summation is over all $(k,l) \in N_{ij}$, and $N_{ij}$ denotes the nearest-neighbor. Typically, several free parameters are involved in the specification of U(C). The Ising model (1925), which is the earliest and best-known lattice system , can be thought of the special case of (4.20) in which C is binary ( $C_{ij}$ =(0,1)), homogeneous ( strictly stationary ) and isotropic ( rotational invariant); Its potential function is

$$U(C) = \alpha \sum_{(1,k) \in N_{ij}} C_{1k} + \beta \left( \sum_{(1,k) \in N_{ij}} C_{1k} * C_{1+1k} + \sum_{(1,k) \in N_{ij}} C_{1k} * C_{1k+1} \right) \qquad (4.21)$$

for some parameters $\alpha$ and $\beta$, which measure, respectively, the external field and bonding strengths.

For our contextual classification case, in which U(C) is a function of pattern configurations, the expression for the Ising model (4.21) is not suitable. But we still assume that the image is homogeneous and isotropic.

Before we derive the canonical potential for the general case, we first describe a useful theorem of potential function (G. R. Grimmett 1973). If the random field $C = \langle c_{ij}, (i,j) \in I \times J \rangle$ is a Markov field, then its potential function is given by

$$V_L(C) = \sum_{L_1 \in W} (-1)^{|W - L_1|} \log P(C_{L_1}) \qquad (4.22)$$

where the summation is over all cliques in W, $C_{L_i}$ designates the configuration which agrees with C on $L_i$, but assigns the value 0 at all sites outside of $L_i$, and $\mu = (-1)^{|W-L_i|}$ is Möbius function, and W- $L_i$ , the difference of W and $L_i$, is {x | x is in W and x is not in $L_i$ }.

From the above theorem we see that given a Markov random field on a finite set I x J, the local characteristics ( see III-1) do uniquely determine this potential function; the canonical potential can then be determined from these local characteristics.

Define three types of potential functions as follows:

$$\sum_{(1,k)\,\epsilon\, N_{ij}} V_{\{(1,k)\}}(C_{1k}) = \sum_{(1,k)\,\epsilon\, N_{(ij)}} \log P(C_{1k})$$

$$\sum_{(1,k)\,\epsilon\, N_{ij}} V_{\{(1,k),(1+1,k)\}}(C_{1k}, C_{1+1k}) = 2 * \sum_{(1,k)\,\epsilon\, N_{(ij)}} \log P(C_{1k} \mid C_{1+1k})$$

$$\sum_{(1,k)\,\epsilon\, N_{ij}} V_{\{(1,k),(1,k+1)\}}(C_{1k}, C_{1k+1}) = 2* \sum_{(1,k)\,\epsilon\, N_{(ij)}} \log P(C_{1k} \mid C_{1k+1}) \qquad (4.23)$$

These definitions are based on the factorization of Markov Random Field and neighboring clique assumption. In this model only cliques of size two are involved.

From the above definitions, the problem in this chapter can be stated as follows: the assigned category is determined by minimizing

$$U(D_{ij}, C) = \sum_{(l,k)} \log P(d_{lk} \mid C_c) + U(C)$$

$$\text{where} \quad U(C) = \sum_{(l,k)} \log P(d_{lk} \mid C_c) + \sum_{(l,k) \in N_{(ij)}} \log P(C_{lk})$$

$$+ 2 * \sum_{(l,k) \in N_{(ij)}} \log P(C_{lk} \mid C_{l+1k})$$

$$+ 2 * \sum_{(l,k) \in N_{(ij)}} \log P(C_{lk} \mid C_{lk+1}) \qquad (4.24)$$

Because of the existence of many local extrema, the computation cost of maximizing the posterior probability for Bayes classification is usually computationally high. For example, if a MSS image has N class categories on a M x M lattice , the number of configurations is at least $N^{M^2}$. Hence, the identification of even a near-optimal solution is surprisingly difficult for such a relatively complex function. In the next section we present the implementation of a stochastic relaxation procedure, which overcomes the computational difficulty remarkably well.

## (IV-4) Implementation of the stochastic relaxation context classification:

The method used in the stochastic relaxation context classification is essentially a variant of a Monte Carlo procedure, due to Metropolis et al. (1983). In the Metropolis procedure samples are randomly generated from

a Gibbs distribution at constant temperature. This simulates the behavior of a physical system in thermal equilibrium. The algorithm can be briefly described as follows. For each state $D_{ij}$ of a model D, a random perturbation is made. The change in energy, $\Delta U$ is computed. If $\Delta U \leqslant 0$, the perturbation is accepted, that is the new pattern configuration, which corresponds to the new "energy", $U' = U_o + \Delta U$, replaces the original one. If $\Delta U$ is positive then the perturbation is accepted with probability

$$P(\Delta U) = e^{\frac{-\Delta U}{T}} \tag{4.25}$$

This conditional acceptance is easily implemented by choosing a random number R uniformly distributed between 0 and 1. If $R <= P(\Delta U)$ then the perturbation is accepted; otherwise the existing model is retained. Random perturbation according to these rules eventually causes the system to reach equilibrium, or the configuration $\theta$. corresponding to maximum probability. The technique used here slowly lowers the temperature T during execution of the iterative procedure. If the system is cooled sufficiently slowly and equilibrium conditions are maintained, the model converges to a state with minimum energy or maximum a posterior probability. This was proved by Geman (1985). Geman also pointed out that the most important aspect of any cooling function is that it be slow, especially near the critical temperature where convergence is rapid. The successful choice of an annealing schedule requires experience; ideally, the procedure would be interactive. As

T decreases, samples from the distribution are forced towards the minimal energy configurations. The temperature T(k) used by Geman satisfies the bound

$$T(k) \geqslant \frac{G}{\log(1+k)} \tag{4.26}$$

Where T(k) is the temperature during the $k^{th}$ interation, so that k is the total number of interation. For every k, G is a constant independent of k. When $k \rightarrow \infty$, the configurations generated by the algorithm will be those of minimal energy.

**(IV-5) summary of the stochastic relaxation context classification procedure:**

In summary, the stochastic relaxation context classification procedure an be implemented as follows:

(i) Evaluate training statistics. This includes the mean vector and covariance required for the Gaussian class conditional distribution.

(ii) Preclassify the image using a pixel independent or context free Bayes classification technique.

(iii) Evaluate the transition probabilities: $P(C_{ij} \mid C_{i,j+1})$ and $P(C_{ij} \mid C_{i+1,j})$ from the preclassification results.

(iv) Using equation (4.24),(4.25),(4.26) perform the stochastic relaxation context classification. The experimental results with both simulated and real multispectral remote sensing data will be presented in next chapter.

## (IV-6) Improved Scheme:

Now we have a desirable stochastic relaxation procedure, in which samples are randomly generated from a Gibbs distribution at a controlled temperature T. As T changes, samples from the distribution are forced towards the minimal energy configuration. Geman (1985) proved the convergence properties of this algorithm, and showed how to reduce the computational difficulty. As we mentioned before, for an MSS image which has N class categories and a M x M size, the number of configuration is at least $N^{M^2}$. In Gemans' scheme the pattern samples are randomly collected from a huge pattern configuration space. In contrast to our proposed method, his method had nothing to do with reduing the pattern configuration space. Experimental results showed that for a significant improvement in classification accuracy, the number of iterations was still sizeable.

In order to further reduce the computational complexity, it is important to reduce the size of the huge pattern configuration space or to place some constraints on the pattern generation procedure. We now describe

how we can use the homogeneous assumption to control the pattern configuration sampling procedure.

Most Landsat and aerial photograph images are divided into a number of elementary regions at the classification stage. Each region is finite, fairly homogenous, and has similar spectral properties over its entire ground surface. These homogeneous regions correspond to uniform objects ( categories ) on the earth's surface. We believe that some smooth or homogenous pattern configurations are much more probable than others, and some irregular patterns have very low probabilities.

This fact gives us a strategy for the iterative procedure in that we may use these most probable homogeneous patterns at the first stage of the iteration procedure. At the second stage, we randomly generate the pattern configuration and skip irregular patterns which have low probability.

We should note that the global procedure is still random; we only set very limited number of special pattern configurations into the beginning stage. From then on a Markov radom process is generated by annealing procedure. Therefore, this scheme is still a stochastic relaxation procedure.

Let T(t) be any decreasing sequence of temperatures for which (a) $T(t) \rightarrow 0$ as $t \rightarrow \infty$ (b) $T(t) \geq \frac{N\Delta}{\log t}$ for all $t \geq t_0$ and some integer $t_0 \geq 2$. And let the annealing procedure generates a process { D(t), t=1,2,...}.

Gemman proved that the distribution of D(t) converges to equilibrium distribution $\pi$, as $t \to \infty$ regardless of starting configuration. The only assumption is that we continue to visit every site.

The convergence of our modified procedure can be proved as follows. Let process $\{ D_1(t), t=1,2,...\}$ be generated by the annealing procedure described above, and process $\{ D_0(1), D_0(2), ....D_0(s) \}$ be a process with limited numbers of states. We create a new process D(t), which has

$$D(t) = \begin{cases} D_0(T) & \text{if } t \leqslant s \\ D_1(t-s) & \text{if } t > s \end{cases}$$

where $D_0(t)$ is a sequence of special pattern configurations at the first stage, and $D_1(t-s)$ is a process generated by the annealing procedure. Gemman states that the distribution of $D_1(t)$ converges to equilibrium distribution, as $t \to \infty$ regardless of starting configuration. Because the D(t) has same statistic property except having different starting configurations, new processes D(t) should converge to the same equilibrium distribution as $D_0(t)$, when $t \to \infty$.

First, we assume that uniform pattern configurations have higher occurring probabilities, and that they are generated and tested at the beginning of the iterative procedure. The assignment of these uniform labels is based on the labels in the neighborhood assigned at the classification stage. So the number of these uniform pattern configurations is equal to the

number of categories in the neighborhood.

Subsequent to the above testing, we assume that some simple pattern configurations (Figure 4-1) also have higher occurring probabilities. These are assigned and tested again.

The pattern in Figure 4-1-a has upper and lower parts. The assignment of labels in each part is also based on the labels in that part assigned at the preclassification stage. Similarly, Figures 4-1-b,c and d show three other simple patterns.

After these steps, a random pattern generator is introduced in the relaxation procedure.

In order to restrict irregular patterns, we employ a measure of the irregularity as follows:

$$IR = \frac{\text{NUMBER OF CLASSES IN THE NEIGHBORHOOD}}{\text{NUMBER OF PIXELS IN THE NEIGHBORHOOD}} \qquad (4.27)$$

After we give a threshold, the irregularity measurement of each pattern is calculated and compared with the threshold. If the measurement is larger than the threshold, the pattern is too irregular and the procedure will skip testing and generate the next pattern instead.

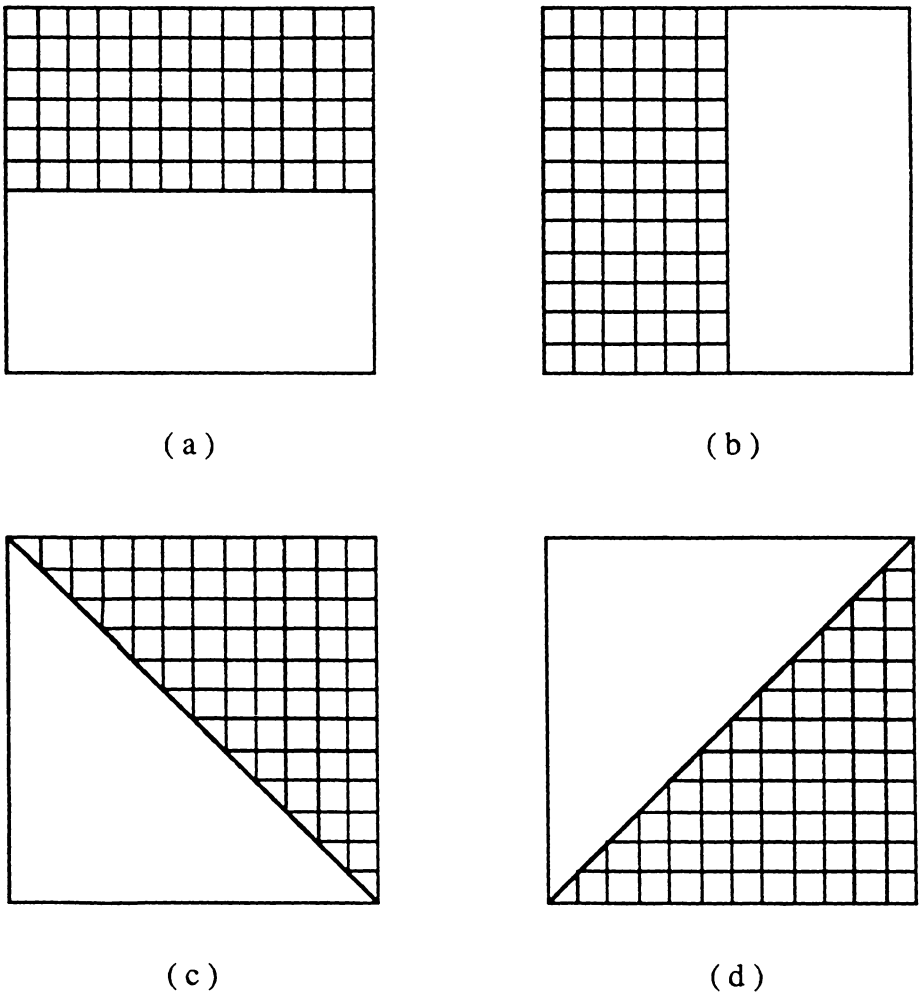


Figure 4.1 Four simple pattern configurations for the improved scheme. The pattern (a) has upper and lower parts; (b) has left and right parts; (c) has upper right and bottom left parts; and (d) has upper left and bottom right parts. The uniform pattern and the above four simple patterns configurations, which are assumed to have higher occurring probability, are generated and tested at the beginning of the iterative procedure.

## (IV-7) Parallel Algorithm :

Although the computational cost of the stochastic relaxation scheme is much more expensive than the conventional context free classification methods, it is highly parallel in the sense that it is implementable by simple and identical neighbor operators.

The performance of the each neighborhood operator is independent with other neighborhood operators in the entire image. The amount of time required for each iteration of the entire image is proportional only to the number of pixels in the image, if all operators are executed in parallel.

This important property allows the algorithm to run naturally in a fully parallel architecture.

A more modest degree of parallelism was noted by Geman (1985). A graph associated with the MRF is divided into collections of sites with each collection assigned to an independently running (asynchronous) processor.

Each such processor would execute a raster scan update of its assigned sites. Communication requirements will be small if the division of the graph respects the natural topology of the scene, provided of course that the neighborhood systems are reasonably local. Such an implementation with five or ten micro- or mini-computers represents a straightforward applica-

tion of available technology. Chen (1985) noted that data flow computer architecture should be useful for the stochastic relaxation (annealing) algorithm for Markov Random Fields.

## (IV-8) Object detection using contextual information

Another application of the stochastic relaxation method is to detect objects using contextual information. It requires segmenting the original gray tone image into a binary image using a contextual threshold technique.

The raw data of targets is only one band of those data available. So having additional contextual information becomes extremely valuable. The pixel independent maximum likelihood classification result is used for the preclassified labels.

At the first preclassification stage, two texture features, (ie. entropy and the inverse difference moment) and the original measurement value are selected to constitute the feature space. Training samples are selected within homogeneous background regions and target regions to obtain mean values and covariances matrices. The transition probabilities are estimated from the preclassification result. The experimental results are shown in the chapter V.

## (IV-9) Summary

We have developed a new multispectral image context classification algorithm with the Markov Random Field assumption, with which remotely sensed data are more accurately classified compared to traditional context free classifiers. This new approach of multispectral image context classification is based on a stochastic relaxation algorithm and the Markov-Gibbs Random Field. The implementation of the relaxation algorithm is one form of optimization using annealing. In this chapter, we have first motivated a Bayesian context decision rule, then introduced a Markov-Gibbs model for the original LANDSAT MSS image. Then we developed a new contextual classification algorithm, in which maximizing the posterior probability (MAP) is based on the stochastic relaxation and annealing method. An improved algorithm has been presented to speed the stochastic relaxation procedure. It has greatly reduced the number of iterations by using some special pattern configurations at the beginning of the iterative procedure. The algorithm is highly parallel and exploits the equivalence between Gibbs distributions and Markov Random Fields (MRF).

# CHAPTER V: EXPERIMENT RESULTS OF THE CONTEXTUAL CLASSIFICATION ALGORITHMS

In order to show accuracy in the improvement of dynamic programming and stochastic relaxation context classification methods, and to comparatively study the three dynamic programming context decision making algorithms, several experimental results based on both simulated and real multispectral remote sensing data are illustrated.

In the section V-1, we show comparative experimental results of three different contextual decision rules described in chapter III, and a Bayes context-free decision rules using pseudo-random images. Classification results are illustrated in the form of error-reject curves. They are set at five different SNR and transition probability values, and for each case 25 simulated test images are generated with different random number seeds, final results show the average and standard deviation of these 25 experimental results. On the basis of the statistical test of these experimental data, conclusions about the performances of three different contextual decision rules and Bayes context-free decision rules are given out.

In section V-2 we discuss another simulated data generating method. For each pixel, simulated data vectors are produced by a Gaussian random number generator having the same mean vector and covariance matrix as the class associated with the pixel on the ground truth map. Finally, the

experimental results of real and simulated remote sensing images are illustrated in section V-3. The classification results in the five study arears show that the contextual classification results yield significant improvement over the context free classification.

## (V-1) First simulation method :

Classification accuracy can vary with different kinds of input data sets for a given classification algorithm. It is difficult to evaluate the effectiveness of classification algorithms using small sets of real image data.

A desirable way to evaluate the effectiveness of classification algorithms is to use simple simulated data sets. In this subsection, we illustrate two kinds of simulated data experiments: one is directly generated from a Markov Random Field model, and another one is generated from the ground truth of a real remote sensing image.

Following Devijvers'(1985) suggestion we generate a data set as follows: A two-dimensional Markov pseudo-random source is selected to be a simple, idealized model of simulated image production. Specifically, we adopted a 6-category model. The transition probability for two 4-adjacent pixels with same category label is p and for two 4-adjacent pixels with different category labels is $\frac{(1-p)}{5}$, that is

$$P(C_{ij} \mid C_{i-1j}, C_{ij-1}) = P(C_{ij} \mid C_{i-1j}) P(C_{ij} \mid C_{ij-1}) \tag{5.1}$$

$$P(C_{ij} \mid C_{i-1j}) = \begin{cases} p & \text{if } C_{ij} = C_{i-1j} \\ \frac{(1-p)}{5} & \text{otherwise} \end{cases} \tag{5.2}$$

$$P(C_{ij} \mid C_{ij-1}) = \begin{cases} p & \text{if } C_{ij} = C_{ij-1} \\ \frac{(1-p)}{5} & \text{otherwise} \end{cases} \tag{5.3}$$

Figures 5.1 and 5.2 show two example simulated images.

A common way to depict the classification results is in the form of error-reject curves, in which the classification accuracy is a function of signal to noise (SNR) and the entropy of the pixels in the image.

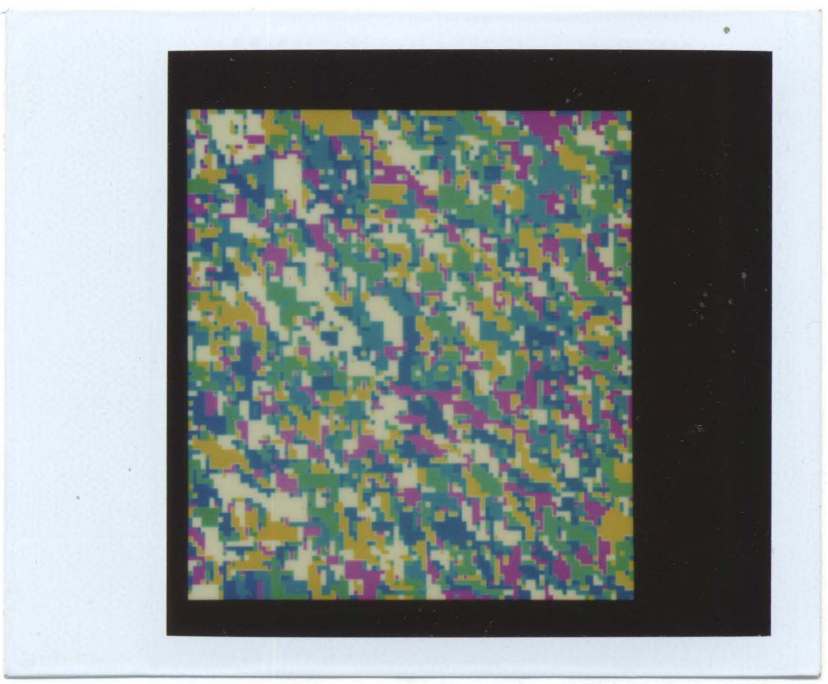

Figure 5.1 Markov pseudo-random image with transition probability p = 0.4

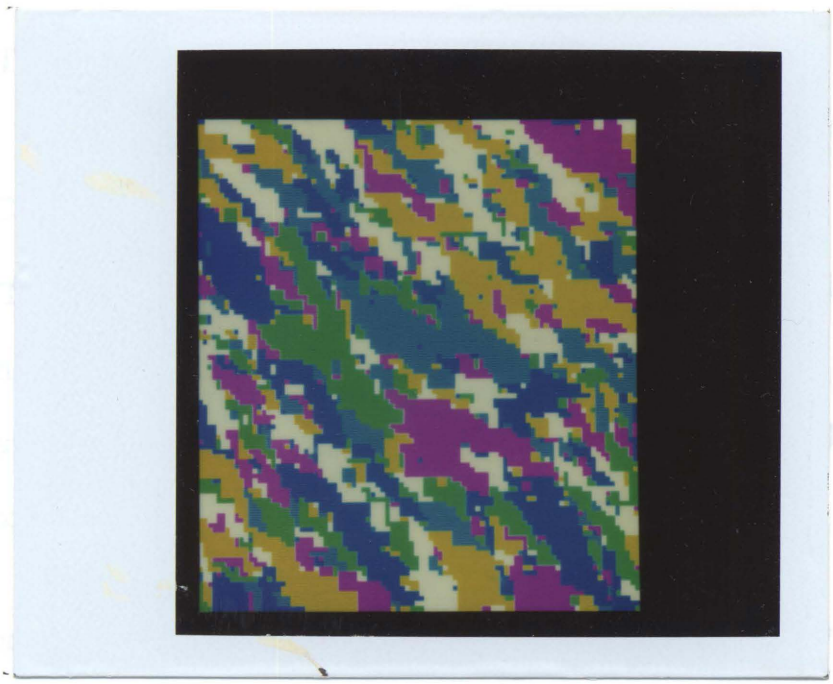

Figure 5.2 Markov pseudo-random image with transition probability p = 0.7

The pseudo-random Markov image is generated in the following way:

i) The top left pixel in the image is assigned a class randomly with equal probability for each possible class.

ii) The selection of classes for successive pixels is performed in raster scan order.

iii) The class of each pixel is determined by the transition probabilities indexed by the states of the nearest north and west neighbors $P( C_{ij} \mid C_{ij-1} , C_{i-1j} )$.

iv) The selection strategy is to partition the unit interval proportionally to the transition probability from nearest north and west neighbors to the current pixel, to generate a random number, and then to chose the new state, according to the subinterval in which the number falls.

Given a category, the conditional distributions were chosen to be 2-dimensional normal ie., $P_i(X) \sim N( m , \Sigma )$. $\Sigma$ is the 2 x 2 identity matrix. Each class mean vector m is located at the vertex of a regular hexagon centered at (128,128) inscribed within a circle of radius R. The signal-to-noise ratio (SNR) can be defined as

$$SNR = \frac{S_0}{N_0}$$

where $N_0$ is the average noise power, and $S_0$ is average signal power. Ignoring context, the signal to noise ratio (SNR) in our problem is measured

by the ratio of the traces of the between class and within class covariance matrix. In our configuration SNR = $R^2$. From this choice for an experimental model, there are only two parameters, namely p and SNR.

Our comparative experiments consist of pseudo-random image classifications using the three different contextual decision rules described in the previous section, and of a Bayes context-free decision rules.

The error rate (probability of misrecognition) and the reject rate are commonly used to describe the performance of a recognition system. An error occurs when a pattern from one class is identified as that from a different class. A reject occurs when the recognition system withholds its recognition decision, and the pattern is rejected for exceptional handling, such as rescan or manual inspection. Because of uncertainties and noise inherent in any pattern recognition task, errors are generally unavoidable. The option to reject is introduced to safeguard against excessive misrecognition. However, when the rejection option is exercised, some would-be correct recognitions are also converted into rejects. Thus the problem of best error-reject tradeoff must be considred.

Classification results in this section are illustrated in the form of error-reject curves, because the performance of the Bayes context-free decision rule and contextual decision rules are characterized by acceptance, rejection, and error probabilities, and probability of correct decision. The

reject ratio is defined by ratio of number of rejected pixels in classification procedure, in which probabilities are less than certain threshold, to total number of pixels in the images.

In order to prove that the experimental results are not by chance, 25 simulated test images generated with different random number seeds are used. The final results show the average and standard deviation of these 25 experimental results.

These test images each with 100 by 100 pixels, are used in the experiments. They are set at five different SNR and transition probability values:

$(SNR,p) \in \{ (4,4),(9,0.4),(9,0.7),(4,0.55),(9,0.55)\}$

For each specified value of SNR and transition probability, classification methods based on the three contextual decision rules and on the Bayes context-free decision rule are performed.

The means and standard deviations of the overall classification accuracies of four different dynamic programming approaches and pixel independent Bayes' classifier are shown in Figures 5.3.a - 5.7.a. The final error-reject curves obtained from the average of repeated experimental results with 25 different random seeds are shown in Figures 5.3.b - 5.7.b.

In order to qualify our estimation in some way to indicate the reliabil-

ity or precision of an estimation process, we show a confidence interval. It is a range of value within which the true value of parameter $\theta$ ( ie. total classification accuracy ) is included with some probability. To obtain an interval, we need to find two statistics $t_1$ and $t_2$ such that the probability statement $P( t_1 \leqslant \theta \leqslant t_2 ) = 1 - \alpha$ is true.

The interval for the total classification accuracy at confident level $\alpha$ is

$$t_1 \leqslant \theta \leqslant t_2$$

where $\theta$ is normally distributed with unknown mean $\mu$ and variance $\sigma^2$.

If a random sample of n observations is taken and $\bar{y}$ computed, then

$$\frac{\bar{y}-\mu}{\frac{s}{\sqrt{n}}} \quad N(0,1)$$

where $s^2$ is the sample variance.

Expressing this as a probability statement, we have

$$P( -t \leqslant \frac{\bar{y}-\mu}{\frac{s}{\sqrt{n}}} \leqslant t ) = 1 - \alpha$$

where t is the upper-tail percentage point of the standard normal distribution, such that the probability to the right of t is $\frac{\alpha}{2}$.

It may be rewritten as

$$P(\bar{y} - t\frac{s}{\sqrt{n}} \leqslant \mu \leqslant \bar{y} + t\frac{s}{\sqrt{n}}) = 1 - \alpha$$

Finally we have

$$\bar{y} - t\frac{s}{\sqrt{n}} \leqslant \mu \leqslant \bar{y} + t\frac{s}{\sqrt{n}}$$

It is confidence interval for $\mu$ at confidence level $\alpha$.

Table 5-1 shows the confidence intervals for total classification accuracies using different context decision rules and the free decision rule, respectively. Five data sets with different SNR and transition probability values are included. Each data set has 25 samples with different random seeds.

On the basis of the experimental data provided, the conclusions are

(1) The lower the transition probability of the Markov source, the more effective is the context decision rule.

(2) One-step look-ahead has a fairly low computational cost, and yields a significant improvement over the context free rule.

(3) The complete context algorithm ( forward-backward algorithm ) always provides an answer as good as or better than the one-step look-ahead.

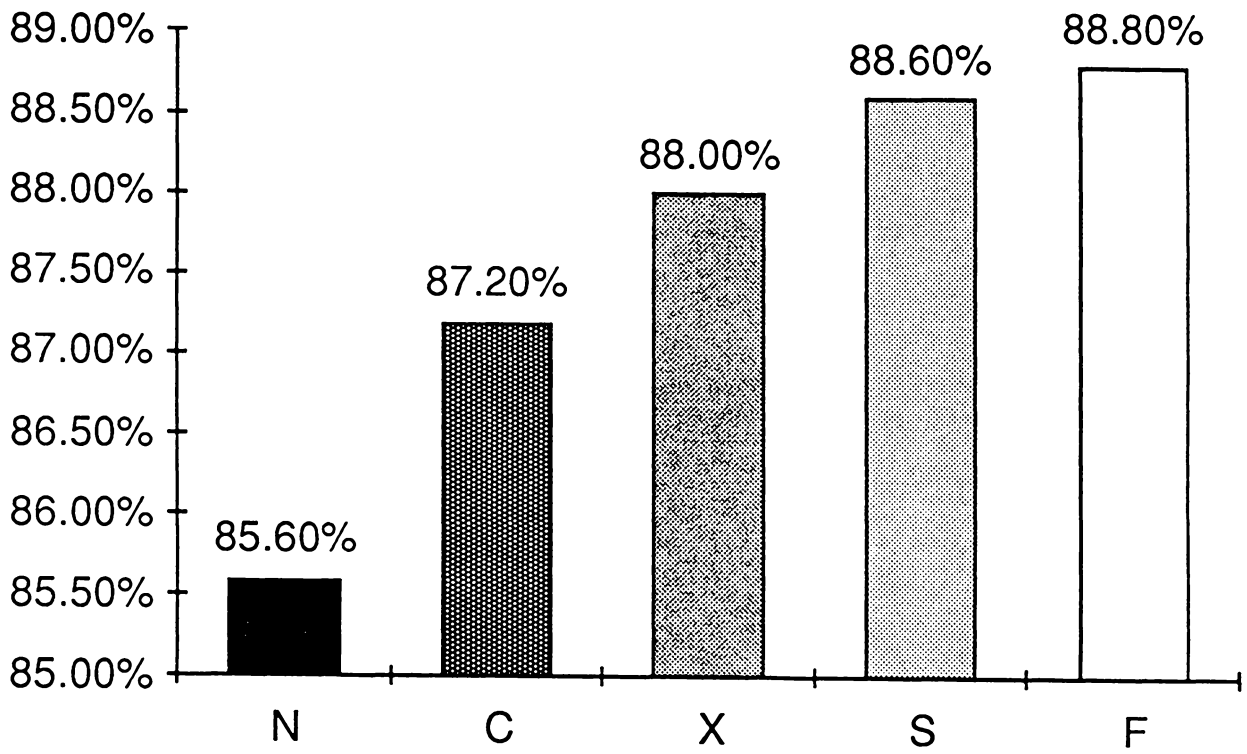


Figure 5.3a Comparision of mean values of the overall classification accuracies of simulated images with 25 different random seeds, SNR = 9 and P = 0.4. N : context free classification; C : no look-ahead context classification; X : one step look-ahead context classification; S : two pass full context classification; F : four pass full context look-ahead algorithm.

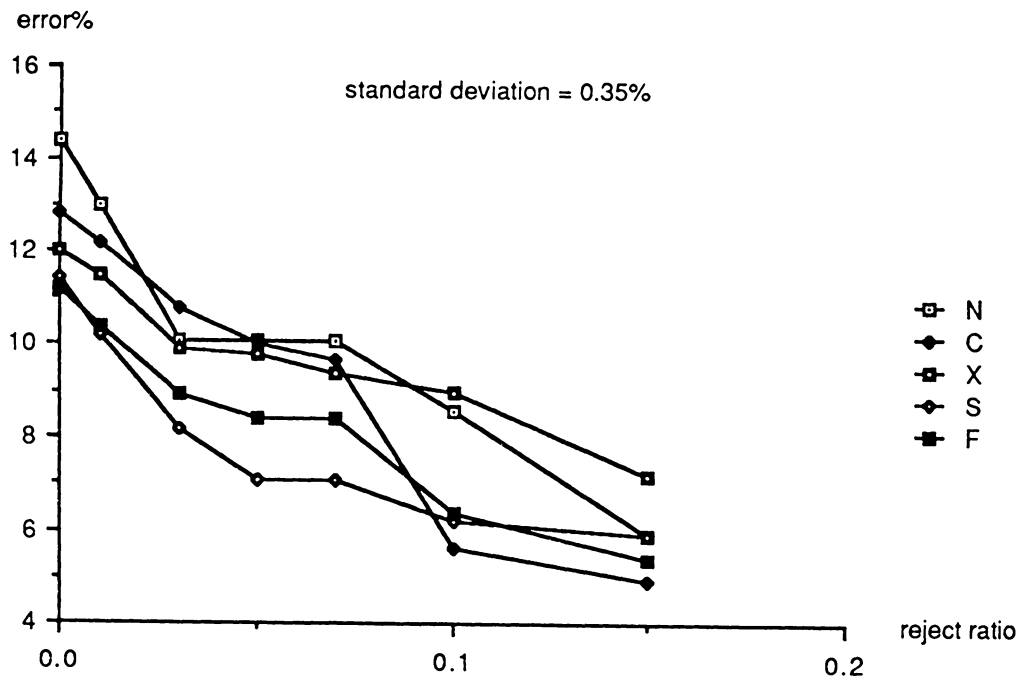


Figure 5.3b Error-reject curves, SNR = 9, P = 0.4; Circle : two pass full context classification, Square: one step look-ahead context classification, Diamond: no look-ahead context classification , Triangle: context free classification, Cross : four pass full context look-ahead algorithm.

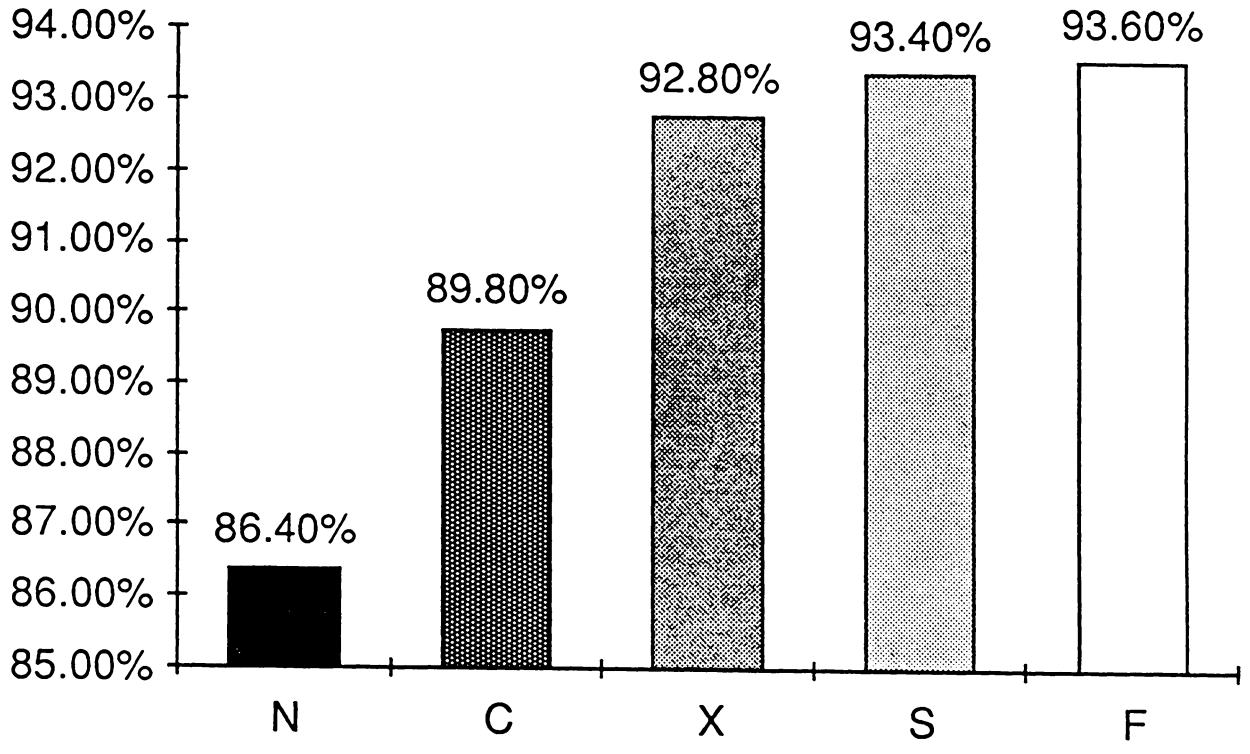


Figure 5.4a Comparision of mean values of the overall classification accuracies of simulated images with 25 different random seeds, SNR = 9 and P = 0.7. N : context free classification; C : no look-ahead context classification; X : one step look-ahead context classification; S : two pass full context classification; F : four pass full context look-ahead algorithm.

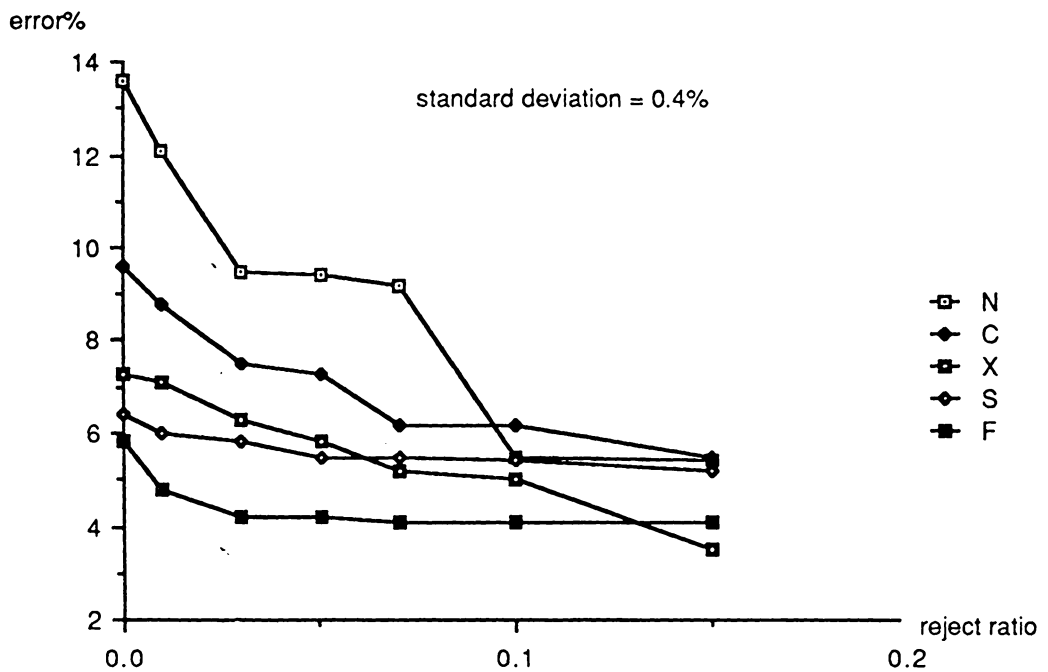


Figure 5.4b Error-reject curves, SNR = 9, P = 0.7; Circle : two pass full context classification, Square: one step look-ahead context classification, Diamond: no look-ahead context classification , Triangle: context free classification, Cross : four pass full context look-ahead algorithm.

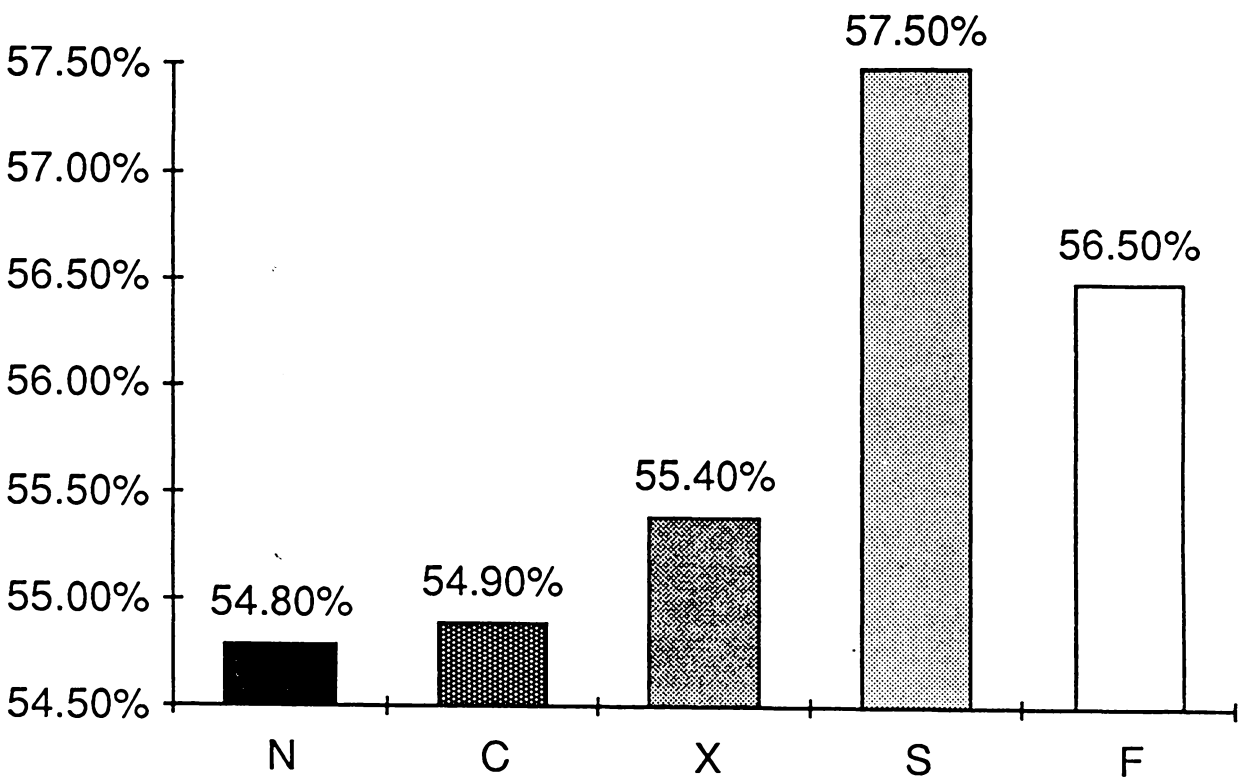


Figure 5.5a Comparison of mean values of the overall classification accuracies of simulated images with 25 different random seeds, SNR = 4 and P = 0.4. N : context free classification; C : no look-ahead context classification; X : one step look-ahead context classification; S : two pass full context classification; F : four pass full context look-ahead algorithm.

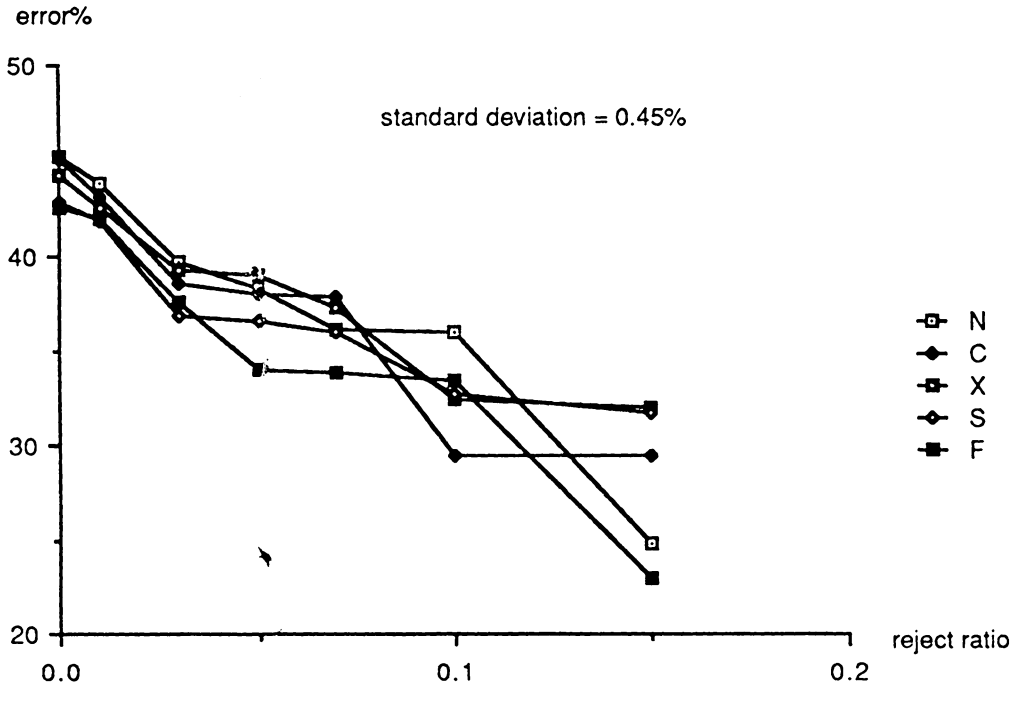


Figure 5.5b Error-reject curves, SNR = 4, P = 0.4; Circle : two pass full context classification, Square: one step look-ahead context classification, Diamond: no look-ahead context classification , Triangle: context free classification, Cross : four pass full context look-ahead algorithm.

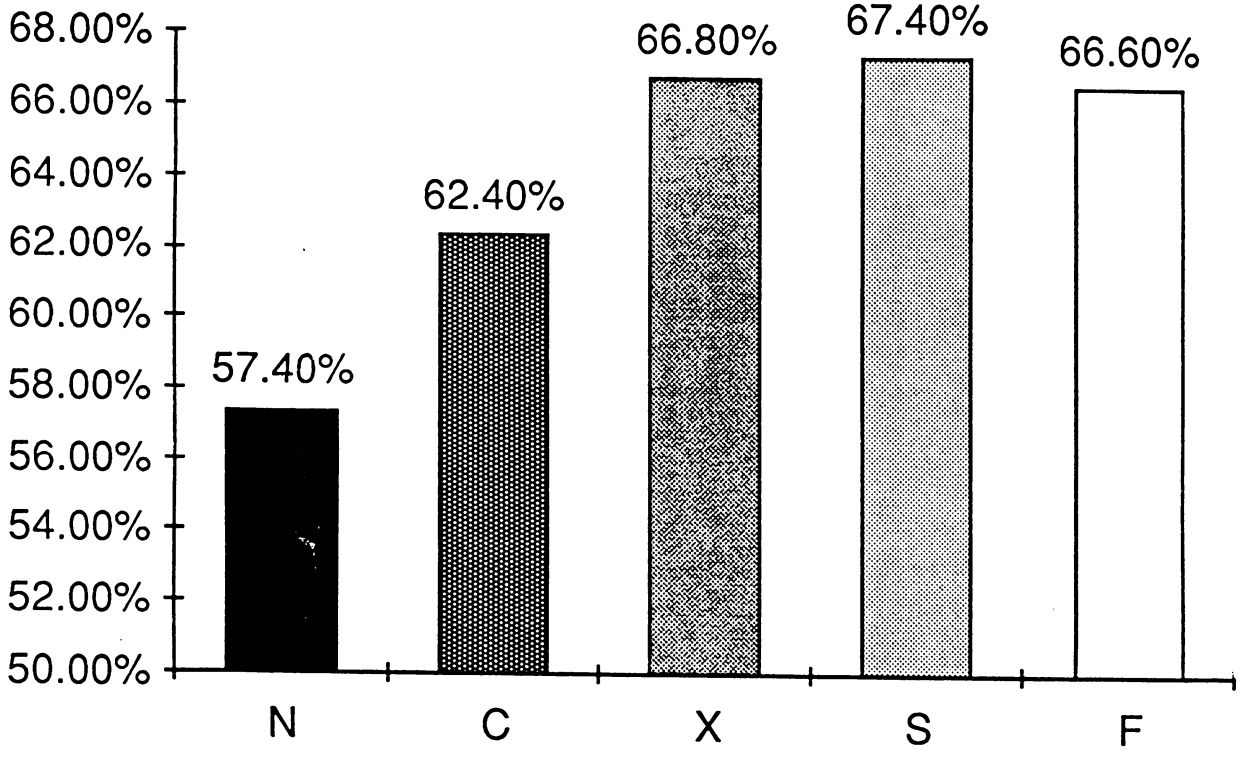


Figure 5.6a Comparison of mean values of the overall classification accuracies of simulated images with 25 different random seeds, SNR = 4 and P = 0.7. N : context free classification; C : no look-ahead context classification; X : one step look-ahead context classification; S : two pass full context classification; F : four pass full context look-ahead algorithm.

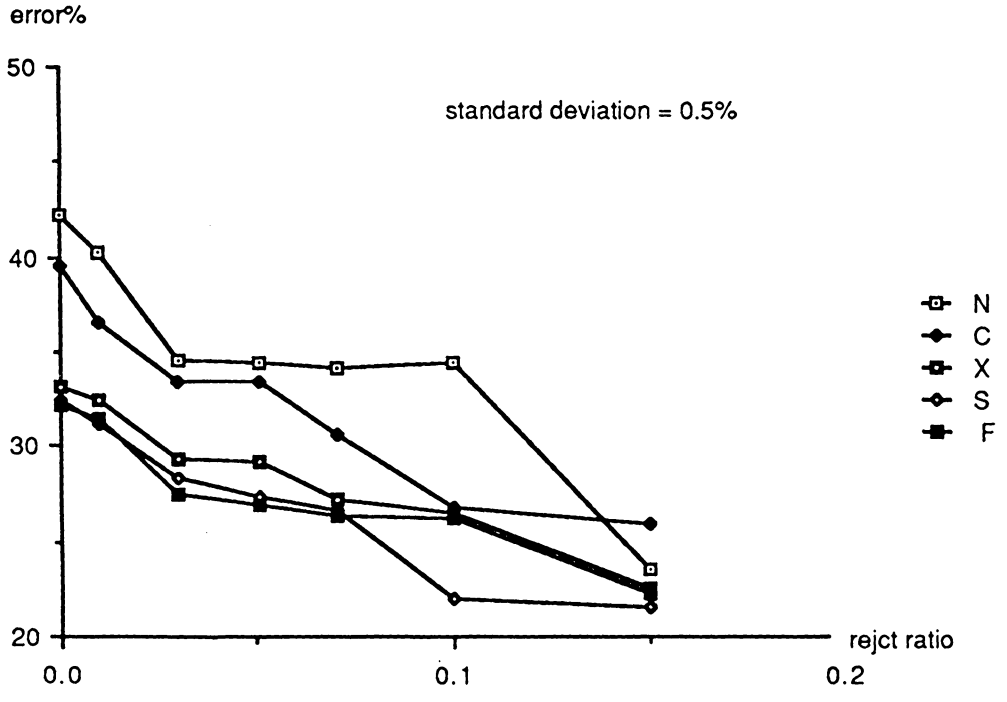


Figure 5.6b Error-reject curves, SNR = 4, P = 0.55; Circle : two pass full context classification, Square: one step look-ahead context classification, Diamond: no look-ahead context classification , Triangle: context free classification, Cross : four pass full context look-ahead algorithm.

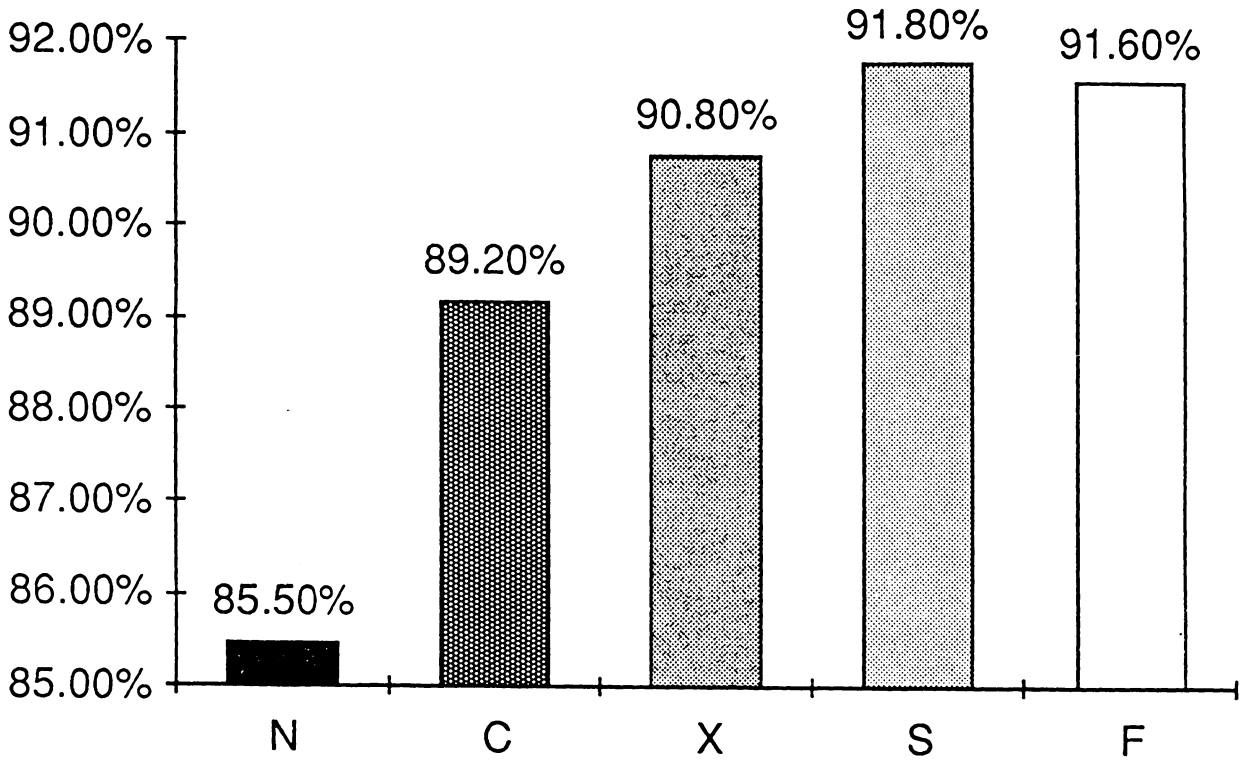


Figure 5.7a Comparision of mean values of the overall classification accuracies of simulated images with 25 different random seeds, SNR = 9 and P = 0.55. N : context free classification; C : no look-ahead context classification; X : one step look-ahead context classification; S : two pass full context classification; F : four pass full context look-ahead algorithm.

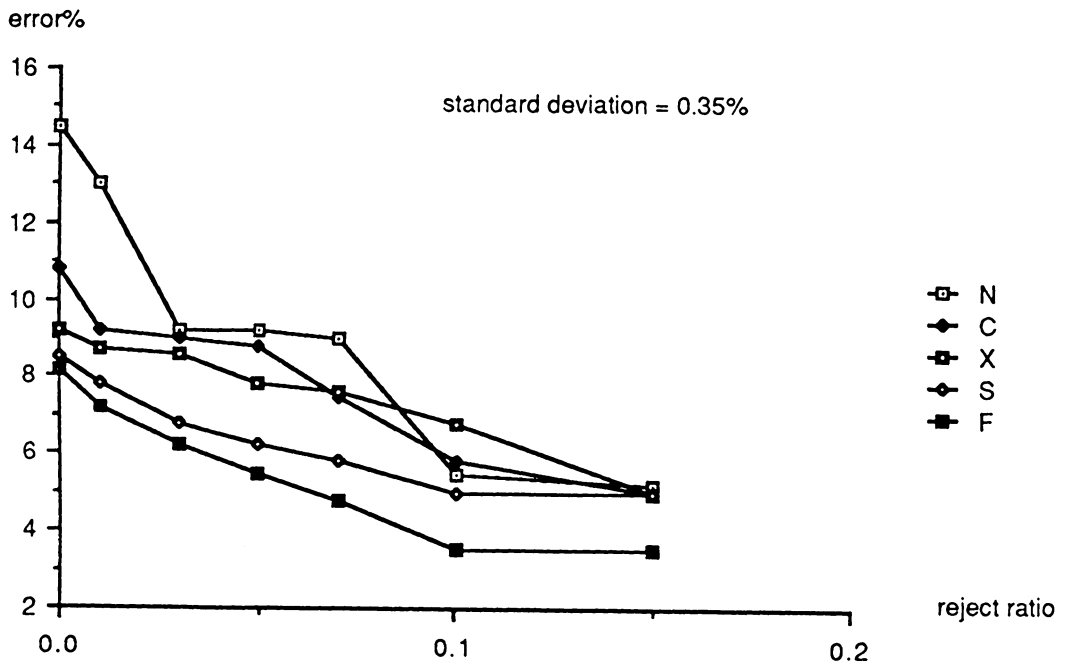


Figure 5.7b Error-reject curves, SNR = 9, P = 0.55; Circle two pass full context classification, Square: one step look-ahead context classification, Diamond: no look-ahead context classification , Triangle: context free classification, Cross : four pass full context look-ahead algorithm.

**Table 5.1 Confidence intervals at the 0.05 level for total accuracy of classification results**

| goup | type | n | mean | var | lower limit | upper limit | range *0.001 |
|---|---|---|---|---|---|---|---|
| A | N | 25 | 85.6% | 0.004 | 85.4% | 85.8% | (-.16,.16) |
| A | C | 25 | 87.2% | 0.0035 | 87.0% | 87.4% | (-.14,.14) |
| A | X | 25 | 88.0% | 0.0035 | 87.8% | 88.2% | (-.14,.14) |
| A | S | 25 | 88.6% | 0.0035 | 88.4% | 88.8% | (-.14,.14) |
| A | F | 25 | 88.8% | 0.0045 | 88.6% | 89.0% | (-.18,.18) |
| B | N | 25 | 86.4% | 0.004 | 86.2% | 86.6% | (-.16,.16) |
| B | C | 25 | 89.8% | 0.0035 | 89.6% | 90.0% | (-.14,.14) |
| B | X | 25 | 92.8% | 0.0045 | 92.6% | 93.0% | (-.18,.18) |
| B | S | 25 | 93.4% | 0.0035 | 93.2% | 93.6% | (-.14,.14) |
| B | F | 25 | 93.6% | 0.0040 | 93.4% | 93.8% | (-.16,.16) |
| C | N | 25 | 54.8% | 0.0050 | 54.6% | 55.0% | (-.20,.20) |
| C | C | 25 | 54.9% | 0.0045 | 54.7% | 55.1% | (-.18,.18) |
| C | X | 25 | 55.4% | 0.0040 | 55.2% | 55.6% | (-.16,.16) |
| C | S | 25 | 57.5% | 0.0055 | 57.3% | 57.7% | (-.22,.22) |
| C | F | 25 | 56.5% | 0.0045 | 56.3% | 56.7% | (-.18,.18) |
| D | N | 25 | 57.4% | 0.0060 | 57.1% | 57.7% | (-.24,.24) |
| D | C | 25 | 62.4% | 0.0060 | 59.4% | 65.4% | (-.24,.24) |
| D | X | 25 | 66.8% | 0.0055 | 66.6% | 67.0% | (-.22,.22) |
| D | S | 25 | 67.4% | 0.0065 | 67.1% | 67.7% | (-.26,.26) |
| D | F | 25 | 66.6% | 0.0055 | 66.4% | 66.8% | (-.22,.22) |
| E | N | 25 | 85.5% | 0.0030 | 85.4% | 85.6% | (-.12,.12) |
| E | C | 25 | 89.2% | 0.0035 | 89.0% | 89.4% | (-.14,.14) |
| E | X | 25 | 90.8% | 0.0030 | 90.6% | 91.0% | (-.12,.12) |
| E | S | 25 | 91.8% | 0.0035 | 91.6% | 92.0% | (-.14,.14) |
| E | F | 25 | 91.6% | 0.0035 | 91.4% | 91.8% | (-.14,.14) |

* GROUP :
A : SNR = 9 ; p = 0.4
B : SNR = 9 ; p = 0.7

C : SNR = 4 ; p = 0.4
D : SNR = 4 ; p = 0.7
E : SNR = 9 ; p = 0.55

* TYPE :
N : context free classification.
C : no look-ahead context classification.
X : one step look-ahead context classification.
S : two pass full context classification.
F : four pass full context classification.

**(5-2) Second simulation method :**

Another simulated data generating method proposed by P.H. Swain, Siegal and Smith (1979) is as follows: Use the ground truth ( or classification map ) and associated estimated mean vectors and covariance matrices of the classes (developed in performing the no-context classification). For each pixel, simulated data vectors are produced by a Gaussian random number generator having the same mean vector and covariance matrix as the class of associated with the pixle on the ground truth map. Thus the pixel in the simulated data set has the following characteristics:

(1)Each pixel in the simulated data set represents the same class as in the ground truth data.

(2)All classes have multivariate Gaussian distributions with parameters typical of those found in the ground truth data.

(3)All pixels measurement values are class-conditionally independent of adjacent pixels.

(4)There are no mixture pixels.

Data simulated in this manner are an idealization of real remote sensing data, but the spatial organization of the simulated data is consistent with a real world scene, and the overall characteristics of the data are consistent

with the contextual classifier assumption.

**(5-3) experimental results of real and simulated remote sensing images :**

The technique is first illustrated using a simulated image, which is generated from a digital remote sensing data collected by the Landsat MSS. The experimental data, which was a subset of the 13 April 1976 MSS scene of Roanoke, VA, was selected as the first study area. It was classified by a Bayes context free classification method in order to compare the results. The following ground cover classes were used:

(1) Class 1: Urban or Built-up Land

(2) Class 2: Agricultural Land

(4) Class 4: Forest land

(5) Class 5: Water ( Only a small amount )

(6) Class 6: Wetland ( Only a small amount )

(7) Class 7: Barren land

Because the study area was selected from the Roanoke, VA mountainous region (longitude from 79 52' to 80 00' W; latitude from 37 15' to 37 23' N), the land cover of this region is a complex pattern of diverse spectral classes occurring in small parcels. The most easily classified of this land cover class—open water—is not represented in this test area. Thus, this area is a difficult area for the conventional pixel independent classification

technique. The accuracy of context free classification for such remote sensing images, including Bayesian classifiers are 60%; AMOEBA (Bryant, 1979), and ISODATA (Duda and Hart 1973). Such accuracies are not unusual for scenes of this complexity.

The test image has four bands of digital multispectral data. The mean vectors $m_i$ and the covariance matrices $\Sigma_i$ for each class i are calculated. Then a simulated image having the approximate characteristics can be generated by the Gaussian Model.

As mentioned before, the key step of the contextual classification scheme we presented in this dissertation is minimizing expression (3.2) for two pass forward-backward algorithm.

$$P( C_{ij} \mid D_{IJ} ) = \frac{\dfrac{P( C_{ij} \mid D_{IJ}^{ij} )P( C_{ij} \mid E_{IJ}^{ij} )}{P( C_{ij})P( d_{ij} \mid C_{ij})}}{\dfrac{P( D_{IJ})}{P( D_{IJ}^{ij})P( E_{IJ}^{ij})}}$$

(3.6) for four pass algorithm, (3.11) for no look-ahead, and (3.13) for fixed look-ahead algorithm, respectively.

The prior probability of each class is calculated from the preclassification results. The transition probability $P( C_{ij} \mid C_{ij-1} , C_{i-1j})$ in the recursive equation 3.3 can be obtained from maximum likelihood esti-

mation or a "robust" estimation (See chapter 6). The conditional probability $P( d_{ij} \mid C_{ij})$ is estimated from the training sets and ground truth. In this experiment means and covariance matrices of each category were calculated from the ground truth data. The class conditional probability $P(d_{ij} \mid C_{ij})$ are assumed multi-dimensional normal:

$$P( d \mid C ) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \Sigma \mid^{\frac{1}{2}}} e^{-\frac{1}{2}( d-m)^T \Sigma^{-1} ( d - m )} \tag{5.4}$$

where n is the dimension of the feature space.

The context free and context classification results with both simulated and real multispectral remote sensing data are shown in Figure 5.6-5.23 and Table 5.2-5.3. The contingency table and classification images lead us to the conclusion that the context classification provides better results than the context free classification algorithm.

By visually examining these results, one can easily tell how good the performances are within each class, and also along the boundaries between classes. At first sight, we see that the Bayesian context-free classifiers results are quite noisy. The Markov context classification seems to "clean up" the picture significantly. It can be seen that many small isolated pixels are eliminated, and that each area is much more homogeneous in the contextual classification results. Boundaries remain correctly placed. The MSS four

band image, ground truth map, which had been classified by professional analysts, are given in Figure 5.8 and Figure 5.9, respectively. The above comparison and Fig 5.10 indicated that a 5 to 10% improvement of accuracy was obtained by the context classification method. Thus, in addition to the visual improvement, the context classification scheme also improves the classification accuracy.

The second study area is California. Three MSS classification results where sizes of subsets are 130 X 90, 101 X 70, 130 X 60 respectively, are shown in Figure 5.12 to 5.19. These results show that the algorithm was effective in several different areas with varied categories and preclassification accuracies (these areas had about 90% preclassification accuracy).

The third study area is a crop field at Clarke, Orgon (Fig 5.20-5.23). The Landsat MSS image is 12 band data set( Landsat MSS bands 4-7 from three dates). Thomas <1982> showed the accuracy of maximum likelihood classification and his canonical analysis method in the same study area is about 75%. Our context classification scheme raises the classification accuracy to 80.8%.

In order to study the effects of noise, independent zero-mean Gaussian noise $N(0,\sigma^2)$ is added to the 4-bands MSS simulated image at three different noise standard deviations 1, 2 and 3. Then, the noisy image is

classified by the pixel independent Bayes classifier, the dynamic programmingd approach for context classification, and the stochastic relaxation approach to context classification. The overall classification accuracy is measured as the ratio of the number of correctly classified pixels to the total number of classified pixels. It is plotted as a function of the noise standard deviation in Figure 5.26.

It can be see that the Bayes context-free classification is very sensitive to random noise, and that the context classification methods are quite opposite and are superior to the context free classifier.

In order to investigate the performance of seven different classifiers ( pixel independent Bayes' classifier, two-pass and four pass full context look-ahead, one step context look-ahead, no context look-ahead dynamic programming approaches, and stochastic relaxation context classification approach), we apply them to two real images ( subsets of Roanoke, VA and Clarke, OR). The comparitive results of overall classification accuracies are shown in Figure 5.27 and 5.28, respectively. On the basis of the real image experimental results provided, the conclusions are consistent with previous ones obtained in the simulated images, and they yield a significant improvement over the context free rule.

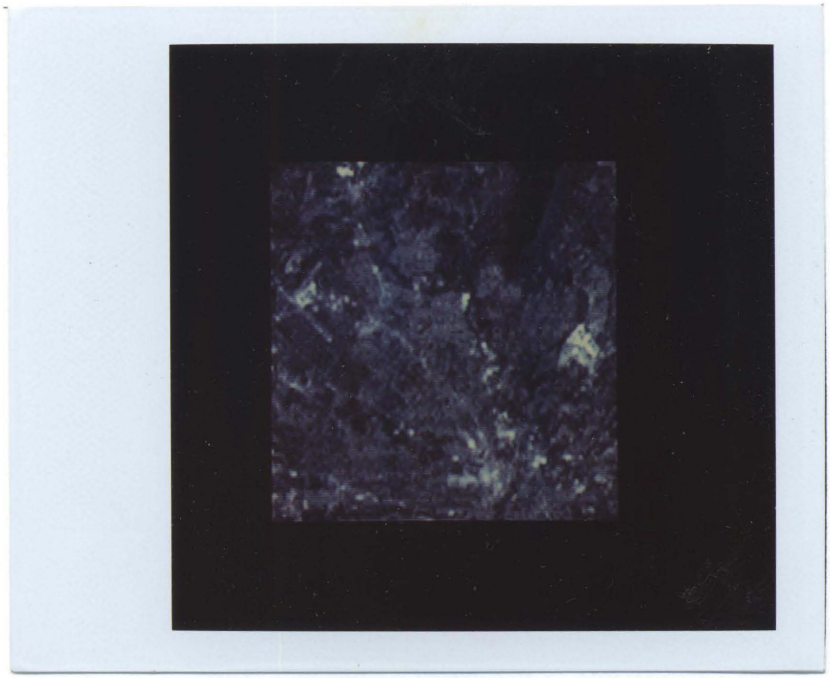

Figure 5.8 First band of MSS scene of Roanoke, VA. 13 April 1976, image size 151 X 151.

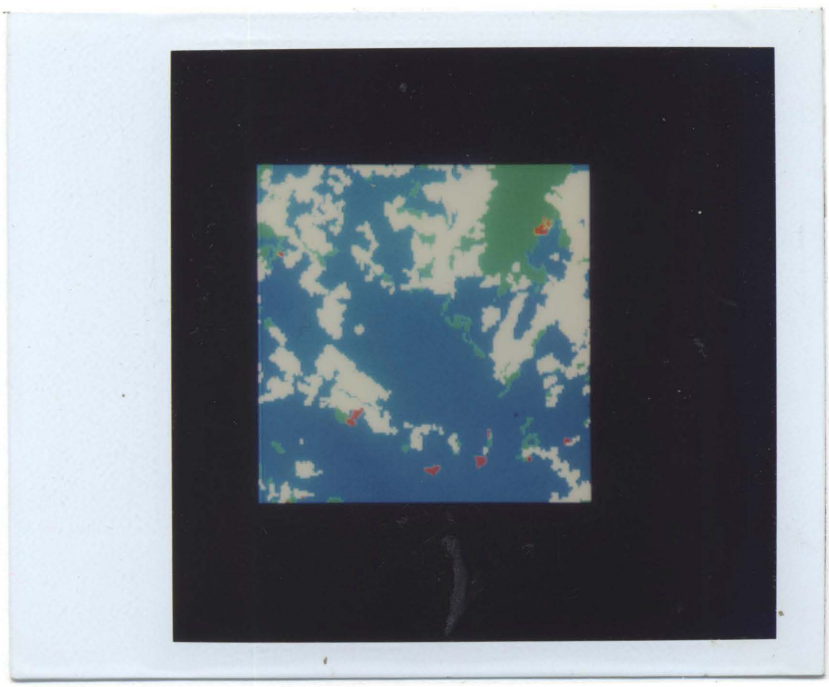

Figure 5.9 Ground truth image of Roanoke, VA. Blue (class 1) - urban or built-up land, white (class 2) - agricultural land,. green (class 4) - forest land.

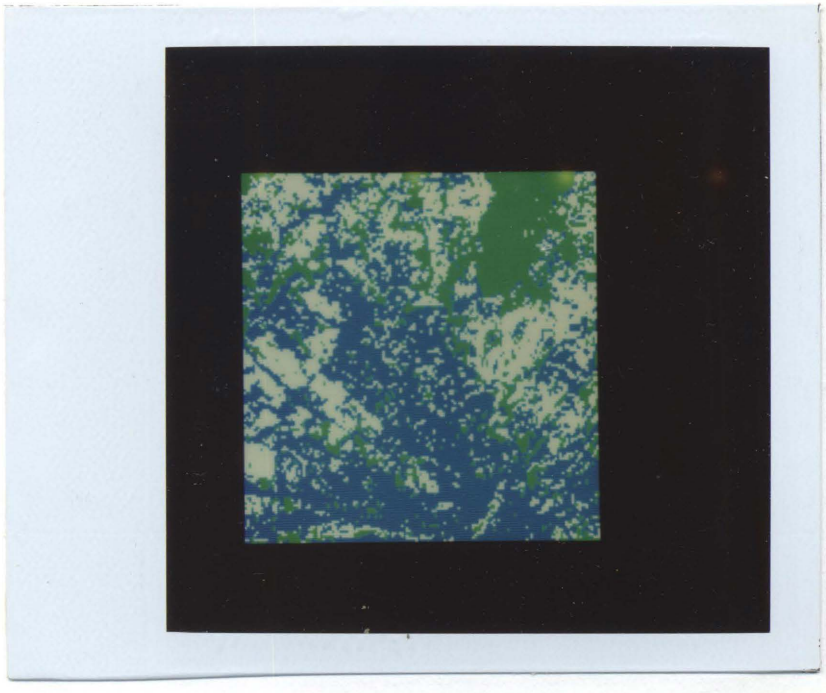

Figure 5.10 Bayes preclassification result of MSS, Roanoke, VA. Blue (class 1) - urban or buit-up land, white (class 2) - agricultural land,. green (class 4) - forest land.

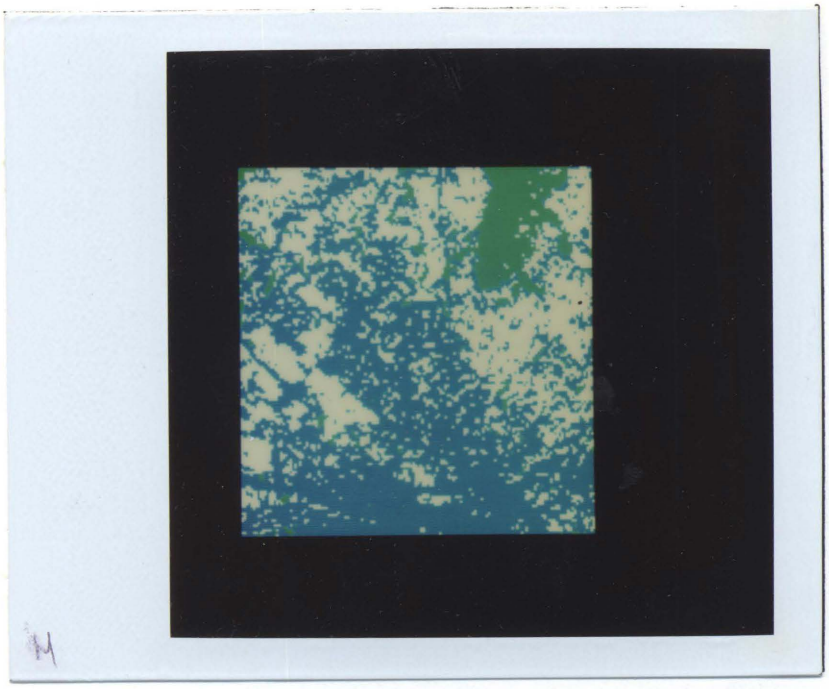

Figure 5.11 Markov contextual classification result of MSS, Roanoke, VA. Blue (class 1) - urban or buit-up land, white (class 2) - agricultural land,. green (class 4) - forest land.

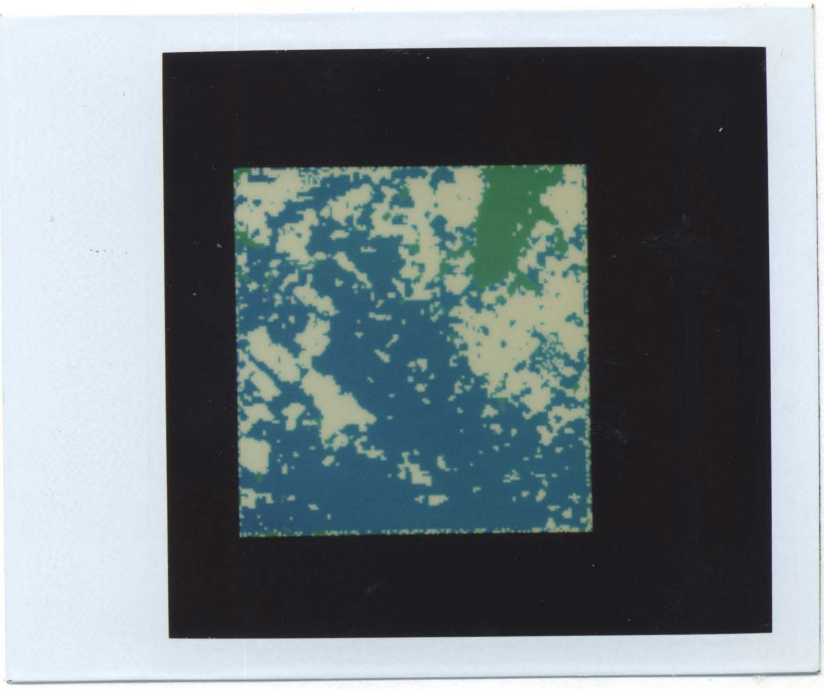

Figure 5.12 Context classification result by stochastic relaxation of MSS, Roanoke, VA. Blue (class 1) - urban or buit-up land, white (class 2) - agricultural land,. green (class 4) - forest land.

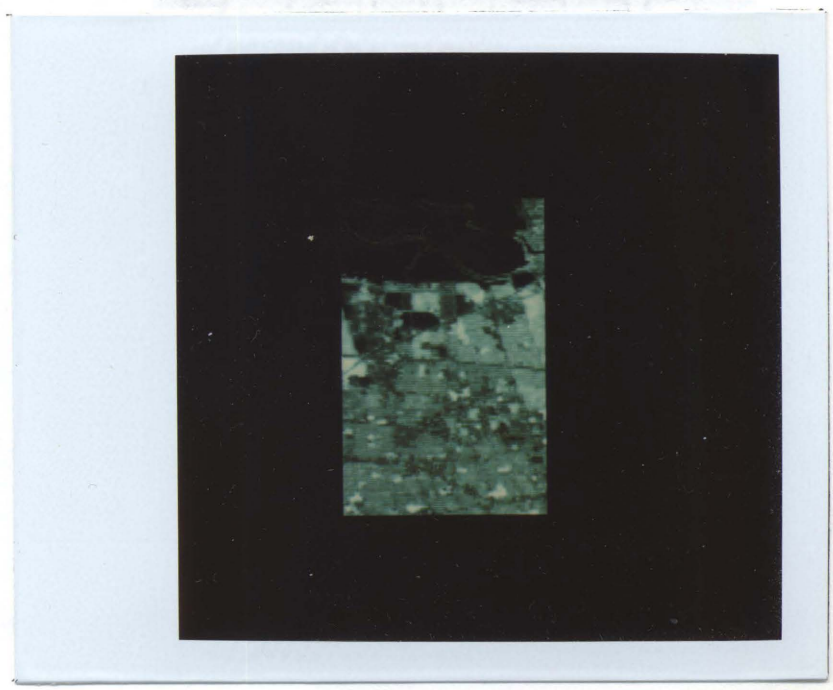

Figure 5.13 Third band of MSS scene of California (I): image size 130 X 90.

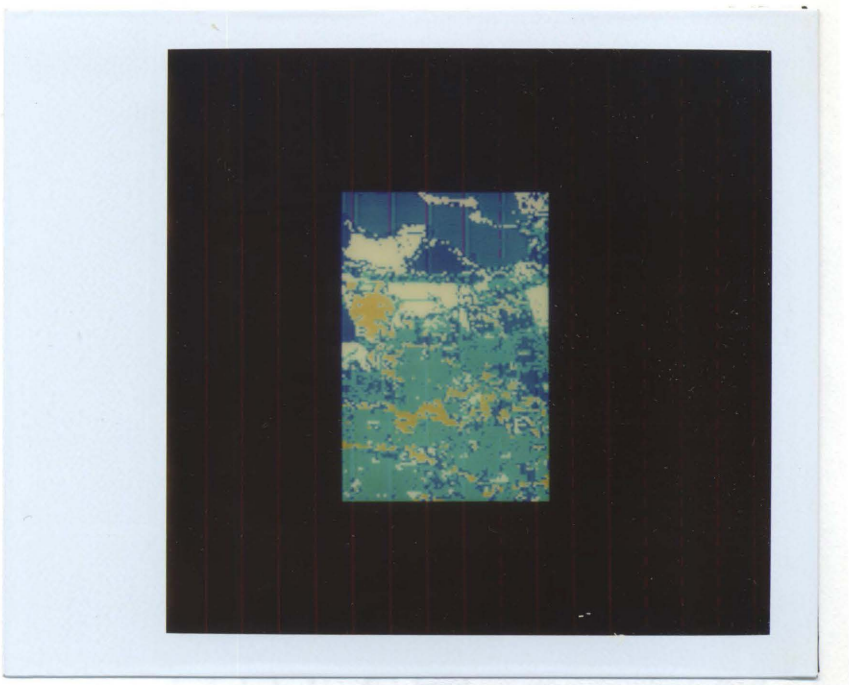

Figure 5.14 Bayes Preclassification result of MSS from California (I)

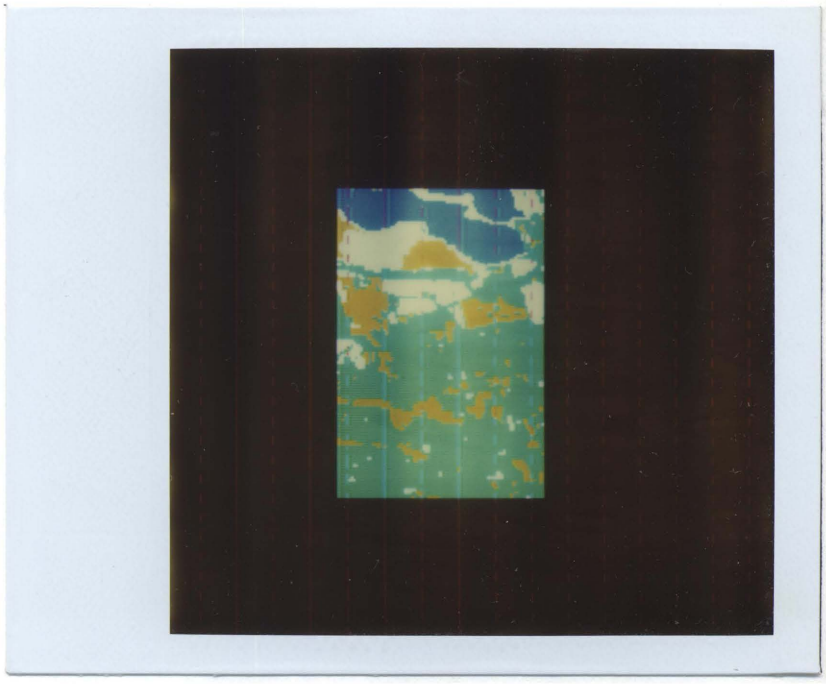

Figure 5.15 Markov contextual classification result of MSS from California (I)

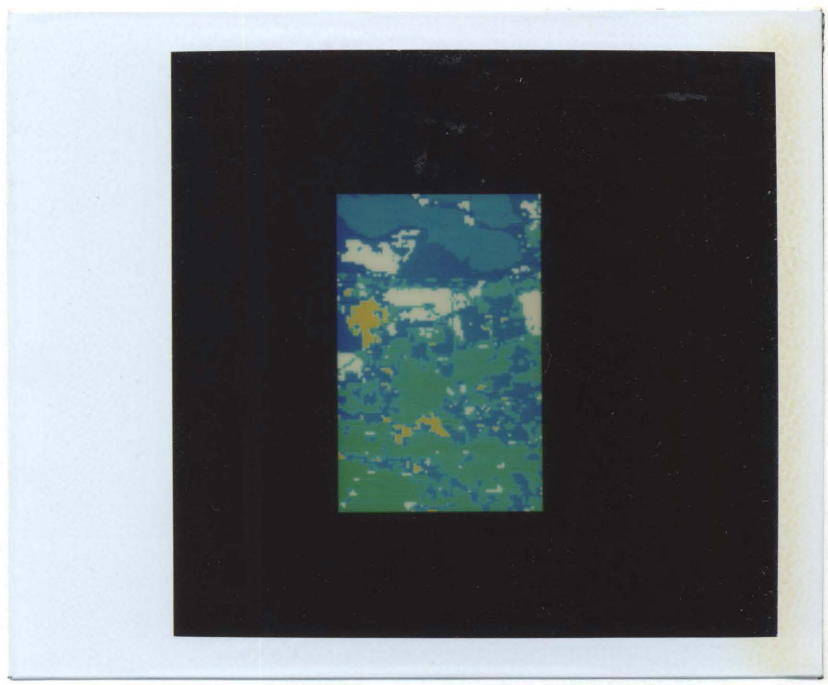

Figure 5.16 Stochastic relaxation classification result of MSS from California (I)

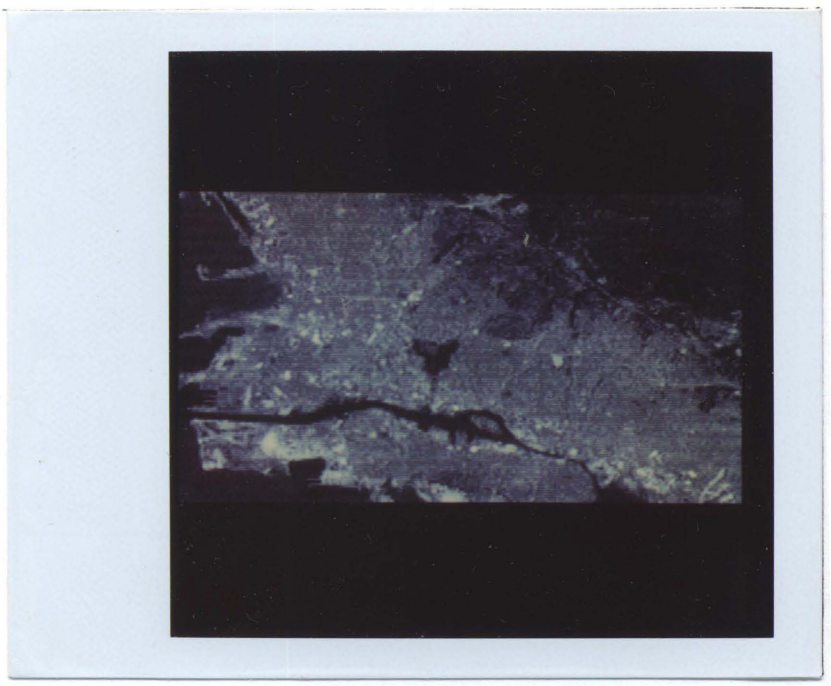

Figure 5.17 Third band of MSS scene of California (H): image size 110 X 70.

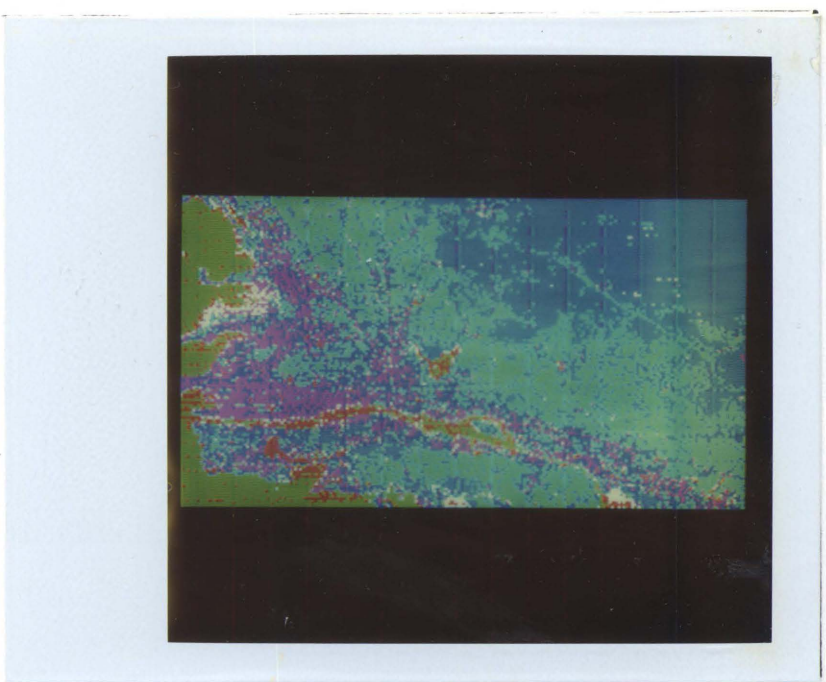

Figure 5.18 Bayes Preclassification result of MSS from California (H)

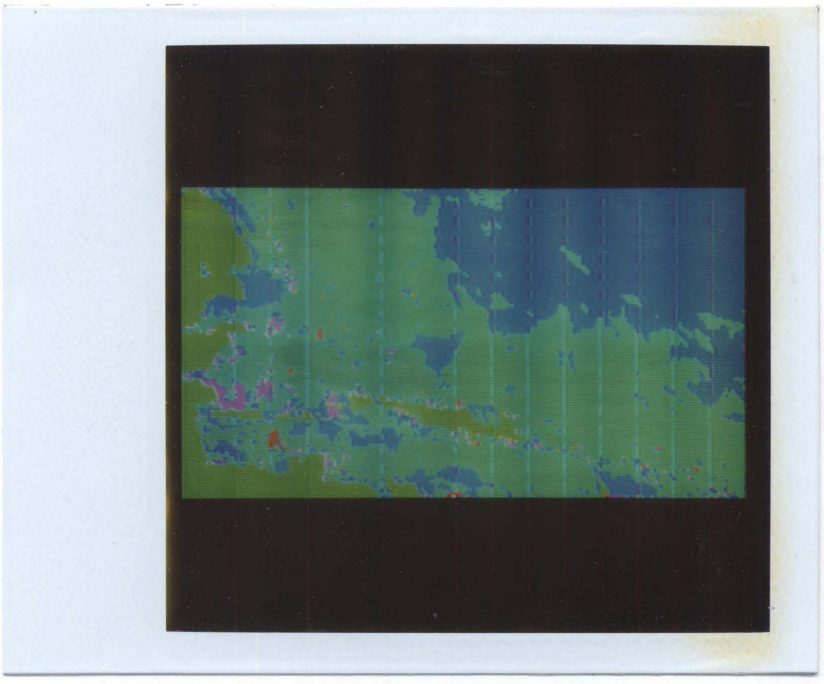

Figure 5.19 Markov contextual classification result of MSS from California (H)

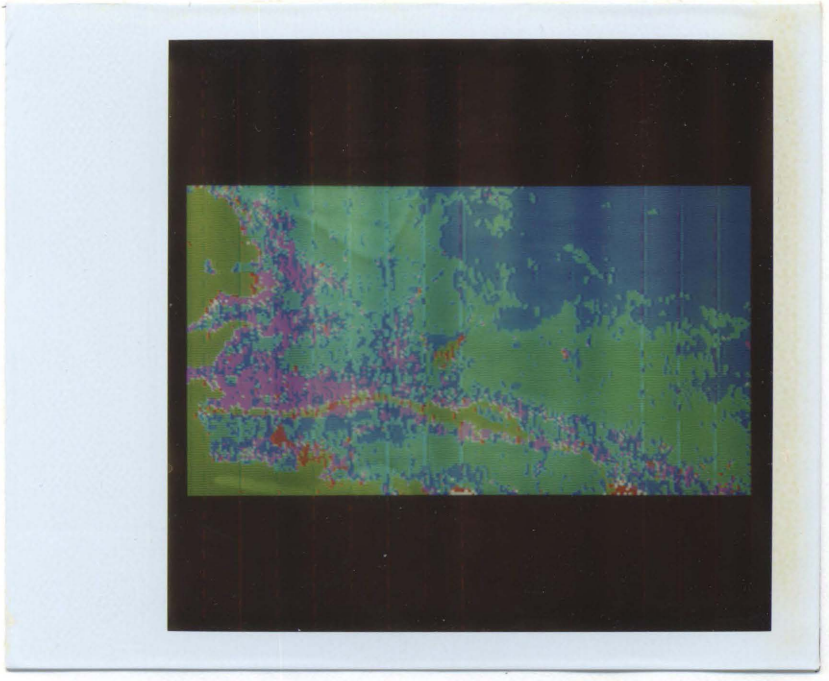

Figure 5.20 Stochstic relaxation context classification result of MSS from California (H)

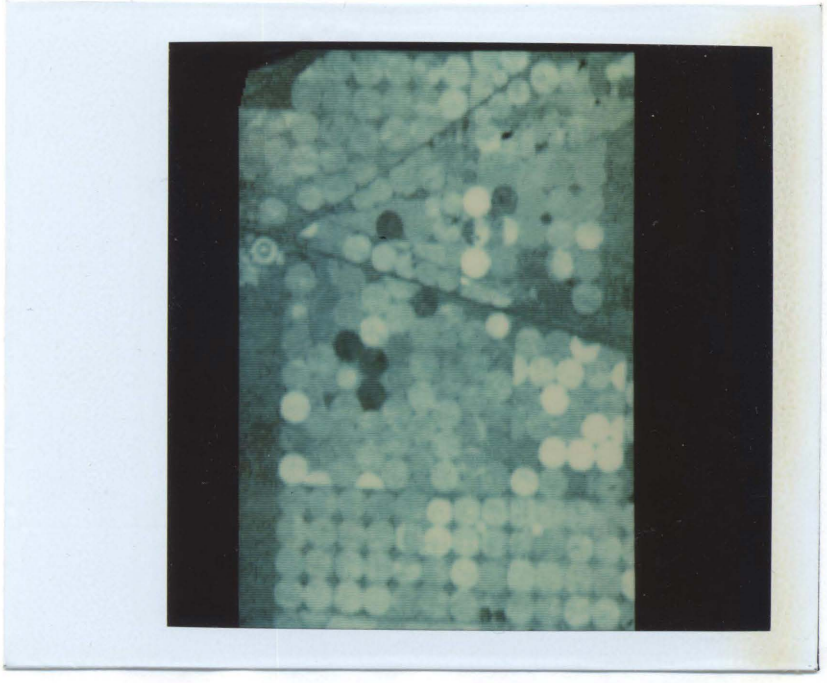

Figure 5.21 First band of MSS scene of crop field at Clarke, Oregon: 1982, image size 150 X 150.

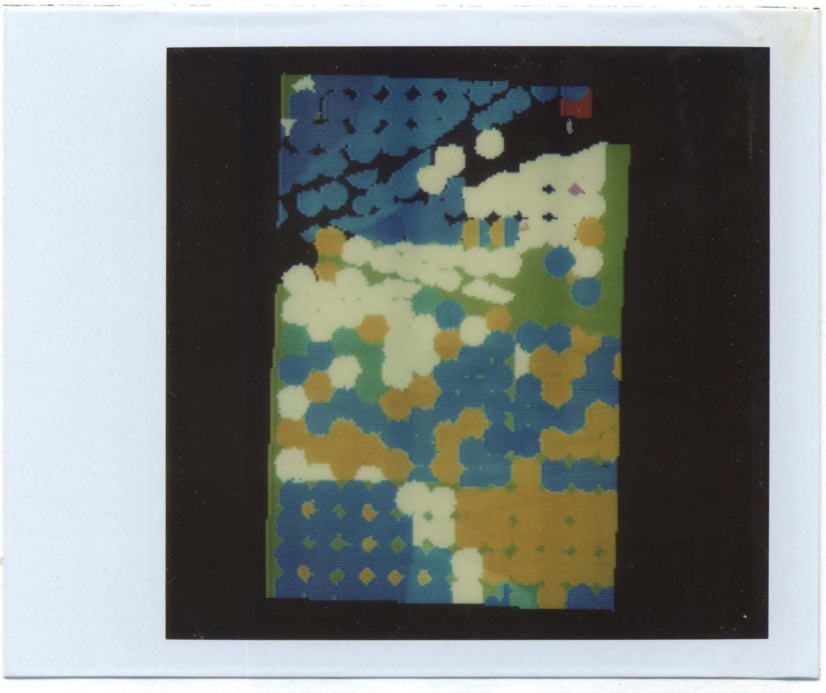

Figure 5.22 Ground truth image of crop field at Clarke, Oregon: 1982, image size 150 X 150.

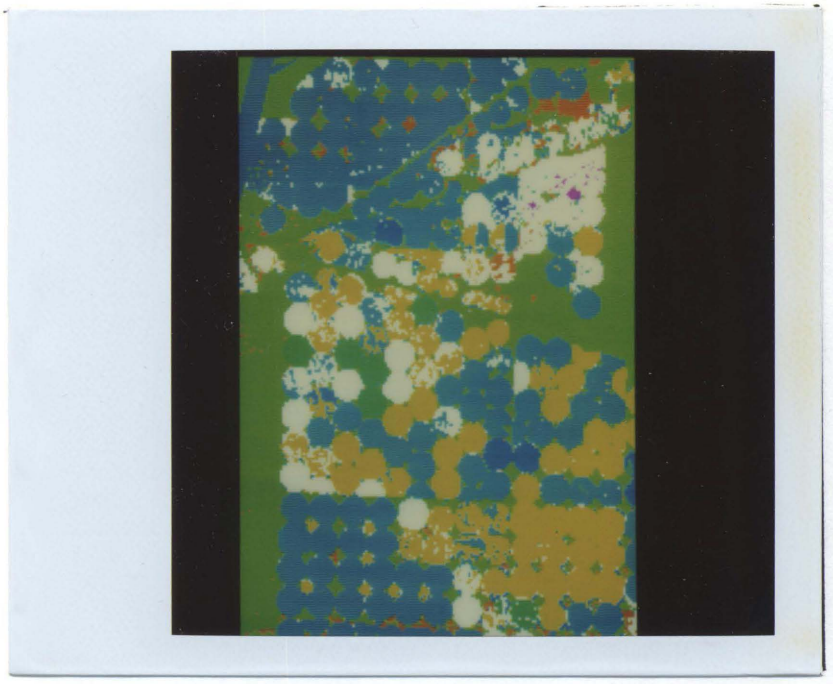

Figure 5.23 Bayes Preclassification result of MSS, Clarke, Oregon.

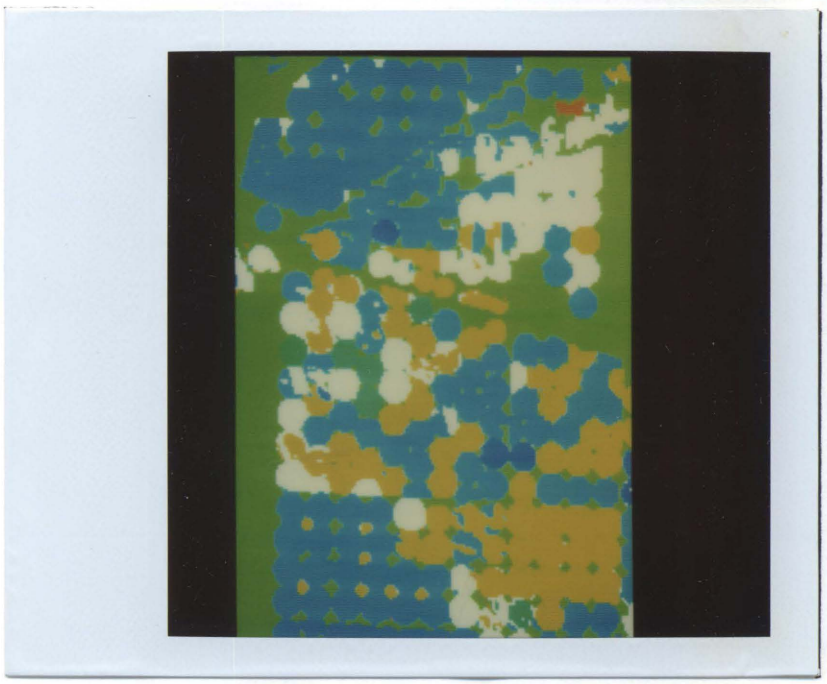

Figure 5.24 Markov contextual classification result of MSS, Clarke, Oregon.

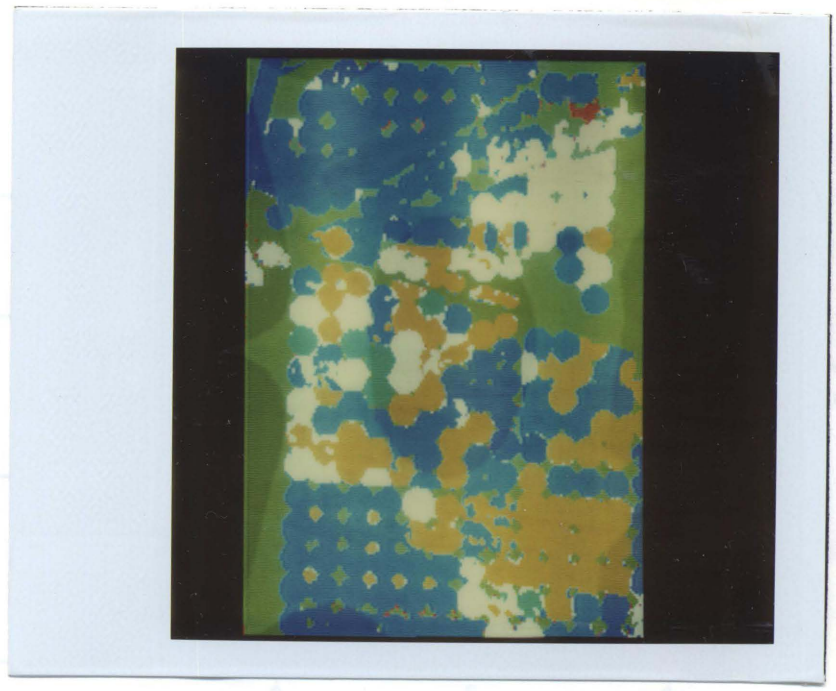

Figure 5.25 Contextual classification result by Stochstic relaxation, Clarke Oregon.

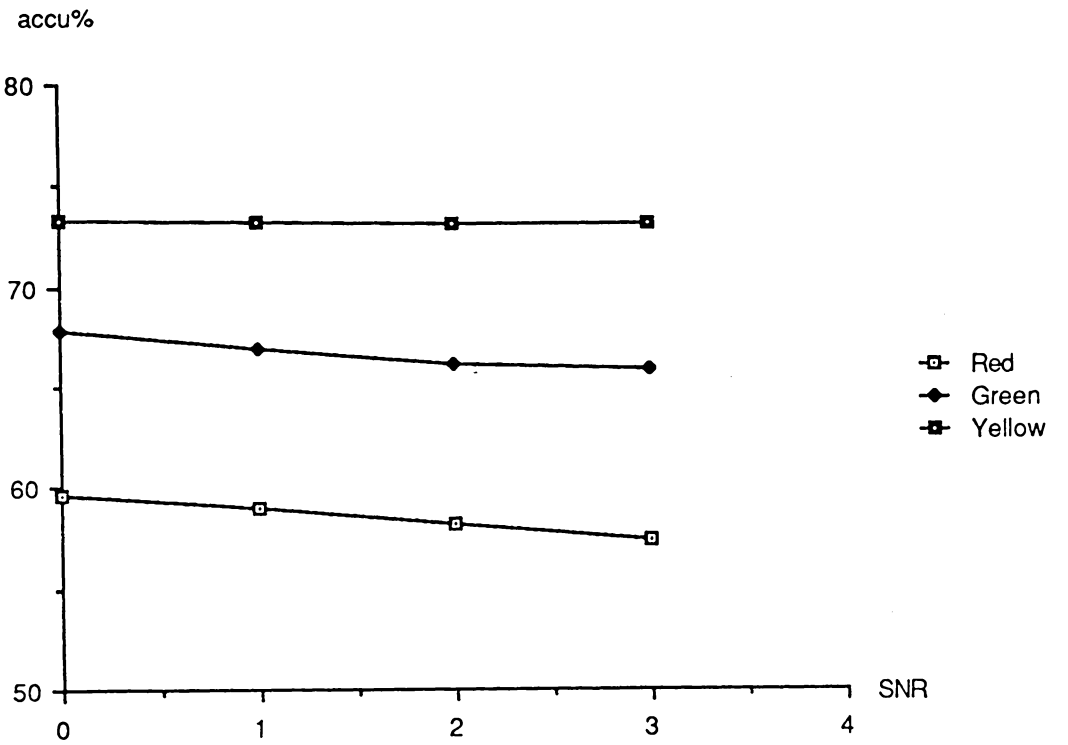


Figure 5.26 Overall classification accuracy curves vs noise level: Yellow line : Contextual classification by stochastic relaxation. Green line : Dynamic programming approach to context classification ( two pass forward-backward algorithm ) Red line: Pixel independent Bayes classification.

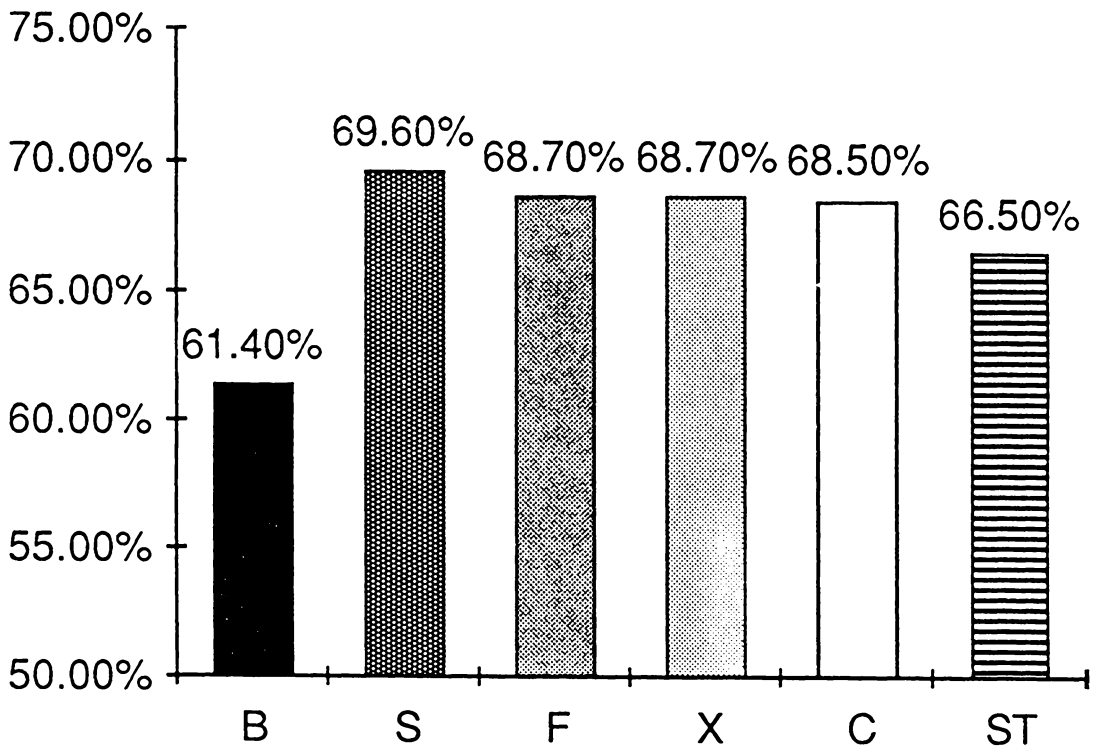


Figure 5.27 Comparison of overall classification accuracies of MSS scene of Roanoke, VA., using different classifiers. B : pixel indepedent Bayes classifies; S: two pass forward-backward look-ahead; F : four pass; X : one step context look-ahead; C : no context look-ahead; ST : stochastic relaxation.

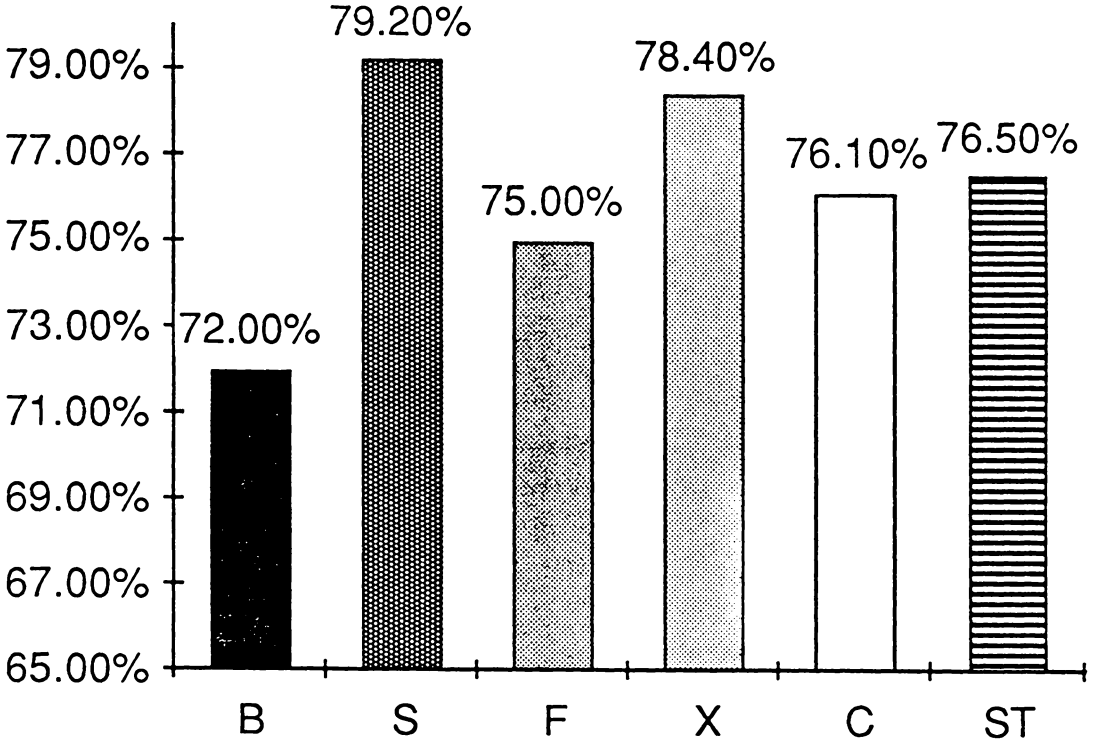


Figure 5.28 Comparison of overall classification accuracies of MSS scene of Clarke, OR., using different classifiers. B : pixel indepedent Bayes classifies; S: two pass forward-backward look-ahead; F : four pass; X : one step context look-ahead; C : no context look-ahead; ST : stochastic relaxation.

Table 5.2 Contingency tables for classification results of 13 April 1976 MSS scene of Roanke, VA. Scale factor of the number of pixels 10 ** 1 .

COL = assigned categories ROW = true categories

(A) Bayes pixel independent classification result

| CLASS | URB | AGR | RNG | FST | TOTAL | ACC(%)* |
|---|---|---|---|---|---|---|
| URB | 760 | 512 | 0 | 162 | 1437 | 52.8% |
| AGR | 116 | 379 | 0 | 83 | 578 | 65.6% |
| RNG | 0 | 0 | 0 | 0 | 0 | - |
| FSN | 15 | 28 | 0 | 210 | 253 | 83.0% |
| TOTA | 894 | 919 | 0 | 455 | 2268 | 59.6** |

(B) Context classification result using a dynamic programming approach

| CLASS | URB | AGR | RNG | FST | TOTAL | ACC(%)* |
|---|---|---|---|---|---|---|
| URB | 999 | 418 | 0 | 20 | 1437 | 69.5% |
| AGR | 183 | 371 | 0 | 24 | 578 | 64.2% |
| RNG | 0 | 0 | 0 | 0 | 0 | - |
| FSN | 58 | 28 | 0 | 167 | 253 | 66.0% |
| TOTA | 1240 | 817 | 0 | 211 | 2268 | 67.8%** |

(C) Stochastic relaxation Context classification result

| CLASS | URB | AGR | RNG | FST | TOTAL | ACC(%)* |
|---|---|---|---|---|---|---|
| URB | 1108 | 317 | 0 | 12 | 1437 | 77.1% |
| AGR | 179 | 385 | 0 | 14 | 578 | 66.6% |
| RNG | 0 | 0 | 0 | 0 | 0 | - |
| FSN | 50 | 35 | 0 | 168 | 253 | 64.4% |
| TOTA | 1337 | 737 | 0 | 194 | 2268 | 73.3%** |

Table 5.2 Contingency tables for classification results of 13 April 1976 MSS scene of Roanke, VA. Scale factor of the number of pixels 10 ** 1 . (continue )

* Classification accuracy.

** Overall classification accuracy : ratio of the number correctly classified pixels to the number of total classified pixels.

URB – Urban or built-up Land

AGR – Agricultural Land

RAN – Rangeland

FSN – Forest Land

Table 5.3 Contingency tables for classification results of test image "Clark". Scale factor of the number of pixels 10 ** 1 .

COL = assigned categories ROW = true categories

(A) Pixel independent Bayes classification result

| CLASS | WHT | ALF | POT | CRN | BNS | APL | PAS | RNG | TOTAL | ACC(%)* |
|---|---|---|---|---|---|---|---|---|---|---|
| WHT | 1017 | 47 | 30 | 5 | 4 | 0 | 10 | 75 | 1188 | 85.5% |
| ALF | 71 | 382 | 135 | 10 | 13 | 6 | 12 | 39 | 668 | 57.1% |
| POT | 40 | 32 | 522 | 5 | 19 | 0 | 2 | 32 | 652 | 84.6% |
| CRN | 1 | 5 | 1 | 65 | 2 | 0 | 0 | 4 | 78 | 83.3% |
| BNS | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 3 | 0% |
| PAS | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 2 | 11 | 81.1% |
| RNG | 15 | 12 | 14 | 2 | 4 | 1 | 9 | 335 | 392 | 85.4% |
| TOTAL | 1146 | 483 | 704 | 89 | 78 | 7 | 42 | 490 | 3040 | 77.5%** |

(B) Context classification result using a dynamic programming approach

| CLASS | WHT | ALF | POT | CRN | BNS | APL | PAS | RNG | TOTAL | ACC(%)* |
|---|---|---|---|---|---|---|---|---|---|---|
| WHT | 1073 | 26 | 26 | 1 | 11 | 0 | 10 | 601 | 1248 | 90.9% |
| ALF | 89 | 390 | 150 | 3 | 1 | 0 | 1 | 34 | 668 | 58.4% |
| POT | 58 | 29 | 534 | 2 | 6 | 0 | 0 | 23 | 652 | 81.9% |
| CRN | 1 | 5 | 1 | 68 | 0 | 0 | 0 | 4 | 79 | 86.1% |
| BNS | 1 | 6 | 2 | 1 | 36 | 0 | 0 | 3 | 49 | 73.5% |
| APL | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0% |
| PAS | 0 | 1 | 0 | 0 | 0 | 0 | 8 | 3 | 11 | 72.7% |
| RNG | 19 | 16 | 15 | 1 | 3 | 0 | 1 | 339 | 394 | 86.1% |
| TOTAL | 1681 | 605 | 777 | 82 | 58 | 0 | 23 | 1064 | 3040 | 80.5%** |

Table 5.3 Contingency tables for classification results of test image "Clark". Scale factor of the number of pixels 10 ** 1 .

COL = assigned categories ROW = true categories

(C) Stochastic relaxation Context classification result

| CLASS | WHT | ALF | POT | CRN | BNS | APL | PAS | RNG | TOTAL | ACC(%)* |
|---|---|---|---|---|---|---|---|---|---|---|
| WHT | 1080 | 23 | 25 | 1 | 1 | 0 | 0 | 58 | 1118 | 90.9% |
| ALF | 91 | 378 | 155 | 1 | 0 | 0 | 0 | 41 | 666 | 56.8% |
| POT | 54 | 23 | 544 | 1 | 3 | 0 | 0 | 29 | 654 | 83.2% |
| CRN | 1 | 5 | 1 | 65 | 0 | 0 | 0 | 6 | 78 | 83.3% |
| BNS | 2 | 5 | 2 | 0 | 35 | 0 | 0 | 4 | 48 | 73.9% |
| APL | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0% |
| PAS | 0 | 1 | 0 | 0 | 0 | 0 | 7 | 3 | 11 | 63.7% |
| RNG | 17 | 11 | 14 | 1 | 0 | 0 | 1 | 349 | 392 | 89.1% |
| TOTAL | 1643 | 573 | 787 | 72 | 47 | 0 | 10 | 1158 | 3040 | 80.8%** |

* Classification accuracy.
** Overall classification accuracy : ratio of the number correctly classified pixels to the number of total classified pixels.

WHT – Wheat
ALF – Alfalfa
POT – Potatoes
CRN – Corn
RNS – Beans
APL – Apples
PAS – Pasture ( irrigated )
RNG – Rangeland

## (5-4) Experiments in object detection using contextual information

:

Now we demonstrate the stochastic relaxation method to the application of object detection.

The experimental data are one band thermal images, so the additional contextual information becomes extremely valuable in the process. The maximum likelihood classification result is used as the preclassified labels.

As mentioned in chapter IV, at the first preclassification stage, two texture features (ie. entropy and the inverse difference moment) and the original measurement value are selected for the feature. Training samples are selected within homogeneous background regions and target regions to obtain mean values and covariances matrices. The transition probabilities are estimated from the preclassification result.

The test image ( Figure 5.29) is an 8-12 micron thermal image of size 200 X 200 taken at Grafenwoehr, Germany which contains one object - tank in the center. Figure 5.30 is a first stage segmentation result by using the pixel independent Bayes classifier. A second stage segmentation result using the contextual classification method is shown in Figure 5.31. We can see that the contextual classifier serves effectively again to eliminate the noise, and "smooth" the object boundaries. Within the homogeneous regions,

the contextual classification also produces better results by eliminating the misclassifed patterns. That is, the contextual method determines a better threshold for segmenting the original image.

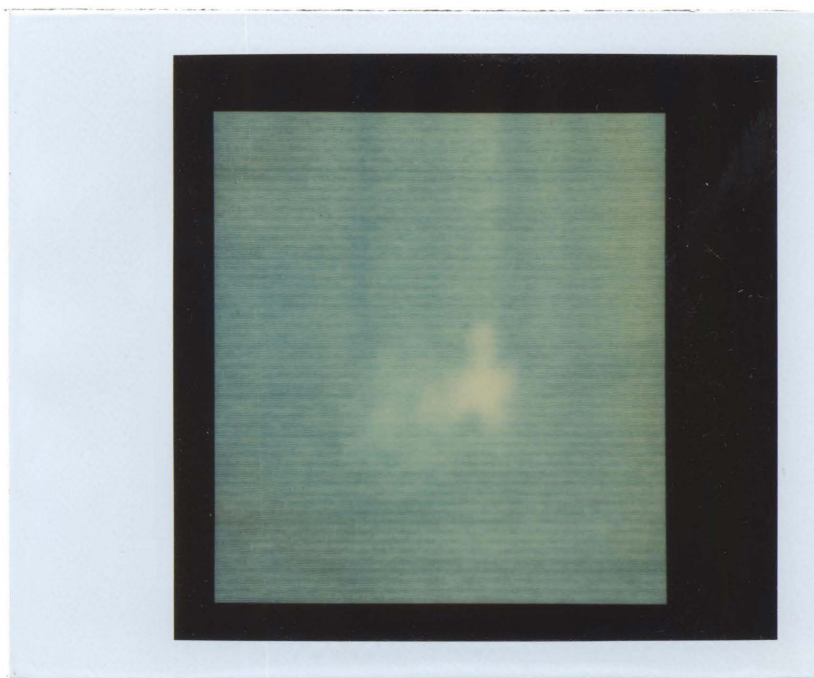

Figure 5.29 Experiment of segmentation using stochastic relaxation: The test image is 8- 12 micron thermal image of size 200 X 200 taken at Grafenwoehr, Germany. This image contains one object at the center.

Figure 5.30 Experiment of segmentation using stochastic relaxation: preclassification result of test image using Bayes pixel independent classification method. blue — object, black — background.

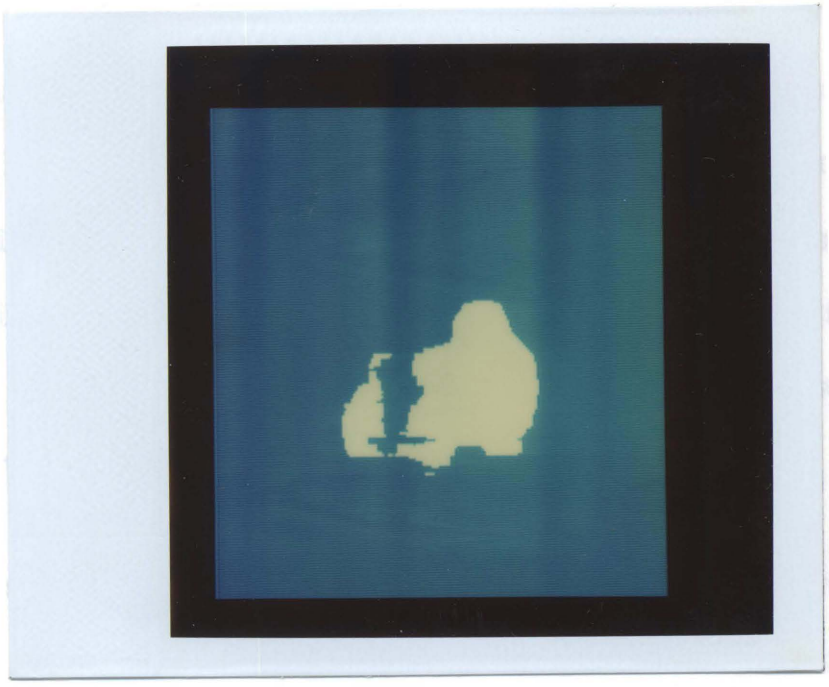

Figure 5.31 Experiment of segmentation using stochastic relaxation: contextual classification result of test image using stochastic relaxation method. blue — object, black — background.

## CHAPTER VI: ESTIMATION OF TRANSITION PROBABILITY

A very important problem using the Markov Random Field Model (MRF) in image processing and pattern recognition is how to estimate the MRF parameters. In chapter III and IV, we introduced two context classification schemes. The dynamic programming approach is based on a recursive procedure for optimal estimation of the state of a two-dimensional discrete Markov Random Field. Estimating the transition probabilities $P( C_{ij} \mid C_{ij+1} , C_{i+1j} )$ efficiently and accurately is an important step in utilizing the technique. In the stochastic relaxation scheme, the decision making minimizes the potential function. Here, also estimating the transition probabilities $P( C_{ij} \mid C_{ij-1} )$ and $P( C_{ij} \mid C_{i-1j} )$ is also an important aspect in utilizing the technique.

In this chapter we describe how to convert a 2-D Markov random field estimation problem into two 1-D problems, and how the maximum likelihood estimation is employed to estimation the parameter for each 1-D problem. Then the theoretical analysis and simulated experiments show that the maximum likelihood estimation technique produce unsatisfactory results when the image contains discontinuities such as edges and boundaries of regions. In order to solve this problem, a space-varying estimation method for transition probability is proposed in section 4. A technique for making sample size determinations for a desired parameter range at specified

confidence level is also discussed in this section. It guarantees the best compromise between increasing the number of observations and maintaining the homogeneity. Finally, simulated and real experimental results of the space varying estimation techniqure are shown in the section 5.

## (VI-1) Transition probability estimation problem in 2-D MRF

For a 2-D Markov Random Field, we have

$$P( C_{ij} \mid C_{ij-1} , C_{i-1j} ) = \frac{P( C_{ij-1} , C_{i-1j} \mid C_{ij} )P( C_{ij} )}{P( C_{ij-1} , C_{i-1j} )}$$

We will use an assumption employed in section III-2 to simplify the above expression. In our four neighbor system given the true interpretation of pixel ( i,j ) the categories of diagonal pixels are independent of each other.

$$P( C_{ij-1} , C_{i-1j} \mid C_{ij} ) = P( C_{ij-1} \mid C_{ij} )P( C_{i-1j} \mid C_{ij} )$$

$$P( C_{ij+1} , C_{i+1j} \mid C_{ij} ) = P( C_{ij+1} \mid C_{ij} )P( C_{i+1j} \mid C_{ij} )$$

then we have

$$P(C_{ij} \mid C_{ij-1}, C_{i-1j} ) = \frac{P( C_{ij-1} , C_{i-1j} \mid C_{ij} )P( C_{ij} )}{P( C_{ij-1} , C_{i-1j} )}$$

$$= \frac{P(C_{ij-1} \mid C_{ij} )P( C_{i-1j} \mid C_{ij} )P( C_{ij} )}{P( C_{ij-1} , C_{i-1j} )}$$

$$= \frac{\frac{P(C_{ij} \mid C_{ij-1})P(C_{ij-1})}{P(C_{ij})} * \frac{P(C_{ij} \mid C_{i-1j})P(C_{i-1j})}{P(C_{ij})} * P(C_{ij})}{P(C_{ij-1}, C_{i-1j})}$$

$$= \frac{P(C_{ij} \mid C_{ij-1})\, P(C_{ij} \mid C_{i-1j}) P(C_{ij-1}) P(C_{i-1j})}{P(C_{ij}) P(C_{ij-1}, C_{i-1j})}$$

$$P(C_{ij-1}, C_{i-1j}) = \sum_{C_{kl}} P(C_{kl}, C_{ij-1}, C_{i-1j})$$

$$= \sum_{C_{kl}} P(C_{i-1j}, C_{ij-1} \mid C_{kl}) P(C_{kl})$$

$$= \sum_{C_{kl}} P(C_{i-1j} \mid C_{kl}) P(C_{ij-1} \mid C_{kl}) P(C_{kl})$$

$$= \sum_{C_{kl}} \frac{P(C_{kl} \mid C_{i-1j}) P(C_{i-1j})}{P(C_{kl})} \frac{P(C_{kl} \mid C_{ij-1}) P(C_{ij-1})}{P(C_{kl})} P(C_{kl})$$

$$= P(C_{ij-1}) P(C_{i-1j}) \sum_{C_{kl}} \frac{P(C_{kl} \mid C_{ij-1}) P(C_{kl} \mid C_{i-1j})}{P(C_{kl})}$$

From above result, finally we have

$$P(C_{ij} \mid C_{ij-1}, C_{i-1j}) =$$

$$= \frac{P(C_{ij} \mid C_{ij-1}) P(C_{ij} \mid C_{ij-1})}{P(C_{ij}) \sum_{C_{kl}} \frac{P(C_{kl} \mid C_{ij-1}) P(C_{kl} \mid C_{i-1j})}{P(C_{kl})}} \qquad (6.1)$$

Equation (6.1) shows that the 2-D MRF transition probability $P(C_{ij} \mid C_{i-1j}, C_{ij-1})$ can be decomposed to two 1-D transition

probabilities $P( C_{ij} \mid C_{ij-1})$ and $P( C_{ij} \mid C_{i-1j})$. Thus the evaluation of a 2-D MRF transition probability becomes two 1-D problems.

**(VI-2) Notation :**

(1) S : designates number of classes.

(2) $m_{ij}$ : designates number of transitions from class i to class j. (transition count)

(3) $m_i$ : designates total number of transitions from class i.

(4) n : designates number of transitions

$$n = \sum_{i=1}^{s} \sum_{j=1}^{s} m_{ij}$$

(5) $P_{ij}$: designates an estimator of the true transition probability.

(6) $f( m_i , m_{ij} )$ : probability density function of $(m_i , m_{ij} )$

(7) $f( P_{ij} )$ : probability density function for the maximum likelihood estimator $P_{ij}$.

(8) $\tilde{f}(P_{ij} )$ : an approximation to $f(P_{ij})$.

(9) $\sigma_i^2$ : variance of $m_i$.

(10) $\sigma_{ij}^2$ : variance of $m_{ij}$.

(11) $\sigma_{i,ij}$ : covariance between $m_i$ and $m_{ij}$.

(12) $Y_{ij}$ : normalizing equation for transition counts.

(13) $Z_i$ : normalizing equation for the state occupacy counts.

(14) $\gamma_i^2$ : variance of $Z_i$.

(15) $\gamma_{ij}^2$ : variance of $Y_{ij}$.

(16) $\gamma_{i,ij}$ : covariance between $Z_i$ and $Y_{ij}$.

(17) $\delta(\hat{P}_{ij})$ : error term for $\tilde{f}(\hat{P}_{ij})$

(18) $p_i$ : steady-state probability.

$$p_i = \frac{p_{ij}}{p_{ij} + p_{ji}}$$

(19) $\overline{m}_i$ : $\overline{m} = n * p_i$ .

(20) $\overline{m}_{ij}$ : $\overline{m} = n * p_i * p_{ij}$.

Conventions:

(a) When a parameter has a superscript star, it designates true value. Parameters without superscripts are estimated values.

1. $P_{ij}^{*}$ : true value.

2. $\hat{P}_{ij}$ : estimated value.

(b) Capital letters denote random variables. Small letters denote a sample observation of the random variable.

## (VI-3) Maximum likelihood estimator:

Bartlett (1952) initiated work on the estimation problem by developing a maximum likelihood estimator for the transition probabilities. Using the notation described in the previous section, the Maximum likelihood estimation estimator for the transition probability can be denoted as follows :

$$\hat{P}_{kl}=\frac{m_{kl}}{m_k} \tag{6.2}$$

From this classical result, we know that $\hat{P}_{kl}$, the maximum likelihood estimation of $P( C_k | C_l )$, is equal to $\frac{m_{kl}}{m_k}$ . When $\sum_{k=1}^{s}\sum_{l=1}^{s} m_{kl} \rightarrow \infty$, $\hat{P}_{kl} = \frac{m_{kl}}{m_k} \rightarrow P_{kl}$ .

Most of digital image processing and contextual pattern classification algorithms are designed with the assumption that image processing is based on stationary spatial statistics. In view of the great variety of real world scenes, the stationary and homogeneous assumptions are not supposed to be exactly true. Thus, there is no guarantee that the estimation of statistical

parameters is still satisfactory and reliable for the implementation of the algorithms. However, we still hope that the minor error in the mathematical model should cause only a small error in the final conclusions. Unfortunately, this does not always hold.

The maximum likelihood estimation technique we used here is basically a global operation, and the result is a global statistic parameter ( ie., an average of transition count value through the entire image). When the image contains discontinuities such as edges and boundaries of regions, the assumption of homogeneity and the global operation produce unsatisfactory results.

To show an extreme example in our application problem, a three class pseudo-vandom Markov image is used as test image. The left half part of the simulated image has high transition probability from class k to class l; and low transition probability from class k to class l, (k,l = 1,2,3 and $k \neq l$ ).

$$P_{lf} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

but the right half part has a reverse situation.

$$P_{rt} = \begin{bmatrix} 0.25 & 0.375 & 0.375 \\ 0.375 & 0.25 & 0.375 \\ 0.375 & 0.375 & 0.25 \end{bmatrix}$$

The maximum likelihood estimation result is given by

$$P_{kl} = \begin{matrix} 0.525 & 0.238 & 0.238 \\ 0.238 & 0.525 & 0.238 \\ 0.238 & 0.238 & 0.525 \end{matrix}$$

Figure 6.2 shows the estimation curves of transition probability vs x-coordinate of the test image. From it, we can see that the maximum likelihood estimator introduces a large error into the estimation of transition probability matrix.

## (VI-4) Space-varying Estimation Method for Transition Probability:

Various researchers proposed estimation of local statistics for digital processing of nonstationary images. Some researchers (N.E. Nahi and Habibi 1975) considered partitioning the image into homogeneous regions. Another (S.A. Saralar and R.J. Defigweeidi 1981) use a continuously adaptive technique where the parameters are, during a scan, are updated based upon a local window. Wallis (1979) and Lee (1980) use local statistics for noise filtering.

In order to solve the local statistics estimation problem, we introduce a "robust" estimation method called a space-varying estimation technique, which should have an optimal or nearly optimal efficiency for the assumed model. It should be robust in the sense that smaller deviation from the model assumptions should impair the performance only slightly, and somewhat larger deviations from the model should not cause a catastrophe.

Let $p_{kl}(i,j)$ be transition probability from class k to class 1 at pixel (i,j). The proposed space-varying estimation method can be described in a general manner by an expression of the following form.

$$p_{kl}(i,j) = \frac{\sum_{(u,v)\epsilon N} W(i,j,u,v) d_{kl}(u,v)}{\sum_{(u,v)\epsilon N} W(i,j,u,v) d_{k}(u,v)} \qquad (6.3)$$

where N is the window from which the sample data are taken, and pixel (i,j) is a central point in the block, pixel (u,v) is a sampling point within the block window N, and $d_{kl}(u,v)$ called unit transition from class k to class 1 is a function of transition count. It has

$$d_{kl}(u,v) = \begin{cases} 1 & \text{if there is a state change from class k} \\ & \text{to class 1 at pixle (u,v)} \\ 0 & \text{otherwise} \end{cases}$$

$d_k(u,v)$ is an unit transition from class k. It has

$$d_k(u,v) = \begin{cases} 1 & \text{if there is a state change from class k} \\ & \text{to class 1 at pixle (u,v)} \\ 0 & \text{otherwise} \end{cases}$$

W(i,j,u,v) is a weighting function that quantifies the importance of pixel (u,v) with respect to the central pixel (i,j).

When W(i,j,u,v) = constant then $P_{kl}$ is position invariant, which is the common maximum likelihood estimation. When the weighting function has the form W(x-u,y-v), it becomes the increasingly popular M-estimators, which can be written as

$$p_{kl}(i,j) = \frac{\sum_{(u,v)\epsilon N} W(x-u,y-v)d_{kl}(u,v)}{\sum_{(u,v)\epsilon N} W(x-u,y-v)d_i(u,v)} \qquad (6.4)$$

where W(.) is usually a positive symmetric decreasing function. The M-estimators have good asymptotic properties and are robust with respect to a variety of distribution function.

The estimation procedure proposed here is one that fits into the general description of equation 6.3, but which has the feature that its computation is very simple, not requiring the solution of nonlinear equations. It is of the form

$$p_{kl}(i,j) = \frac{\sum_{(u,v)\epsilon N} W(u,v)d_{kl}(u,v)}{\sum_{(u,v)\epsilon N} W(u,v)d_{k}(u,v)} \tag{6.5}$$

where W(u,v) is a positive function, whose value is decreased by increasing the distance between pixel (u,v) and central pixel (i,j). For simplicity, we assume $W(u,v) = 1/D(u,v)$, where D(u,v) is the distance between pixel (u,v) and central pixel (i,j).

For the space-varying estimation technique an important problem is choosing a suitable block window size. In view of the sampling theory, to obtain a smaller confidence interval at a high confidence level, more observations are necessary. However, for a typical image, the local statistics of the image are position dependent. A large sample size will affect the accuracy of estimation of these local statistics. A compromise between increasing the number of observations and maintaining the homogeneity is desired.

In order to solve this problem we use the "projected sample determination technique" proposed by Young (1977) to create a nearly minimum size of estimation window, in which the space-varying estimation is guaranteed to satisfy a given accuracy. The technique for determining a projected sample size first required estimation of confidence intervals of transition probabilities.

The major steps and final results of Youngs' method for the confidence interval estimation and projected sample size determination are briefly shown in the following two sections, and the derivation of approximation expressions is described in the Appendix.

**(VI-5) Confidence Intervals estimation :**

Whittle (1955) derived a joint distribution for the transition counts of a Markov process. To develop a sampling distribution for a two-class Markov process, Young (1977) derived a bivariate normal approximation for Whittle's distribution using the normaling expressions from Smirnov (1966). This density function is expressed in terms of $m_i$ and $m_{ij}$, and is denoted by $f(m_i, m_{ij})$. To obtain a sampling ditribution for $\hat{p}_{ij}$, a ration distribution (ie. $\hat{p}_{ij} = \frac{m_{ij}}{m_i}$ is derived from $f(m_i, m_{ij})$) (see Appendix).

Young proved that under following conditions :

$$\text{If} \quad \frac{v - e}{1 - \rho} >> 0 \tag{6.6}$$

and

$$\frac{L^2(m_i) - e}{1 - \rho} >> 0$$

where

$$v = \frac{\overline{m}_i^{\,2}}{\sigma_i^{\,2}} - \frac{2\,\rho\,\overline{m}_i\,\overline{m}_{ij}}{\sigma_i \sigma_{ij}} + \frac{\overline{m}_{ij}^{\,2}}{\sigma_{ij}^{\,2}}$$

$$e = \frac{(1 - \rho)^2 (\overline{m}_i\,\hat{p}_{ij} - \overline{m}_{ij})^2}{(\sigma_{ij}^{\,2} - 2\,\rho\,\hat{p}_{ij}\,\sigma_i\,\sigma_{ij} + \sigma_i^{\,2}\,\hat{p}_{ij}^{\,2})}$$

$$L^2(m_i) = \frac{(m_i - \overline{m}_i)^2}{\sigma_i^{\,2}} - \frac{2\rho(m_i - \overline{m}_i)(m_i\hat{p}_{ij} - \overline{m}_{ij})}{\sigma_i \sigma_{ij}} + \frac{(m_i\hat{p}_{ij} - \overline{m}_{ij})^2}{\sigma_{ij}^{\,2}}$$

$$\overline{m}_i = n * p_i$$

$$\overline{m}_{ij} = n * p_i * p_{ij}$$

$$p_i = \frac{p_{ij}}{p_{ij} + p_{ji}}$$

$$\sigma_i^{\,2} = \frac{n p_{ij} p_{ji} (p_{ii} + p_{ij})}{(p_{ij} + p_{ji})^3}$$

$$\sigma_{i,ii}^{\,2} = \frac{2 * n\, p_{ii} p_{ij} p_{ii}}{(p_{ij} + p_{ji})^3}$$

$$\rho = \frac{2\sqrt{p_{ii}}}{\sqrt{p_{ii}+ p_{jj}(\, p_{ii}p_{jj}+ (\, 1+ p_{ji})^2)}} \tag{6.7}$$

The probability mass of $p_{ij}$ which lies between $t_1$ and $t_2$ has

$$P(t_1<\hat{p}_{ij}< t_2) = \frac{C}{\sqrt{2\pi}} \int_{u(t_1)}^{u(t_2)} e^{\frac{-x^2}{2}} dx \tag{6.8}$$

Where

$$u(t) = \frac{(\overline{m}_i t-\overline{m}_{ij})}{\sqrt{\sigma_{ij}^2-2\rho t\sigma_i\sigma_{ij} + \sigma_i^2 t^2}} \tag{6.9}$$

and C is determined a posteriori, so the probability p is unity. The equation (6.8) is easily calculated using the normal distribution scheme. From standard normal distribution approximation, we have

$$p(t_1<\hat{p}_{ij}<t_2) = 1-G(t_1,t_2) \tag{6.11}$$

and $G(t_1,t_2)$ is defined to be the confidence level for the confidence interval specified by $(t_1,t_2)$.

Let us make the following definition

$$\Phi(z_\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_\alpha} e^{\frac{-x^2}{2}} dx \tag{6.12}$$

Then we can obtain a $z_\alpha$ value such that

$$\Phi(z_\alpha) = \frac{1-G(t_1,t_2)}{2} \tag{6.13}$$

Substituting $z_\alpha$ for u(t) in expression 6.8, the value for $(t_1,t_2)$ can be obtained. Finally, the solution to 6.8 can produce upper/lower limits. If the uper/lower bounds correspond to the maximum/minimum limits for $\hat{p}_{ij}$, i.e. either zero or one, the problem should be resolved for the other limits, abandoning the attempt to locate symmetrical limits.

**(VI-6) s class case :**

Forthmore, Young extended the estimation algorithm from the two-class case to the s class case ( $s > 2$ ). The procedure is that we decompose the s class transition probability estimation matrix into a set of either two or three class transition matrixes, and then using the results of Section VI-5 a sampling distribution for $\hat{p}_{kl}$ will be produced.

By the Markov assumption, the transition from class k to class 1 is unaffected by any of the other classes. Using this property, we can group classes outside of class k and 1 together without affecting the interaction between class k and 1, that is, the transition matrix can be decomposed into a set of 3-class transition matrices without disturbing the statistical properties of those transitions not included in the actual compression.

For the s class case, we have transition matrix as

$$P = \begin{Vmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{Vmatrix} \qquad (6.14)$$

and the transition count matrix is

$$P = \begin{Vmatrix} M_{11} & M_{12} & \cdots & M_{1s} \\ M_{21} & M_{22} & \cdots & M_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ M_{s1} & M_{s2} & \cdots & M_{ss} \end{Vmatrix} \qquad (6.15)$$

Using the assumption we discussed above, the s class transition count matrix can be decomposed into $\binom{s}{2}$ three-class transition count matrices.

$$[M_r' ; r = 1,2 \cdots \binom{s}{2}] \qquad (6.16)$$

Each three-class transition count matrix $M'_r$ is

$$[M'] = \begin{vmatrix} m_{kk} & m_{kl} & m'_{13} \\ m_{lk} & m_{ll} & m'_{23} \\ m'_{31} & m'_{23} & m'_{33} \end{vmatrix} \qquad (6.17)$$

where k,l are the row and column indices of the s class transition count matrix, respectively, and k,l = 1,2 ...s ( $k \neq 1$ ).

The $m'_{13}$', $m'_{23}$, $m'_{31}$ , $m_{32}$, $m_{33}$ are defined as

$$m'_{13} = \sum_{j=1}^{s} m_{kj} \qquad (j \neq k,l)$$

$$m'_{31} = \sum_{i=1}^{s} m_{ik} \qquad (i \neq k,l)$$

$$m'_{23} = \sum_{j=1}^{s} m_{ij} \qquad (j \neq k,l)$$

$$m'_{32} = \sum_{j=1}^{s} m_{il} \qquad (i \neq k,l)$$

$$m'_{33} = \sum_{j=1}^{s} \sum_{i,j=1}^{s} m_{ij} \qquad (j \neq k,l) \qquad (6.18)$$

Young proved that density function $f(m_k, m_{kk})$ can be expressed as follows

$$f( m_k, m_{kk} ) = C' * e^{\frac{-q}{2}} \qquad (6.19)$$

where

$$C' = \frac{1}{2\pi\sqrt{\sigma_k^2 \sigma_{kk}^2 - \sigma_{k,kk}^2}}$$

$$q = [\, r,w \,]\, B_{11}^{-1}\, [\, r,w \,]^t$$

$$B_{11} = \begin{vmatrix} \sigma_{kk}^2 & \sigma_{k,kk} \\ \sigma_{k,kk} & \sigma_k^2 \end{vmatrix} \qquad (6.20)$$

$$[r,w] = \begin{vmatrix} (m_{kk} - \overline{m}_{kk}) \\ (m_{kl} - \overline{m}_{kl}) \\ (m_{lk} - \overline{m}_{lk}) \\ (m_{ll} - \overline{m}_{ll}) \\ (m_{k} - \overline{m}_{k}) \\ (m_{l} - \overline{m}_{l}) \end{vmatrix} \tag{6.21}$$

In a similar fashion $f(m_k,m_{kl})$, $f(m_l,m_{lk})$, and $f(m_l,m_{ll})$ can be obtained. Since each of these has a bivariate normal density function form (6.19), the method we discussed in the section VI-5 is directly applicable to develop confidential interval. Finally, the results of all members of the set [ $M'_r$ ; r =1,2 ...$(\frac{s}{2})$] are produced by the same technique. The experimental results will be shown in the section (VI-8).

(VI-7) Projected Sample Size Determination :

We should note that the confidence intervals estimation technique discussed in the previous section is under certain approximation condition (6.6). The approximation condition 6.6 usually does not hold for determining the confidence intervals of the transition probability estimations. However, this approximation still provides useful information for determining a sufficient sample size, which should be our new projected sample size required for a desired estimation accuracy.

This section we will present how to use the previous results to determine the projected sample size (Young, 1977).

If the confidence intervals for transition probability estimate obtained from the previous technique are wide, and we want to obtain a smaller confidence interval at a given confidence level, more observations are necessary. It is important to first determine the projected sample size required for the given confidence level.

Let us denote the projected number of observations by n'. From previous discussion, we know that $\overline{m}_i$, $\overline{m}_{ij}$ $\sigma_i^2$ , $\sigma_{ij}^2$ and $\sigma_{i,ij}$ increase as sampling size n increases. From the above facts, Young argued that equation (6.9) can be rewritten as

$$u(t) = \frac{\sqrt{n'}}{\sqrt{n}} = \frac{\overline{m}_i t - \overline{m}_{ij}}{\sqrt{\sigma_{ij}^2 - 2\rho\sigma_i\sigma_{ij}t + t^2\sigma_i^2}} \tag{6.22}$$

Solving (6.22) for n', we have

$$n' = \frac{n\, u^2(\sigma_{ij}^2 - 2\rho\sigma_i\sigma_{ij}t + t^2\sigma_i^2)}{(\overline{m}_i t - \overline{m}_{ij})^2} \tag{6.23}$$

The u(t) values are derived from a one-sided normal table such that

$$\Phi[u(t_1)] = 1 - \frac{(1-\frac{\alpha}{c_0})}{2}$$

$$\Phi[u(t_2)] = 1 - [-\frac{(1-\frac{\alpha}{c_0})}{2}] \tag{6.24}$$

Where $\alpha$ is confident level, and $c_0$ is determined by

$$\frac{1}{c_0} = \int_{u(0)}^{u(1)} \Phi(x)\, dx \tag{6.25}$$

and $(t_1,t_2)$ are desired range given by the user in our estimation problem.

Substitute these values into (6.23) for u and using the corresponding values $t_1$ and $t_2$, we get $n'_1$ and $n'_2$, respectively. Sometimes these two values are different because of the distribution's skewness, and this can interpreted to mean that substantially more observations are required to bring the lower limit up than to bring the upper limit down or vice versa. We will conservatively choose the larger one as the new projected sample size n'.

## (VI-8) Experiments results for space-varying estimation techniqure:

To understand the performance of the space-varying estimation technique, we examine its behavior on simulated data sets and on real remote sensing images. Subsection (VI-8.a) and (VI-8.b) first illustrate the performance of confidence intervals and projected sample size determination for two-class case and multi-class case, respectively. Subsection (VI-8.c) illustrates the performance of contextual stochastic relaxation approach using both space-varying estimation and space-invarying estimation methods.

### (VI-8.a) Simple 2-class simulated experiment:

This section describes a simple two-class simulated experiment, which demonstrates how the estimation procedure works and how the confidence intervals and projected sample size are determined.

A pseudo-random Markov image with two-categories is used for our simulated data experiments. The simulated image is generated the same way as in chapter III.

The transition matrix $p_{ij}^{*}$ :

$$p_{ij}^{*} = \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix}$$

Initially, the size of image we used is 20 x 20, (ie. n = 400), and the transition count matrix, which is obtained from measuring the simulated data, is

$$M = \begin{bmatrix} 71 & 116 \\ 111 & 63 \end{bmatrix}$$

From the Equation 6.2, We get maximum likelihood estimation as

$$\hat{p}_{ij} = \begin{bmatrix} 0.379 & 0.62 \\ 0.63 & 0.36 \end{bmatrix}$$

the corresponding steady-state probabilities

class 1: $p_1 = 0.88$

class 2: $p_2 = 0.12$

To compute confidence interval we use Equations in sections VI-5 and VI-7.

for $\hat{p}_{11}$ :

$\hat{\bar{m}}_1 = 187$

$\hat{\sigma}_1^2 = 53.19$

$\hat{\bar{m}}_{11} = 71$

$\hat{\sigma}_{11}^2 = 76.79$

$\hat{\rho}_1 = 0.85$

and for $\hat{p}_{22}$ :

$\hat{\bar{m}}_2 = 174$

$\hat{\sigma}_2^2 = 53.19$

$\hat{\bar{m}}_{22} = 111$

$\hat{\sigma}_{22}^2 = 71.74$

$\hat{\rho}_2 = 0.84$

Using the technique we discussed in the previous sections, we have confidence intervals for confidence level $\alpha = 0.5$:

$\hat{p}_{11}$: $P( 0.30 < \hat{p}_{11} < 0.45 ) = 0.5$ and $\hat{p}_{11} = 0.38$

$\hat{p}_{12}$: $P( 0.57 < \hat{p}_{12} < 0.70 ) = 0.5$ and $\hat{p}_{12} = 0.62$

$\hat{p}_{21}$: $P( 0.57 < \hat{p}_{21} < 0.72 ) = 0.5$ and $\hat{p}_{21} = 0.63$

$\hat{p}_{22}$: $P( 0.28 < \hat{p}_{22} < 0.43 ) = 0.5$ and $\hat{p}_{22} = 0.36$

Their ranges are:

$\hat{p}_{11}$: ( 0.08, 0.07 )

$\hat{p}_{12}$: ( 0.07, 0.08 )

$\hat{p}_{21}$: ( 0.07, 0.08 )

$\hat{p}_{22}$: ( 0.08, 0.07 )

Obviously, the confidence intervals do not satisfy the given confidence level. However, this approximation still provides useful information for

determining a sufficient sample size.

In order to determinant the projected sample size, we choose a range from - 0.05 to + 0.05 at the $\alpha = 0.5$ level, and t = ($\hat{p}_{11}$ + 0.05, and $\hat{p}_{11}$- 0.05). Substitute the values $m_i$, $m_{ij}$, $\sigma_i$, $\sigma_{ij}$. into equation (6.23).

We have n' = 900 for $t_1$ , and n'= 723 for $t_2$.

For the sake of satisfaction of the desired accuracy, the next step is to increase the sample size such that n = 900. Then we have the new values as follows

$$M = \begin{bmatrix} 179 & 252 \\ 252 & 158 \end{bmatrix}$$

$$P = \begin{bmatrix} 0.415 & 0.585 \\ 0.614 & 0.385 \end{bmatrix}$$

and

| for $\hat{p}_{11}$ | for $\hat{p}_{22}$ |
|---|---|
| $\hat{\bar{m}}_1 = 431$ | $\hat{\bar{m}}_2 = 410$ |
| $\hat{\sigma}_1^2 = 140.3$ | $\hat{\sigma}_2^2 = 140.28$ |
| $\hat{\bar{m}}_{11}^2 = 179$ | $\hat{\bar{m}}_{22}^2 = 158$ |
| $\hat{\sigma}_{11}^2 = 201.3$ | $\hat{\sigma}_{22}^2 = 180.35$ |

$\hat{\rho} = 0.87$ $\hat{\rho} = 0.85$

Now the approximation (6-20) is applicable in determining the $\hat{p}_{11}$ and $\hat{p}_{22}$ confidence intervals, and they are

$\hat{p}_{11}$: $P( 0.37 < \hat{p}_{11} < 0.46 ) = 0.5$ and $\hat{p}_{11} = 0.41$

$\hat{p}_{12}$: $P( 0.54 < \hat{p}_{12} < 0.63 ) = 0.5$ and $\hat{p}_{12} = 0.59$

$\hat{p}_{21}$: $P( 0.57 < \hat{p}_{21} < 0.66 ) = 0.5$ and $\hat{p}_{21} = 0.61$

$\hat{p}_{22}$: $P( 0.34 < \hat{p}_{22} < 0.43 ) = 0.5$ and $\hat{p}_{22} = 0.39$

Their ranges are:

$\hat{p}_{11}$: ( 0.05, 0.05 )

$\hat{p}_{12}$: ( 0.04, 0.05 )

$\hat{p}_{21}$: ( 0.05, 0.05 )

$\hat{p}_{22}$: ( 0.05, 0.04 )

This experimental result demonstrates that the technique works well in an extreme case allowing us to determine confidence intervals given n transitions, and to determine the projected sample size required for a desired estimation accuracy.

**(VI-8.b) Simulated experiment for the multi-class case :**

In this subsection we present a simulated experimental result to illustrate how the decomposition technique and estimation procedure work for the multi-class case. The simulated image is a pseudo-random Markov image with 4-categories. The transition matrix of the simulated image is

$$P_{ij} = \begin{bmatrix} 0.40 & 0.20 & 0.20 & 0.20 \\ 0.20 & 0.40 & 0.20 & 0.20 \\ 0.20 & 0.20 & 0.40 & 0.20 \\ 0.20 & 0.20 & 0.20 & 0.40 \end{bmatrix}$$

Initially, we take image size 20 X 20 (n=400). The transition count matrix is

$$M_{ij} = \begin{bmatrix} 30 & 11 & 18 & 20 \\ 8 & 34 & 18 & 21 \\ 19 & 17 & 30 & 27 \\ 19 & 29 & 29 & 43 \end{bmatrix}$$

and the corresponding maximum-likelihood estimation of transition probability is

$$P_{ij} = \begin{bmatrix} 0.385 & 0.140 & 0.231 & 0.244 \\ 0.100 & 0.420 & 0.222 & 0.259 \\ 0.165 & 0.183 & 0.320 & 0.290 \\ 0.174 & 0.204 & 0.266 & 0.394 \end{bmatrix}$$

A four-class transition matrix can be decomposed into six three-class transition matrixes.

From Equations 6.16-6.18 the decomposed transition count matrixes are

$$M_1 = \begin{matrix} 30 & 11 & 37 \\ 8 & 34 & 39 \\ 38 & 35 & 129 \end{matrix} \qquad M_2 = \begin{matrix} 30 & 18 & 30 \\ 19 & 30 & 44 \\ 27 & 47 & 116 \end{matrix}$$

$$M_3 = \begin{matrix} 30 & 19 & 29 \\ 19 & 43 & 47 \\ 27 & 48 & 99 \end{matrix} \qquad M_4 = \begin{matrix} 34 & 18 & 29 \\ 17 & 30 & 46 \\ 29 & 47 & 111 \end{matrix}$$

$$M_5 = \begin{matrix} 34 & 21 & 26 \\ 18 & 43 & 48 \\ 28 & 46 & 97 \end{matrix} \qquad M_6 = \begin{matrix} 30 & 27 & 36 \\ 29 & 43 & 37 \\ 36 & 40 & 83 \end{matrix}$$

Using the technique we discussed in the section (VI-6), we have the results shown in table 6-1.

From Table 6-1 it is observed that the transition probabilities $\hat{p}_{ij}$ have large ranges, which are unsatisfying for the specified accuracy. Since the sample size projections require using the estimates for the actual values, the sample size projections are made assuming that more transitions are necessary to reduce the confidence interval ranges of the transition matrices. In order to improve the estimation accuracy, we should increase the sample size again.

We will choose the largest projected sample size in the Table 6-1 as the new observation size for the following reason : If the largest projected sample size value is chosen then since the $\hat{p}_{ij}$ variances are decreasing functions of n, all the other estimates will have a range less than or equal to the

confidence interval at this largest projected sample size.

Consequently, we have new window size 45 X 45 ( n' = 2025), and its results are shown in the Table (6-2). From this table we see that all estimates are within the desired range at the 0.05 confidence level. The experimental result demonstrates that the technique also works well for the multi-class case. In the next section, we will show how to use the projected sample size determination technique in the space-varying estimation method.

### (VI-8.c) Experimental Results For space varying Estimation Technique

As mentioned in section VI-2, for some extreme cases such as the simulated image in Figure 6-2, a large estimation error is introduced when using the common maximum likelihood estimator.

But the space-varying estimation technique can achieve better classification results. In order to show the performance of the "space-varying" estimation, a simulated image of size 50 X 50 pixels is used as first test image. It consists of two equal size pseudo-random Markov images. (Figure 6-2)

The left part of image has transition probability matrix $p_1$ as

$$P_l = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

and the right part of image

$$P_r = \begin{bmatrix} 0.25 & 0.375 & 0.375 \\ 0.375 & 0.25 & 0.375 \\ 0.375 & 0.375 & 0.250 \end{bmatrix}$$

The maximum likelihood estimation result is given by:

$$P_{ij} = \begin{bmatrix} 0.525 & 0.238 & 0.238 \\ 0.238 & 0.525 & 0.238 \\ 0.238 & 0.238 & 0.525 \end{bmatrix}$$

The "space-varying" estimations for each pixel are calculated from the equation (6.5), and the numerical results are shown in the figure 6.3.

From the figure, it is observed that the estimation accuracy using the robust technique is much better than maximum-likelihood estimation for most pixels, except for those pixels near the boundary.

Then we apply the stochastic relaxation contextual classification method to the test image using transition probability from both the common maximum likelihood estimator and the "space-varying" estimator. Tables 6.3 shows the contingency tables for classification results of the pixel independent Bayes' context free classification, and the stochastic relaxation contextual classification methods using the maximum likelihood

transition probability estimation and the "space-varying" estimation, respectively. They indicated that the contextual classifier based on the maximum likelihood estimation gained a 1.0 % increase in overall classification accuracy over the pixel independent Bayes' classifier, and that the "space-varying" contextual classification gained a 2.3% increase over the pixel independent Bayes' classification. Obviously, the accuracy improvement of a "space-varying" estimation technique is significant.

In order to examine the behavior of the technique, we use the Landsat MSS data, which was a subset of the 13 April 1976 MSS scene of Roanoke, VA., and has been used in chapter V. From the previous section we konw that the local statistics of the image (eg. transition probability matrix) are position depedent. A large sample size will affect the accuracy of estimation of these local statistics. But to obtain a smaller confidence interval at a high confidence level, more observations are neccessary. In order to find a possible minimum sample size under a given confidence interval at the given confidence level, we use the technique discussed in section V-8 to determine the projected sample's size. Then the minimum size of block window in which the number of pixel is equal to the projected number of observations is created for the space-varying estimation technique. For this test image, a window size of 11 x 11 is satisfied to the confidence interval ( -0.1, +0.1 ) at the given confidence level of 0.05. The experimental results, shown in table 6.4, indicate that the space-varying estimation technique can improve

the classification accuracy, especially for discontinuous images. (see Figure 6.4)

The second real test image is a subset of MSS image for Clarke Oregon. The overall classification accuracies of pixel independent classifier and stochastic relaxation context classifiers using the maximum likelihood transition probability estimation and "space-varying estimation" are shown in table 6.5. Examining the results in table 6.5, it can again be seen that the context classification using "space-varying" estimation is superior to the maximum - likelihood estimation. (See Figure 6.5)

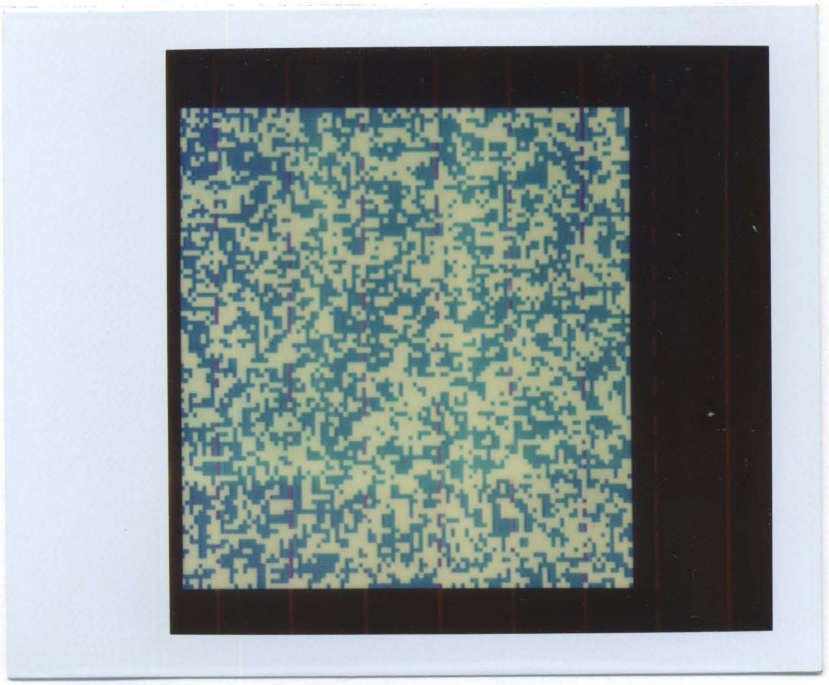

Figure 6.1 2-class simulated experiment of transition probability estimation: This simulated image is a pseudo-random Markov image with 2-category. The transition matrix:

$$P_{ij} = \begin{matrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{matrix}$$

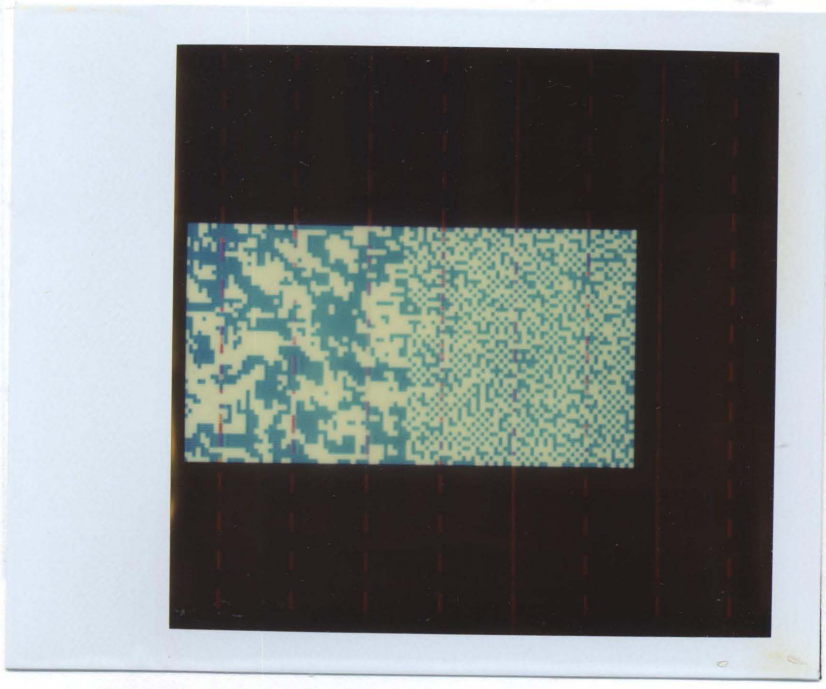

Figure 6.2 Simulated experiment of transition probability estimation using Robust estimation technique. This image has two parts in which there are two extreme different transition probability matricies.

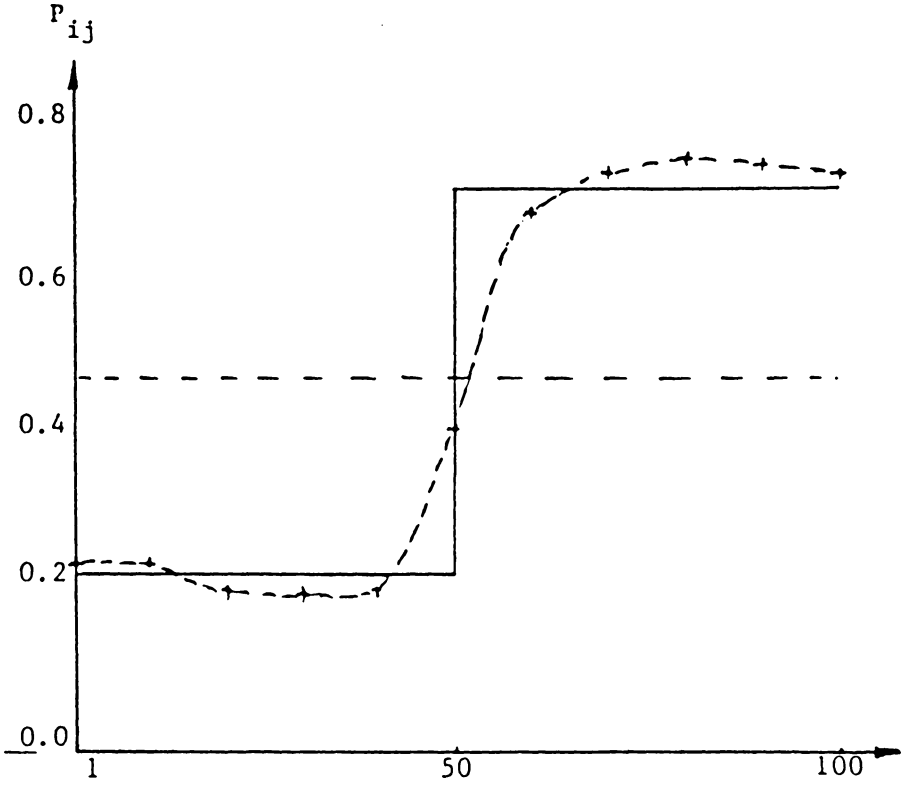


Figure 6.3 Estimation curves of transition probability vs x-coordinate of the test image. curve 1 - true of Pij curve 2 - maximum likelihood estimation curve 3 - space varying estimation

Table 6.1 Confidence intervals at the 0.05 level and projected sample sizes necessary for P ij image size 19 X 19 ( ie. n = 361 )

| | Lower limit | Upper limit | range | projected | Pij | Actural |
|---|---|---|---|---|---|---|
| P11 | 0.294 | 0.457 | (.090,.090) | 1183 | 0.439 | 0.40 |
| P12 | 0.069 | 0.212 | (.072,.072) | 741 | 0.141 | 0.20 |
| P13 | 0.146 | 0.315 | (.080,.080) | 1036 | 0.230 | 0.20 |
| P14 | 0.156 | 0.324 | (.080,.080) | 1087 | 0.244 | 0.20 |
| P21 | 0.031 | 0.166 | (.070,.070) | 654 | 0.099 | 0.20 |
| P22 | 0.330 | 0.508 | (.090,.090) | 1123 | 0.420 | 0.40 |
| P23 | 0.141 | 0.303 | (.080,.080) | 944 | 0.222 | 0.20 |
| P24 | 0.174 | 0.344 | (.080,.080) | 1039 | 0.259 | 0.20 |
| P31 | 0.137 | 0.272 | (.070,.070) | 652 | 0.204 | 0.20 |
| P32 | 0.115 | 0.251 | (.070,.070) | 662 | 0.183 | 0.20 |
| P33 | 0.244 | 0.401 | (.080,.080) | 891 | 0.323 | 0.40 |
| P34 | 0.212 | 0.369 | (.080,.080) | 885 | 0.290 | 0.20 |
| P41 | 0.115 | 0.234 | (.060,.060) | 509 | 0.174 | 0.20 |
| P42 | 0.104 | 0.226 | (.060,.060) | 543 | 0.165 | 0.20 |
| P43 | 0.198 | 0.334 | (.070,.070) | 660 | 0.266 | 0.20 |
| P44 | 0.321 | 0.468 | (.070,.070) | 781 | 0.394 | 0.40 |

Table 6.2 Confidence intervals at the 0.05 level and projected sample sizes necessary for Pij image size 34 X 34 ( ie. n = 1156 )

| | Lower limit | Upper limit | range | projected | Pij | Actural |
|---|---|---|---|---|---|---|
| P11 | 0.355 | 0.451 | (.050,.048) | 1009 | 0.402 | 0.40 |
| P12 | 0.155 | 0.236 | (.040,.040) | 759 | 0.196 | 0.20 |
| P13 | 0.148 | 0.229 | (.040,.043) | 767 | 0.189 | 0.20 |
| P14 | 0.171 | 0.250 | (.050,.050) | 814 | 0.213 | 0.20 |
| P21 | 0.125 | 0.213 | (.040,.040) | 870 | 0.170 | 0.20 |
| P22 | 0.313 | 0.409 | (.048,.048) | 1074 | 0.362 | 0.40 |
| P23 | 0.178 | 0.267 | (.045,.045) | 917 | 0.223 | 0.20 |
| P24 | 0.200 | 0.290 | (.045,.045) | 955 | 0.246 | 0.20 |
| P31 | 0.165 | 0.237 | (.040,.040) | 622 | 0.201 | 0.20 |
| P32 | 0.154 | 0.229 | (.037,.037) | 650 | 0.190 | 0.20 |
| P33 | 0.374 | 0.460 | (.044,.044) | 898 | 0.418 | 0.40 |
| P34 | 0.150 | 0.220 | (.040,.040) | 668 | 0.189 | 0.20 |
| P41 | 0.170 | 0.214 | (.040,.040) | 734 | 0.205 | 0.20 |
| P42 | 0.138 | 0.218 | (.040,.040) | 751 | 0.179 | 0.20 |
| P43 | 0.195 | 0.276 | (.040,.040) | 742 | 0.235 | 0.20 |
| P44 | 0.335 | 0.425 | (.050,.050) | 948 | 0.381 | 0.40 |

Table 6.3 Contingency tables for classification results of simulated test image of Figure 6.2

COL = assigned categories ROW = true categories

(A) Bayes classification result

| CLASS | 1 | 2 | 3 | TOTAL | #ERR | ACC(%)* |
|---|---|---|---|---|---|---|
| 1 | 998 | 45 | 45 | 1088 | 90 | 91.73% |
| 2 | 27 | 653 | 55 | 735 | 82 | 88.85% |
| 3 | 27 | 2 | 648 | 677 | 29 | 95.4% |
| TOTAL | 1052 | 700 | 748 | 2500 | 201 | 91.96% |

(B) Context classification result using maximum likelihood estimation technique

| CLASS | 1 | 2 | 3 | TOTAL | #ERR | ACC(%)* |
|---|---|---|---|---|---|---|
| 1 | 1046 | 16 | 26 | 1088 | 42 | 96.1% |
| 2 | 47 | 647 | 41 | 735 | 88 | 88.1% |
| 3 | 40 | 4 | 633 | 677 | 44 | 93.5% |
| TOTAL | 1133 | 667 | 700 | 2500 | 174 | 93.0% |

(C) Context classification result using space-varying transition probability estimation technique.

| CLASS | 1 | 2 | 3 | TOTAL | #ERR | ACC(%)* |
|---|---|---|---|---|---|---|
| 1 | 1053 | 26 | 9 | 1088 | 35 | 96.8% |
| 2 | 36 | 687 | 12 | 735 | 48 | 93.5% |
| 3 | 41 | 19 | 617 | 677 | 60 | 91.4 |
| TOTAL | 1130 | 732 | 638 | 2500 | 143 | 94.3% |

Table 6.4 Contingency tables for classification results of 13 April 1976 MSS scene of Roanoke, VA. Scale factor of the number of pixels 10 ** 1

COL = assigned categories ROW = true categories

(A) Bayes classification result

| CLASS | URB | AGR | RNG | FST | TOTAL | ACC(%)* |
|---|---|---|---|---|---|---|
| URB | 398 | 210 | 0 | 635 | 635 | 63.0% |
| AGR | 86 | 181 | 0 | 36 | 303 | 60.0% |
| RNG | 0 | 0 | 0 | 0 | 0 | - |
| FST | 20 | 7 | 0 | 34 | 61 | 56.0% |
| TOTAL | 504 | 398 | 0 | 97 | 999 | 61.4% |

(B) Context classification result using maximum likelihood estimation technique

| CLASS | URB | AGR | RNG | FST | TOTAL | ACC(%)* |
|---|---|---|---|---|---|---|
| URB | 454 | 173 | 0 | 8 | 635 | 71.0% |
| AGR | 94 | 200 | 0 | 9 | 303 | 66.0% |
| RNG | 0 | 0 | 0 | 0 | 0 | - |
| FST | 26 | 15 | 0 | 19 | 60 | 32.0% |
| TOTAL | 574 | 388 | 0 | 36 | 998 | 67.5% |

(C) Context classification result using space-varying transition probability estimation technique.

| CLASS | URB | AGR | RNG | FST | TOTAL | ACC(%)* |
|---|---|---|---|---|---|---|
| URB | 533 | 99 | 0 | 4 | 636 | 83.3% |
| AGR | 122 | 168 | 0 | 13 | 304 | 55.0% |
| RNG | 0 | 0 | 0 | 0 | 0 | - |
| FST | 18 | 1 | 0 | 39 | 61 | 67.0% |
| TOTAL | 737 | 170 | 0 | 53 | 1000 | 74.0% |

Table 6.5 Contingency tables for classification results of MSS scene of Clark, OR.

COL = ASSIGNED CAT ROW = TRUE CAT

(A) Pixel independet Bayes' classification result:

| CLASS | WHT | ALF | POT | CRN | PAS | RNG | TOTL | ACC(%) |
|---|---|---|---|---|---|---|---|---|
| WHT | 1560 | 31 | 110 | 8 | 10 | 132 | 1851 | 84.3% |
| ALF | 44 | 295 | 241 | 25 | 167 | 125 | 897 | 32.3% |
| POT | 88 | 46 | 1219 | 25 | 45 | 138 | 1561 | 78.1% |
| CRN | 4 | 8 | 5 | 351 | 3 | 22 | 393 | 89.3% |
| PAS | 0 | 0 | 2 | 0 | 15 | 1 | 18 | 83.3% |
| RNG | 15 | 14 | 13 | 27 | 12 | 99 | 180 | 55.0% |
| TOTL | 1711 | 394 | 1590 | 436 | 252 | 517 | 4900 | 72.2% |

(B) Context classification result by stochastic relaxation using maximum likelihood estimation technique:

| CLASS | WHT | ALF | POT | CRN | PAS | RNG | TOTL | ACC(%) |
|---|---|---|---|---|---|---|---|---|
| WHT | 1643 | 9 | 112 | 5 | 2 | 80 | 1851 | 88.7% |
| ALF | 46 | 299 | 296 | 21 | 116 | 119 | 897 | 33.3% |
| POT | 72 | 22 | 1361 | 21 | 24 | 61 | 1561 | 87.1% |
| CRN | 7 | 8 | 6 | 351 | 2 | 19 | 393 | 89.3% |
| PAS | 0 | 0 | 3 | 0 | 14 | 1 | 18 | 77.7% |
| RNG | 26 | 6 | 42 | 21 | 7 | 78 | 180 | 43.3% |
| TOTL | 1794 | 344 | 1820 | 419 | 165 | 358 | 4900 | 76.4% |

Table 6.5 Contingency tables for classification results of MSS scene of Clark, OR.

(C) Context classification result by stochastic relaxation using space-varying transition probability estimation technique:

| CLASS | WHT | ALF | POT | CRN | PAS | RNG | TOTL | ACC(%) |
|---|---|---|---|---|---|---|---|---|
| WHT | 1845 | 1 | 2 | 0 | 0 | 3 | 1851 | 99.6% |
| ALF | 251 | 584 | 11 | 8 | 0 | 43 | 897 | 65.1% |
| POT | 507 | 13 | 1015 | 9 | 0 | 17 | 1561 | 65.1% |
| CRN | 64 | 14 | 5 | 288 | 0 | 22 | 393 | 73.3% |
| PAS | 0 | 0 | 0 | 0 | 18 | 0 | 18 | 100.% |
| RNG | 105 | 14 | 2 | 4 | 0 | 55 | 180 | 58.3% |
| TOTL | 2772 | 626 | 1035 | 309 | 18 | 140 | 4900 | 77.6% |

* Classification accuracy.
** Overall classification accuracy : ratio of the number correctly classified pixels to the number of total classified pixels.

WHT – Wheat
ALF – Alfalfa
POT – Potatoes
CRN – Corn
RNS – Beans
APL – Apples
PAS – Pasture ( irrigated )
RNG – Rangeland

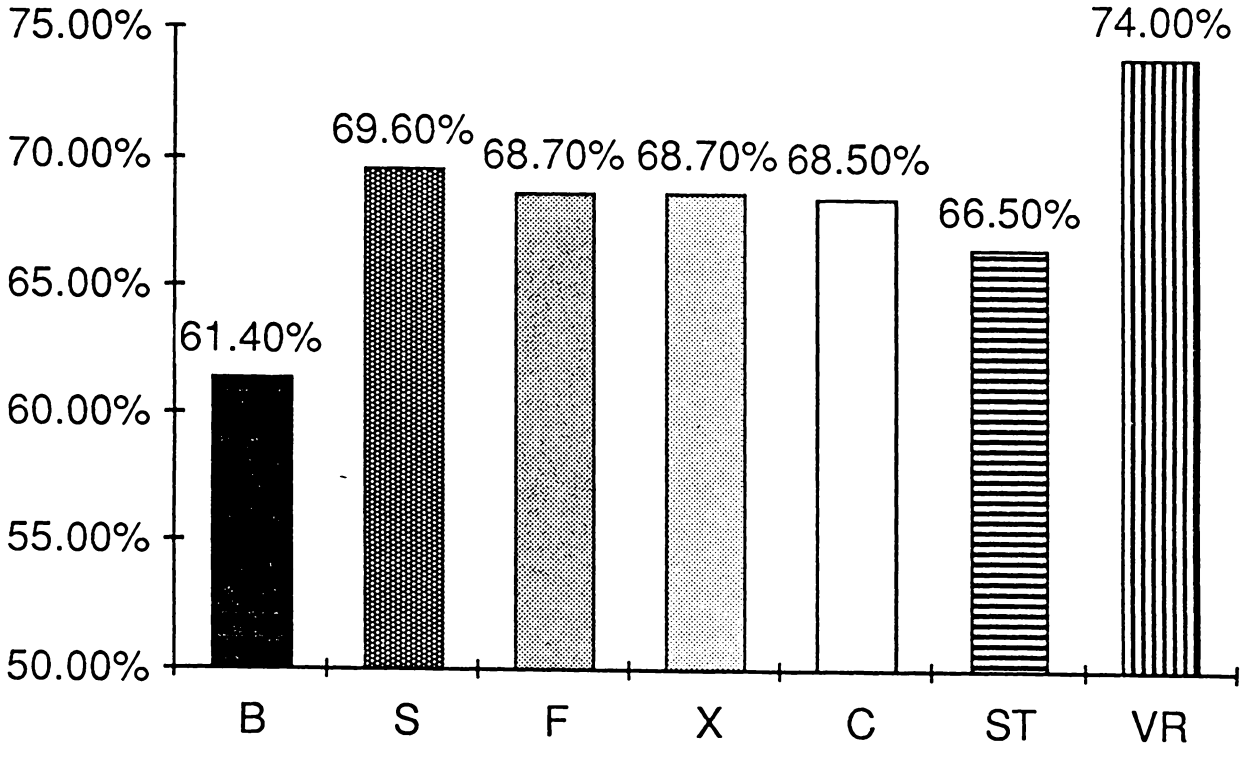


Figure 6.4 Comparison of overall classification accuracies of MSS scene of Roanoke, VA., using different classifiers. B : pixel indepedent Bayes classifies; S: two pass forward-backward look-ahead; F : four pass; X : one step context look-ahead; C : no context look-ahead; ST : stochastic relaxation. VR : contextual stochastic relaxation classification using space-varying estimation technique.

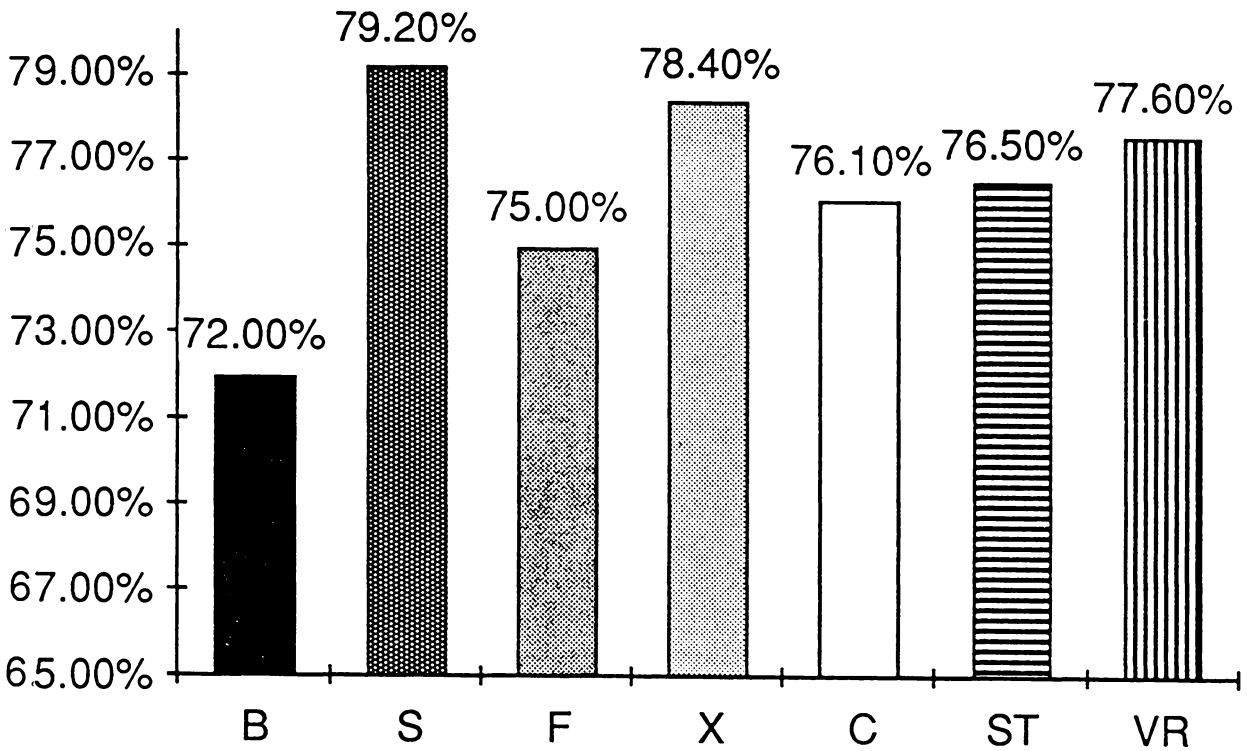


Figure 6.5 Comparison of overall classification accuracies of MSS scene of Clarke, OR., using different classifiers. B : pixel indepedent Bayes classifies; S: two pass forward-backward look-ahead; F : four pass; X : one step context look-ahead; C : no context look-ahead; ST : stochastic relaxation. VR : contextual stochastic relaxation classification using space-varying estimation techniq.

# CHAPTER VII: CONCLUSION

In this dissertation, most of the efforts have been aimed to developing a feasible stochastic model for contextual pattern classification, and to implementing the developed contextual algorithm in the classification of remotely sensed MSS data. Other efforts have been aimed at estimating the transition probability matrix and its confidence intervals.

To simulate the human performance in information-processing and decision making, context information must be fully employed in the decision making process. The importance of the research in this dissertation is justified by its theoretical and practical progress and contributions in the above area.

First a Markov Random Field assumption is adopted as the mathematic model for our contextual classification algorithms. From this model we have described one way of developing a context classification scheme from a Bayesion framework. We discussed three contextual decision rules for minimum error probability under the 2-D discrete Markov Random Field, and the rules are characterized by the range of their look-ahead capabilities. We demonstrated comparative experiments using those different contextual decision rules. Our results show that the one-step look-ahead has a fairly low computational cost, and yields a significant improvement over the context free rule. Also, the complete context algorithm always provides an

answer as good as or better than the one-step look-ahead. In addition, we derived a contextual classification process using the context of the whole image for decision making. That is, we use the entire context of the image. The context classifier produces a global result and not a local result.

We also derived an alternative context classification method based on stochastic relaxation algorithm and the Markov- Gibbs Random Field. The implementation of the relaxation algorithm is one form of optimization using annealing. In addition, an improved method illustrated that the computation cost can be greatly reduced.

The estimation of transition probabilities played an important rule in our decision schemes. We formulated a maximum likelihood and robust estimation technique for the transition probability matrix. A confidence interval estimation technique for making sample size determinations for a desired parameter range at a specified confidence level was also demonstrated.

**Suggestions for further research**

The following research problems, which are significant and challenging, are suggested.

1) Most of the existing contextual classification algorithms are compu-

tational intensive. This is still a substantial obstacle to wide use in the remote sensing applications and other pattern recognition areas. There are two suggested ways to solve this problem. First, new and more efficient algorithms are still highly desirable, and parallelism is the key to success in reducing the cost of the contextual classification. Another way is to exploit vectorizing, pipelining and data flow concepts to achieve a new parallelism in computer systems.

2) Not only do we need new parallel processing systems and high parallel algorithms, but we also need the development of effective procedures for mapping parallel algorithms onto existing parallel architectures, and alternatively, for specifying optimal architectures for a given algorithm.

3) Most of the recent research in contextual classification has been the classification of small neighborhoods using local contextual information. It is still a great challenging problem how to effectively use global contextual information in the pattern recognition. Further research on the information extraction processes should extend contextual reasoning based on any or all information available, local or global, about a given scene location. Such processes would incorporate a relatively high- level intelligence in the decision-making operation.

Reference

Abend, K. , T. J. Harley and L. N. Kanal, 1965. "Classification of Binary Random Patterns". IEEE Trans. Info. Theory, Vol. 11, 1965, pp. 538-543.

Ahuja, N., A. Rosenfeld and R. M. Haralick, 1980. "Neighbor Gray Levels as Features in Pattern Classification". Pattern Recognition, Vol. 12, 1980, pp. 251-260.

Averintser, M. B. , 1970. "On a Method of Describing Discrete Parameter Random Fields". Problem Peredachi Informatsii, Vol. 6, 1970, pp. 100-109.

Bajcsy, R. and M. Tavakoli, 1976. "Computer Recognition of Roads from Satellite Pictures". IEEE Trans. on Syst., Man and Cybernetics, Vol. SMC-6, No. 9, 1976, pp. 623-637.

Bartlett, M. S., 1955. "An Introduction to Stochastic Processes". Cambridge, Univ. Press, 1955.

Besay, J. E. , 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems". Journal of the Royal Statistical Society, Vol. B36, 1974, pp. 192-236.

Bryant, J., 1979. "On the Clustering of Multidimensional Pictorial Data". Pattern Recognition, Vol. 11, 1979, pp. 115-125.

Brayer, J. M., P. H. Swain, and K. S. Fu, 1980. "Modelling of Earth

Resources Satellite Data". In Fu, K. S. (ed.) Applications of Syntactic Pattern Recognition, Springer-Verlag, New York, 1980, pp. 50-159.

Burdick, R. C., and R. A. Speirer, 1980. "Development of a Method to Detect Geologic Faults and Other Linear Features from Satellite Images". U.S. Bureau of Mines Report of Investigations, No. 8413, 1980.

Chen, S., 1985. "A Data Flow Computer Architecture For Markov Image Models". IEEE Computer Society Workshop on Computer Architecture for Pattern Analysis and Image Database Management, 1985, pp. 75-79.

Chow, C. K. , 1962. "A Recognition Method Using Neighbor Dependence". IEEE Trans. on Electronic Computers, Vol. EC-11, Oct., 1962, pp. 683-690.

Cooper, D. B. and F. P. Sung, 1983. "Multiple-window Parallel Adaptive Boundary Finding in Computer Vision". IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 5, No. 3, 1983, pp. 299-316.

Deriv, H., H. Elliott, R. Christ and D. Geman, 1983. "Bayes Smoothing Algorithms for Segmentation of Images Modelled by Markov Random Fields". Dept. of Computer Science, University of Massachusetts, Technical Report, August, 1983.

Devijver, P. A. , 1984. "A Comparative Study of Decision Making Algorithms in Hidden Markov Chains". Report R481, Philips Research Lab., Brussels, Belqium, Oct. 1984.

Donaldson, R. W. , and G. T. Toussaint, 1970. "Use of Contextual Constraints in Recognition of Contour Traced Handprinted Characters". IEEE Trans. Computer, Vol. C-19, No. 11, 1970, pp. 1095-1099.

Duda, R. O. and P. E. Hart, 1973. "Pattern Classification and Scene Analysis". John Wiley and Sons, 1973, pp. 405-424.

Edwards, and R. L. Chambers, 1964. "Can a Priori Probabilities Help in Character Recognition?". Association for Computing Machinery, Vol. 11, No. 4, 1964, pp. 465-470.

Elliott, H. and H. Dervin, R. Christ and Geman, 1983. "Application of the Gibbs Distribution to Image Segmentation". Dept. of Computer Science, University of Massachusetts, Technical Report, August, 1983.

Forney, G. D. , 1972. "Maximum Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbolic Interfence". IEEE Trans. Infor. Theory, Vol. IT-18, May, 1972, pp. 283-295.

Forney, G. D. , 1973. "The Viterbi Algorithms". Proc. of The IEEE, Vol. 61, No. 3, 1973, pp. 268-278.

Fu, K. S. , 1976. "Pattern Recognition in Remote Sensing of Earth Resources". IEEE Tran. on Geo. Scien. and Elc., Vol. GE-14, No. 1, Jan., 1976, pp. 10-18.

Fu, K. S. and A. Rosenfield, 1976. "Pattern Recognition and Image Processing". IEEE Trans. Computer, Vol. C-25, Dec., 1976, pp. 1336-1346.

Fu, K. S. and T. S. Yu, 1980. "Statistical Pattern Classification Using Contextual Information". Research Studies Press, John Wiley & Sons, Ltd., 1980.

Fu, K. S., 1976. "Syntactic(linguistic) Pattern Recognition". In Fu, K. S. (ed.) Digital Pattern Recognition, Springer-Verlag, New York, 1976.

Geman, S. and D. Geman, 1984. "Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images". IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 6, No. 6, 1984, pp. 721-724.

Grimmertt, G. R. , 1973. "A Theorem about Random Fields". Bull. London Math. Soc., Vol. 5, 1973, pp. 81-84.

Goetz, A. F., F. C. Billingsley, A. R. Gillespie, M. J. Abrams, 1975. "Applications of ERTS-1 Images and Image Processing to Regional Geologic Problems and Mapping in Northern Arizona". Jet Propulsion Lab Tech. Rept. 32-1597, Pasadena, California, pp. 188.

Groffeath, 1973. "Markov Random Fields". Unpublished manuscript.

Groffeath, 1976. "Introduction to Random Fields". In Knapp and Snell (ed.) Denumerable Markov Chains, Springer-Verlag, New York, 1976.

Gurney, C. M. , and Townshend, 1983. "The Use of Contextual Information in the Classification of Remote Sensing Data". Photogrammetric Engineering and Remote Sensing, Vol. 49, No. 1, Jan., 1983, pp. 10-15.

Gurney, C.M., 1980. "Threshold Selection for Line Detection Algorithms". IEEE Trans. on Geosci. and Remote Sensing, Vol. GE-18, No. 2,

1980, pp. 204-211.

Haralick, R. M. , K. Shanmugam and I. Dinstein, 1973. "Textural Features for Image Classification". IEEE Trans. on Syst. Man and Cybernetics, Vol. SMC-3, 1973, pp. 610-621.

Haralick, R. M. , 1983. "Decision Making in Context". IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-5, No. 4, July, 1983, pp. 417-428.

Haralick, R. M. , M. C. Zhang and J. B. Campbell, 1984. "Multispectral Image Context Classification Using the Markov Random Field". PECORA Proceedings, Oct., 1984, pp. 190-200.

Hassner, M. and J. Sklansky, 1980. "The Use of Markov Random Field as Models of Texture". Computer Graphics and Image Processing , Vol. 12, 1980, pp. 357-370.

Hansen, F. R. and H. Elliott, 1982. "Image Segmentation Using Simple Markov Field Models". Computer Graphics and Image Processing, Vol. 20, 1980, pp. 101-132.

Hinton, G. E., T. J. Sejnowski and D. H. Ackley, 1984. "Boltzmann Machines : Constraint Satisfaction Networks that Learn". Technical Report, CMU-CS-84-119.

Ho, Y. C. and A. K. Agrawala, 1968. "On Pattern Classification Algorithm Introduction and Survey". Proc. of The IEEE, Vol. 56, Dec. 1968, pp. 2101-2114.

Ising, E., 1925. "Beitray Sur Theorie Des Ferromagnetismus". Zeitschrift Physik, Vol. 31, 1925, pp. 253-258.

Itten, K. I. and F. Fasler, 1978. "Thematic Adaptive Spatial Filtering of Landsat Land Use Classification Results". Proc. Conf. on Remote Sensing of Environment, 1978, pp. 1035-1042.

Jain, A. K. and E. Angel, 1974. "Image Restoration, Modeling, and Reduction of Dimensionally". IEEE Trans. Computer, Vol. C-23, 1974, pp. 11-16.

Kanal, L. N. , 1972. "Interactive Pattern Analysis and Classification Systems: A survey and Commentary". Proc. of The IEEE, Vol. 60, Oct., 1972. pp. 1200-1215.

Kemeny, J. G., J. L. Snell and Knapp, 1976. "Finite Markov Chains". Springer-Verlag, New York, 1976.

Kettig, R. L. and D. A. Landgrebe, 1976. "Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects". IEEE Trans. on Geoscience Electronics, Vol. GE-14, No. 1, 1976, pp. 19-21.

Kirkpatrick, S. and C. D. Gellatt, Jr., and M. P. Vecchi, 1982. "Optimization by Simulated Annealing". Science, No. 222, 1982, pp. 671-680.

Kittler, J. and J. Fogein, 1985. "Contextual Classification of Multispectral Pixel Data". Computer Graphics and Image Processing, Vol. 3, 1985, pp. 13-29.

Kinderman, R. P., 1973. "Random Fields : Theorems and Examples". Journal of Undergraduate Math., Vol. 5, 1973, pp. 25-34.

Landgrebe, D. A., 1980. "The Development of a Spectral-Spatial Classification for Earth Observational Data". Pattern Recognition, Vol. 12, 1980, pp. 165-176.

Lee, J. S., 1980. "Digital Image Enhancement and Noise Filtering by Use of Local Statistics". IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 2, Mar., 1980, pp. 165-169.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953. "Equations of State Calculations by Fast Computing Machines". Journal of Chem. Phys., Vol. 21, 1953, pp. 1087-1091.

Moran, P. A., 1973. "Necessary Conditions for Markovian Processes on a Lattice". Journal of Applied Probability, Vol. 10, 1973, pp. 605-612.

Monotot, L., 1977. "Digital Detection of Linear Features in Satellite Imagery". Proc. Int. Symp. on Image Process., Graz, Austria, 1977, pp. 149-153.

Nagy, G. , 1972. "Digital Image Processing Activities in Remote Sensing for Earth Resource". Proc. of The IEEE, Vol. 60, No. 10, 1972, pp. 1170-1200.

Nagy, G., T. Matsuyama and H. Mori, 1978. "A Structuaral Analysis of Complex Aerial Photographics". Proc. of 6th IJCAI, Tokyo, 1979, pp. 75-87.

Nahi, N. E. and A. Habihi, 1975. "Decision-Directed Recursive Image Enhancement". IEEE Trans. Circuits and System on Digital Fieltering and Image Processing, Vol. CAS-22, Mar., 1975, pp. 286-293.

Nack, M. L., 1977. "Dectification and Registration of Digital Images and the Effect of Cloud Detection". Proc. Symp. on Mach. Process. of Remotely Sensed Data, Purdue Univ., Indiana, 1977, pp. 12-13.

Neuhoff, D. L. , 1975. "The Viterbi Algorithm as an Aid in Text Recognition". IEEE Trans. Inf. Theory, Vol. 21, 1975, pp. 222-226.

Preston, C. J. , 1974. "Gibbs States on Countable Sets". Cambrige Univ. Press, 1974.

Raviv, J. , 1967. "Decision Making in Markov Chain Applied to the Problem of Pattern Recognition". IEEE Trans. Inform. Theory, Vol. It-13, Oct., 1967, pp. 535-551.

Rosenfeld, A. and A. C. Kak, 1976. "Digital Picture Processing". Academic Press, New York.

Rota, G. C., 1964. "On the Foundations of Combinational Theory, I. Theory of Mobius Functions". Z'tschr. Wahrsch' Theorie & Verw. Geb., Vol. 2, 1964, pp. 340-368.

Rubinstein, B. Y. , 1981. "Simulation and the Monte Carlo Method". Series in Probability and Mathematical Statistics, John Wiley and Son, New York, 1981.

Spitzer, F. , 1971. "Markov Random Field and Gibbs Ensembles". American Mathematical Monthy, Vol. 28, 1971, pp. 142-154.

Sherman, 1973. "Markov Random Field and Gibbs Random Field". Israel Journal of Math. ,Vol. 14, 1973, pp. 92-103.

Sminov, N. V. , O. V. Sarmanov, and V. K. Zahrov, 1966. "A Local Limit Theorem for Transition Numbers in a Markov Chain, and its Applications". Soviet Math., Doklady, Vol. 7, No. 3, 1966, pp. 151-157.

Saralar, S. A. and R. J. DeFiguereido, 1981. "Adaptive Image Restoration by a Modified Kalman Filtering Approach". IEEE, Acoust., Speech and Signal Processing, Vol. Ass-29, Oct., 1981, pp. 1033-1042.

Strickland, R. N. , 1985. "Estimation of Local Statistics for Digital Processing of Nonstationary Image". IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. Assp-33, No. 2, April, 1985, pp. 465-468.

Strong, J. P. and A. Rosenfeld, 1973. "A Region Technique for Scene Analysis". Communication of the ACM, Vol. 16, No. 4, 1973, pp. 237-246.

Swain, P. H. , S. B. Vardeman and J. C. Tiltor, 1981. "Contextual Classification of Multispectral Image Data". Pattern Recognition, Vol. 13, No. 6, 1981, pp. 428-441.

Swain, P. H. , H. J. Siegel and B. W. Smith, 1979. "A Method for Classifying Multispectral Remote Sensing Data Using Context". Machine Processing of Remote Sensed Data Symposium, 1979, pp. 343-353.

Tekalp, A. M. , H. Kaufman, and J. W. Woods, 1985. "Fast Recursive Estimation of the Parameters of a Space-Varying Autoregressive Image Model". IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. Assp-33, No. 2, April, 1985, pp. 496-471.

Tilton, T. C. , S. B. Vardeiman and P. H. Swain, 1982. "Estimation of Context for Statistical Classification of Multispectral Image Data". IEEE Trans. on Geo. and Remote Sensing, Vol. GE-20, No. 4, Oct., 1982, pp. 445-451.

Toussaint, G. T. , 1978. "The Use of Context in Pattern in Recognition". Pattern Recognition, Vol. 10, 1978, pp. 189-204.

VanderBrug, G. J., 1976. "Line Detection in Satellite Imagery". IEEE Trans. on Geosci and Electronics, Vol. GE-14, No. 1, 1976, pp. 37-44.

Wallis, R. , 1976. "An Approach to Space Variant Restoration and Enhancement of Images". Proc. Symp. Current Mathematical Problems in Image Science. Naval Postgraduate School, Monterey, CA. Nov., 1976, pp. 78-84.

Waszka, J., C. Dyer and A. Rosenfeld, 1976. "A Comparative Study of Texture Measure for Terrain Classification". IEEE Trans. Syst. Man Cybernetics, Vol. SMC-6, No. 4, 1976, pp. 269-285.

Welch, J. R. and K. G. Salter, 1971. "A Context Algorithm for Pattern Recognition and Image Interpretation". IEEE Trans. Syst., Man, and Cybernetics, Vol. SMC-1, No. 1, 1971, pp. 24-30.

Welch, J. R. and K. G. Salter, 1971. "A Context Algorithm for Pattern Recognition and Image Interpretation". IEEE Trans. on Syst., Man and Cybernetics, Vol. SMC-1, No. 1, 1971, pp. 24-30.

Wharton, S., 1982. "A Contextual Classification Method for Recognizing Land Use Patterns in High Resolution Remotely Sensed Data". Pattern Recognition, Vol. 15, No. 15, 1982, pp. 317-324.

Woods, J. W. , 1972. "Two-Dimensional Discrete Markovian Field". IEEE Trans. On Information Theory, Vol. IT-18, 1972, pp. 232-240.

Wong, E. , 1973. "Recent Progress in Stochastic Processes - a Survey". IEEE Trans. On Information Theory, Vol. IT-19, No. 3, May, 1973, pp. 262-274.

Wharton, S. W. , 1982. "A Contextual Classification Method For Recognizing Land Use Patterns in High Resolution Remotely Sensed Data". Pattern Recognition, Vol. 15, No. 4, 1982, pp. 317-324.

Whittle P., 1955. "Some Distribution and Moment Formulae for the Markov Chain". Journal of Royal Stat. Soc, Vol. 17, 1955, Series B, pp. 235-242.

Whittle, P., 1963. "Stochastic Processes in Several Dimensions". Bull. Int. Stat. Inst., Vol. 40, 1963, pp. 974-994.

Young, R. E. , 1977. "Transition Probability Estimation for Discrete State Markov Chains". Thesis, Purdue University, 1977.

**Appendix A : Whittle's distribution and its bivariate normal approximation**

For the two-class case Whittle's distribution is

$$P( m_i , m_{ij} , i,j = 1,2 ) =$$

$$\frac{F\, m_1!\, m_2!}{m_{11}m_{12}m_{21}m_{22}} p_{11}^{m_{11}} p_{12}^{m_{12}} p_{21}^{m_{21}} p_{22}^{m_{22}} \qquad (A.1)$$

where

$$F = \begin{vmatrix} \frac{m_{21}}{m_1} & \frac{m_{21}}{m_1} \\ \frac{m_{12}}{m_2} & \frac{m_{12}}{m_2} \end{vmatrix}$$

Let $p_1$ and $p_2$ be given by

$$p_1 = \frac{p_{21}}{p_{12} + p_{21}}$$

$$p_2 = \frac{p_{12}}{p_{12} + p_{21}} \qquad (A.2)$$

$p_1$ and $p_2$ are steady-state probabilities.

The bivariate normal approximation to Whittle's distribution for the two-state class can be expressed as follows:

$$f(m_1 , m_{11} ) = \frac{1}{2\pi} \frac{( p_{12} + p_{21})^2}{n p_{12} p_{21} (\sqrt{p_{11} p_{22}})} * e^{\frac{-q}{2}} \qquad (A.3)$$

where

$$q = \frac{1}{(1-\rho^2)}\left[\frac{(m_{11}-\bar{m}_{11})^2}{\sigma_{11}^2} + \frac{(m_1-\bar{m}_1)^2}{\sigma_1^2} - \frac{2\rho}{\sigma_1\sigma_{11}}(m_{11}-\bar{m}_{11})(m_1-\bar{m}_1)\right]$$

and

$$\sigma_{11}^2 = \frac{np_{11}p_{12}p_{21}\left[\, p_{11}p_{22} + (\, p_{12}(\, 1 + p_{21}\,)^2\right]}{(p_{12}+p_{21})^3}$$

$$\sigma_1^2 = \frac{np_{12}p_{21}\,(\, p_{11} + p_{12})}{(p_{12}+p_{21})^3}$$

$$\sigma_{1,11}^2 = \frac{2 * n\, p_{11}p_{12}p_{11}}{(p_{12}+p_{21})^3}$$

$$\rho = \frac{2\sqrt{p_{11}}}{\sqrt{p_{11}+\, p_{22}(\, p_{11}p_{22}+\,(\, 1+\, p_{21})^2)}} \qquad \text{(A.4)}$$

$\rho$ represents the correlation coefficient between $m_1$ and $m_{11}$, $\sigma_1^2$ represents the variance of $m_1$, $\sigma_{1,11}$ is covariance between $m_1$ and $m_{11}$, and $\sigma_{11}^2$ is variance of $m_{11}$.

Where $\bar{m}_1 = np_1$,

and $\bar{m}_{11} = n\, p_1p_{11} = \dfrac{n\, p_1\, p_{11}}{(p_{12}+p_{21})}$

For $f(m_1\,,\, m_{12})$ we have

$$\bar{m}_{12} = \frac{np_{12}p_{21}}{(p_{12}+p_{21})}$$

and

$$\sigma_{12}^2 = \frac{np_{12}p_{21}(\, p_{11}\, p_{22}^2 + p_{22}p_{12}^2)}{(p_{12}+p_{21})^3}$$

$$\sigma_{1,12} = \frac{n\, p_{12}\, p_{21}\, (\, p_{22} - p_{11}\,)}{(p_{12}+p_{21})^3}$$

$$\rho = \frac{\sigma_{1,12}}{\sqrt{\sigma_1 \sigma_{12}}} \qquad \text{(A.5)}$$

For f ( $m_2$ , $m_{21}$ ) we have

$$\overline{m}_{21} = \frac{np_{21}p_{12}}{(p_{12}+p_{21})}$$

$$\overline{m}_2 = \frac{np_{12}}{(p_{12}+p_{21})}$$

$$\sigma_2^2 = \sigma_1^2$$

$$\sigma_{21}^2 = \sigma_{12}^2$$

$$\sigma_{2,21} = \sigma_{1,12}$$

$$\rho_{2,21} = \rho_{1,12} \qquad \text{(A.6)}$$

And for f ( $m_2$ , $m_{22}$ ) we have

$$\overline{m}_{22} = \frac{np_{22}p_{12}}{(p_{12}+p_{21})}$$

$$\sigma_{22}^2 = \frac{n\, p_{22}p_{12}p_{21}[p_{11}p_{12}+(1+p_{12})^2]}{(p_{12}+p_{21})^3}$$

$$\sigma_{2,22} = \frac{2np_{22}p_{12}p_{21}}{(p_{12}+p_{21})^3}$$

$$\rho = \frac{2\sqrt{p_{22}}}{\sqrt{(p_{11}+p_{22})[p_{11}p_{22}+(1+p_{12})^2]}} \qquad \text{(A.5)}$$

**Appendix B : Distribution function for the transition probability $p_{ij}$ and its approximation :**

Using the result of distribution f( $m_i$ , $m_{ij}$ ), Young (1977) derived the density function for $f(\hat{p}_{ij})$. The ratio of two normally distributed random variables is given by

$$\hat{p}_{ij} = \frac{m_{ij}}{m_i} \; ; \; (n > m_i > m_{ij} > 0 \; ; 0 < \frac{m_{ij}}{m_i} < 1)$$

and

$$f(\hat{p}_{ij}) = K\, e^{-\frac{e}{2(1-\rho^2)}} \frac{(1-\rho^2)}{a} * [e^{\frac{-g}{2}} - e^{\frac{-h}{2}}]$$

$$+ \frac{b}{a^{\frac{3}{2}}} * \sqrt{(1-\rho^2)} * [\int_{-\sqrt{g}}^{\infty} e^{-\frac{x^2}{2}} dx - \int_{\sqrt{h}}^{\infty} e^{-\frac{x^2}{2}} dx] \qquad \text{(B.1)}$$

where

$$g = \frac{(v - e)}{(1 - \rho^2)}$$

$$h = \frac{(L^2(n) - e)}{(1 - \rho^2)}$$

$$a = \frac{(\sigma_{ij}^2 - 2\rho\,\hat{p}_{ij}\,\sigma_i\,\sigma_{ij} + \hat{p}_{ij}^2\,\sigma_i^2)}{\sigma_{ij}^2\,\sigma_i^2}$$

$$b = -\sigma_{ij}(\rho\,\overline{m}_{ij}\,\sigma_i - \overline{m}_i\sigma_{ij}) + \frac{\hat{p}_{ij}\sigma_i(\rho\overline{m}_i\sigma_{ij} - \overline{m}_{ij}\sigma_i)}{\sigma_i^2\,\sigma_{ij}^2}$$

$$v = \frac{\overline{m}_i^2}{\sigma_i^2} - \frac{2\rho\,\overline{m}_i\,\overline{m}_{ij}}{\sigma_i\sigma_{ij}} + \frac{\overline{m}_{ij}^2}{\sigma_{ij}^2}$$

$$e = \frac{(1 - \rho)^2(\overline{m}_i\,\hat{p}_{ij} - \overline{m}_{ij})^2}{(\sigma_{ij}^2 - 2\rho\,\hat{p}_{ij}\,\sigma_i\,\sigma_{ij} + \sigma_i^2\,\hat{p}_{ij}^2)}$$

$$L^2(m_i) = \frac{(m_i - \overline{m}_i)^2}{\sigma_i^2} - \frac{2\rho(m_i - \overline{m}_i)(m_i\hat{p}_{ij} - \overline{m}_{ij})}{\sigma_i\sigma_{ij}} + \frac{(m_i\hat{p}_{ij} - \overline{m}_{ij})^2}{\sigma_{ij}^2} \qquad (B.2)$$

and

$$k = \frac{1}{2\pi\sqrt{(1 - \rho^2)}\,\sigma_i\sigma_{ij}}$$

The probability mass of $\hat{p}_{ij}$ which lies between $t_i$ and $t_2$ can be stated as follows for $0 < \hat{p}_{ij} < 1$

$$P(t_1 < \hat{p}_{ij} < t_2) = C[(K(1 - \rho^2)\,e^{\frac{-v}{2(1-\rho^2)}} \int_{t_1}^{t_2} \frac{1}{a}\,dt$$

$$+ K^*(1-\rho^2)\int_{t_1}^{t_2}\frac{e^{\frac{-L^2(n)}{2(1-\rho^2)}}}{\alpha}dt + K^*\sqrt{(1-\rho^2)}\int_{t_1}^{t_2}\frac{b}{a^{\frac{3}{2}}}e^{\frac{-e}{2(1-\rho^2)}}$$

$$*\left(\int_{-\sqrt{g}}^{\infty} e^{\frac{x^2}{2}}dx - \int_{\sqrt{h}}^{\infty} e^{\frac{-x^2}{2}}dx\right)]dt \qquad \text{(B.3)}$$

The expression (B.3) has four integrals, The first can be evaluated using a table of integrals . The second does not have closed forms and must be integrated numerically. To integrate the last two integrals, we can begin by replacing the integral in the expression by a polynomial expansion which is good within the entire range of x, $0 < x < \infty$. We can see that the integration $f(\hat{p}_{ij})$ requires the evaluation of a complicated expression. In order to implement the estimation scheme, it would be expedient to use a standard normal distribution. An error will be introduced by using the standard normal approximation.

$$\text{If} \quad \frac{v-e}{1-\rho} >> 0 \qquad \text{(B.4)}$$

and

$$\frac{L^2(n)-e}{1-\rho} >> 0$$

then $f(\hat{p}_{ij})$ have been approximated by

$$\tilde{f}(\hat{p}_{ij}) \approx \left(\frac{1}{2\sqrt{\pi}}\right) * \left(\frac{\beta}{\alpha^{\frac{3}{2}}\sigma_i\,\sigma_{ij}}\right) e^{\frac{-u^2}{2}} \qquad \text{(B.5)}$$

Where

$$u = \frac{(\bar{m}_i \hat{p}_{ij} - \bar{m}_{ij})}{\sqrt{\sigma_{ij}^2 - 2\rho \hat{p}_{ij} \sigma_i \sigma_{ij} + \sigma_i^2 \hat{p}_{ij}^2}} \tag{B.6}$$

If we let

$$u(t) = \frac{(\bar{m}_i t - \bar{m}_{ij})}{\sqrt{\sigma_{ij}^2 - 2\rho t \sigma_i \sigma_{ij} + \sigma_i^2 t^2}} \tag{B.7}$$

$$\text{then } \frac{du(t)}{dt} = \frac{b}{(a^{\frac{3}{2}} \sigma_i \sigma_{ij})}$$

$$P(t_1 < \hat{p}_{ij} < t_2) = \frac{C}{\sqrt{2\pi}} \int_{u(t_1)}^{u(t_2)} e^{\frac{-x^2}{2}} dx \tag{B.8}$$

Where C is determined a posteriori, so the probability p is unity.

The approximated error $\delta(\hat{p}_{ij})$ is

$$\delta(\hat{p}_{ij}) = | f(\hat{p}_{ij}) - \tilde{f}^*(\hat{p}_{ij}) | \tag{B.9}$$

where $\tilde{f}^*(\hat{p}_{ij})$ denots the approximation of $f(\hat{p}_{ij})$.

Young (1977) gave the final inequality as follows,

$$\delta(p_{ij}) \leqslant | e^{\frac{-v}{2(1-\rho^2)}} [\sqrt{1-\rho^2} - \frac{\sqrt{v}}{2}] - e^{\frac{-L^2(n)}{2(1-\rho^2)}} [\sqrt{1-\rho^2} + \frac{\sqrt{v}}{2}] | \tag{B.10}$$

Which is the desired error term for the normal approximation to the

$f(\hat{p}_{ij})$ density function.

In summary, using the bivariate normal approximation, the joint density function $f(m_i, m_{ij})$ can be expressed in terms of $m_i$ and $m_{ij}$. Then distribution for $\hat{p}_{ij}$ may be obtained from $f(m_i,m_{ij})$. Under certain approximation conditions, the probability mass of $\hat{p}_{ij}$ which lies between $t_1$ and $t_2$ is easily calculated using the normal distribution scheme.

The vita has been removed from the scanned document