3-1-1993

CLASSIFICATION OF MULTISPECTRAL IMAGE DATA WITH SPATIAL-TEMPORAL CONTEXT

Byeungwoo Jeon
Purdue University School of Electrical Engineering

David Landgrebe
Purdue University School of Electrical Engineering

CLASSIFICATION OF MULTISPECTRAL IMAGE DATA WITH SPATIAL-TEMPORAL CONTEXT

Byeungwoo Jeon
David Landgrebe

TR-EE 93-15
March, 1993

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907-1285

This work was sponsored in part by NASA under Grant NAGW-925

TABLE OF CONTENTS

Page

ABSTRACT

Pattern recognition technology has had a very important role in many fields of application including image processing, computer vision, remote sensing, etc. The advent of more powerful sensor systems should enable one to extract far more detailed information than ever before from observed data, but to realize this goal requires the development of concomitant data analysis techniques which can utilize the full potential of the observed data.

This report investigates classification using spatial and/or temporal contextual information. Although contextual information has been an important and powerful data analysis clue for the human-analyst, the lack of a good contextual classification scheme especially which can both use spatial and temporal context has not allowed its usefulness to be put to full use.

Two different approaches to spatial-temporal contextual classification are investigated. One is based on ***statistical*** spatial-temporal contextual classification, and the other is based on ***decision fusion*** of temporal data sets which are classified individually with spatial contexts.

In the first approach, a general form of ***maximum a posterior*** spatial-temporal contextual classifier is derived after spatial and temporal neighbors are defined. Joint prior probabilities of the classes of each pixel and its spatial neighbors are modeled by the Gibbs random field. The classification is performed in a recursive manner to allow a computationally efficient contextual classification.

In the second approach based on the decision fusion, each temporal data set is separately fed into the local classifier and a final classification is performed by summarizing the local class decisions with an optimum decision fusion rule which is derived based on the minimum expected cost. The new decision fusion rule is designed to handle not only data set reliabilities but also classwise reliabilities of each data set.

Experimental results with three temporal Landsat Thematic Mapper data show significant improvement of classification accuracy over non-contextual pixelwise classifier. These spatial-temporal contextual classifiers will find their use in many real applications of remote sensing, especially when the classification accuracy is important.

CHAPTER 1
INTRODUCTION

1.1 Classification with Spatial and Temporal Contextual lnformation

For decades, 'the technology of remote sensing has been successfully applied in many interdisciplinary applications of Earth observational data, and multispectral data have been extensively used in the classification. Recent development in sensor technology and solid state devices allows spatially and spectrally far more rich information-bearing data sets. Note that multispectral image data are very complex entities that have not only spectral attributes but also rich spatial and temporal attributes as in Fig. 1.1.

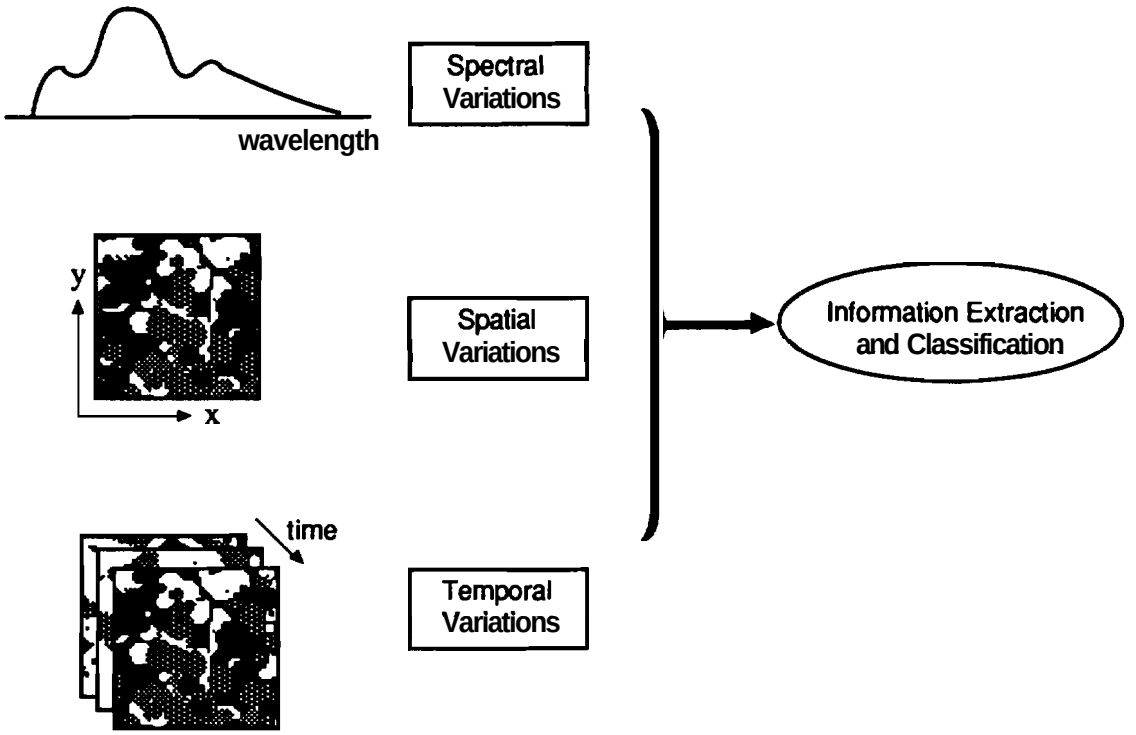


Figure 1.1 Spectral, Spatial and Temporal Variations in Images.

The availability of temporal data sets over the same scene makes it possible to extract valuable temporal characteristics of surface covers, which are of interest in applications requiring to detect spectral or spatial characteristic changes over time. Proper utilization of this spatial and temporal contextual information, in addition to spectral information, can improve the classification performance significantly in many applications compared to the conventional pixel-wise classification.

In part, due to the lack of good framework for using both spatial and temporal attributes in addition to spectral features, conventional approaches in the analysis of remotely sensed data have been mainly limited to pixel-wise classification. The objective of this research is development of a classification algorithm which can utilize both spatial and temporal contextual information in addition to spectral attributes in an efficient and effective way.

Although there has been much research (Kittler and Föglein 84) on the spatial contextual classification and temporal contextual classification, there have been only a few works utilizing both spatial and temporal contextual information. Two different approaches to spatial-temporal contextual classification are investigated. One is based on statistical spatial-temporal contextual classification, and the other is based on a decision fusion approach in multisource classification.

1.2 Organization of the Report

The outline of this report is as follows.

In Chapter 2, a spatial-temporal contextual classifier which finds the best set of class assiignments in the sense of ***maximum a posteriori*** probability (MAP) is formulated. With a few assumptions, this spatial-temporal contextual classifier is simplified into a more manageable form consisting of spatial and temporal contextual classifier parts.

The spatial contextual part in the spatial-temporal contextual classifier derived in Chapter 2 is applied to spatial classification in Chapter 3. Several models are

presented which allow computation of the conditional joint probability and prior probability in spatial contextual classification, with discussion of their computational aspects. Experimental results of this spatial contextual classifier are presented.

Chapter 4 addresses various methodologies in temporal contextual classification with an application for the temporal contextual classifier part introduced in Chapter 2 in mind. A decision fusion-based approach in temporal contextual classification is developed and its performance is compared with that of the conventional data fusion-based classifiers.

The two constituent contextual parts developed in Chapter 3 and 4 are combined for spatial-temporal contextual classification in Chapter 5 and experimental results on various spatial-temporal classifiers discussed so far are compared. The data fusion-based spatial-temporal classifier designed in Chapter 2 is modified to be used in the decision fusion-based approach. After presenting the experimental results on the spatial-temporal contextual classification, there follow conclusions and suggestions for future research regarding the spatial-temporal contextual classification.

CHAPTER 2

DESIGN OF A SPATIAL-TEMPORAL CONTEXTUAL CLASSIFIER

2.1 Introduction

In recent years, considerable research effort has been concentrated on extracting more information from a given data set. In pattern classification problems, this detailed information enables one to go deeper into the, so called, information tree (Landgrebe 78), *i.e.* the more detailed data now becoming available makes it possible to discriminate between classes of greater detail than previously possible. For this purpose, sensors with very fine spectral and spatial resolution are being put to use. Besides the development of new sensors, research is being carried out to find more accurate and powerful data analysis techniques. Most information extraction techniques rely on features pertaining to only one pixel location at a time. Although the spectral variability of a pixel can provide substantial discriminating power due to the increasingly fine spectral resolution now becoming available, confining analysis methods to only a single pixel at a time surely doesn't exploit the full information potential of newly emerging data.

Additional information is available from the relationship between pixels. This is called as "contextual" information. Context as used here is intended to mean spatial, temporal and/or spatial-temporal relationships between pixels. A contextual pattern classifier refers to a classifier which can utilize information from this interpixel relationship. The informative nature of this information source in human perception has such this contextual information an indispensable clue which is extensively relied upon in the manual interpretation of aerial photography. A simultaneous use of this spatial and/or temporal context can push the performance limitation further down so that more accurate and detailed classification result can be obtained.

There can be basically two different types of information which can be extracted from the data (Kittler and Föglein 84). One is interpixel dependency context between class labels, and the other is interpixel correlation context between pixel values. Both contexts exist spatially and temporally. Though contextual information is not restricted to only these two types (for example, contextual information can be obtained from shape, size, or direction, etc.), a main focus of this research will be so confined.

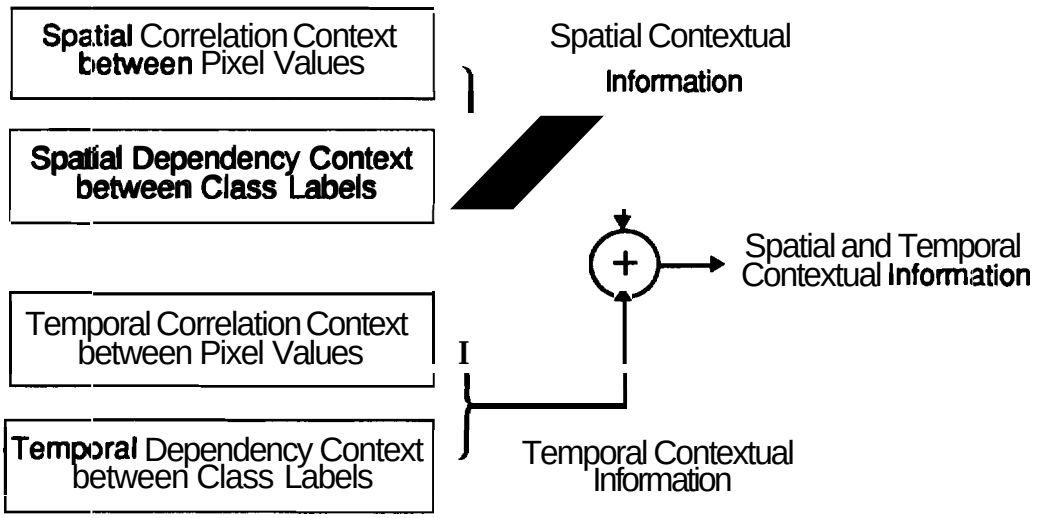


Figure 2.1 Sources of Spatial and Temporal Contextual Information.

The reason for the class label dependency correlation contexts being existent between class labels can be understood in following way. There are certain classes which are more likely to be found adjacently than others. In the same token, some classes are seldom found in proximity. Therefore, non-trivial information can be drawn from the relative assignments of neighboring class labels. Also, in many remotely sensed images, objects on the ground are much larger than the pixel size so that neighboring pixels are very likely to come from the same class and form a homogeneous region. This means that a pixel may be expected to be from the same class as its neighboring pixels. This property is successfully exploited in ECHO (Extraction and classification of Homogeneous Qbjects) classifier (Kettig and Landgrebe 76, Landgrebe 80) which first finds homogeneous regions to perform classification on per object basis. Though this

class label dependency context might not provide more detailed information in a discriminating process, in most of cases[1] a proper treatment of this contextual information can produce a classification result with far fewer errors.

Depending on the purpose of the usage, this interpixel class label dependency context can be divided further into two different types. One type of inter-pixel class label dependency context can be used to impose a local homogeneity of class labels in spatial or temporal proximity. In this case, the class label dependency context will be used for a sort of smoothing of the class label variability inside a local window. In other applications, one can use this class dependency context to impose on classifiers, the statistical likelihood of co-occurrence of the class labels in spatial or temporal proximity*. A good example of this type of usage can be found in a land cover discrimination application in an agricultural area where (temporal) class transition probabilities are used to model the known land use pattern over time and fed into a multi-temporal classification process as the temporal interpixel class dependency context.

In many cases, pixel values (or, feature vectors) exhibit significantly high spatial correlation between spatially adjacent pixels. Spatial correlation coefficients between pixels generally differ according to the distance between pixels and the spectral bands. Proper exploitation of spatial correlation context can make it possible to differentiate classes in more detail than would be possible without additional spatial correlation contextual information, however, the inclusion of spatial correlation factors in classifiers requires paying the price of increased computational complexity as compared to pixelwise classification (Khazenie and Crawford 90, Yu and Fu 83). It also tends to require a more highly trained user.

The spatial correlation which is class-unconditionally computed has generally a higher value and a slower decreasing rate vs. pixel separation than the class-conditionally computed quantity. On the other hand, the class-conditional spatial correlation decays rather quickly if the spatial distance between pixels is increased. This fact was exploited in the ECHO classifier (Kettig and Landgrebe

---

1 An exception can be the case when the relative distribution of class labels itself can indicate a particular information class. This will be discussed in next section in a review of S. W. Warton's work (Warton 82).

2 The meaning of spatial or temporal proximity will be formally defined in section 2.3.

76, Landgrebe 80) in which pixels inside an object are assumed to be spatially independent and the likelihood values of an object are computed as a simple product of the likelihood values of each pixel belonging to that object. This interpixel correlation can also exists temporally. Temporal correlation contexts may be useful in specific applications. But care must be taken in using this temporal correlation context, since there can be potentially significant difference!; in the ternporal data sets, such as the difference of the atmospheric condition.

In this report, attention will be given only to using the two spatial contexts (correlation between class labels and pixel values), and temporal class label dependency context. Before going further to develop a spatial-temporal classification framework, some of the related works in this direction are reviewed.

2.2 Related Works in Spatial and/or Temporal Contextual Classification

A tutorial overview of various techniques for using contextual information in different pattern recognition problems can be found in (Toussaint 78). Among many works in diverse fields of application, J. Kittler and J. Föglein (Kittler and Föglein 84), N. L. Hjort and E. Mohn (Hjort and Mohn 87) and R. M. Haralick (Haralick 83) specifically dealt with the use of contextual information in image classification problems. Especially, J. Kittler and J. Föglein (Kittler ancl Föglein 84) and J. R. G. Townshend (Townshend 83) provide extensive overviews of spatial contextual classifiers designed primarily for remote sensing applications.

2.2.1 Related Works in Spatial Contextual Classification

Broadly speaking, the methodologies to take spatial context into account can be categorized into three different groups (Kittler and Föglein 84) according to how the contextual information is used.

Post-processing approach
Pre-processing approach
Simultaneous processing approach

Post-processing type contextual classifiers perform a post-processing such as filtering or applying syntactic rule after the pixel-wise classification. One example of filters available for post-processing is the majority filter (Drake et *al*. 87) which counts the votes of classification results inside a given-sized window and re-assigns to the center pixel of that window, a particular class which most of the pixels inside the window choose. Small classes mainly composed of scattered noise pixels might be merged to neighboring large classes after the majority filtering (Guo and Moore 91). Another approach which can be categorized into this group is that of (Warton 82, Zhang et *al*. 88) which extracts, in the first pass, new feature vectors composed by class labels of pixels in a given neighborhood after pixelwise classification and then, in the second pass, uses these vectors to obtain final decisions. Contextual information is used in the second pass. These classifiers are especially useful in land-cover classification of urban areas in which information classes consist of several spectrally dissimilar components. For example, a class "residential area" may contain spectrally different components of house, road, lawn, etc. By accounting for the components' frequency distribution, such classes as "high density residential area" and "low density residential area" can be differentiated. However, a common handicap of this category is to try to recover the information already lost in the pixel-wise classification phase, which inevitably confines its success to a certain limit.

The pre-processing type approaches are based on a region growing or object extraction process. A given scene is divided into distinct homogeneous regions by using an appropriate homogeneity test and each homogeneous region is classified on an object or per-field basis. One procedure of this category is ECHO which uses a conjunctive, object-seeking method as the tool for region finding (Kettig and Landgrebe 76, Landgrebe 80). Several varieties of algorithms have been proposed with different statistical measures of homogeneity. In a study of (Kusaka et *al.* 89), primitive regions with nearly uniform colors (*i.e.*, spectral responses) were found with edge-based segmentation. Classification of the primitive regions was obtained using various spatial features computed for each regions. S. L. Sclove (Sclove 81) and H. M. Kalayeh and D. A. Landgrebe (Kalayeh and Landgrebe 87) developed similar object classifiers which could utilize spatial correlation contexts through Markov

random field modeling of feature vectors, but under an assumption that the objects were already extracted. A common problem of these segmentation-based algorithms is that the classification result is heavily dependent on the success of the region finding process, which may be as difficult as the classification itself.

Classifiens of the third type of approaches account for the spectral and spatial contextual information simultaneously to make the most use of the available information. One of the straightforward way of this is the so-called, stacked vector approach, which adds, to the original spectral feature vector, new components of features which can carry spatial contexts. Additional components can be derived, for example, from some texture descriptors such as Fourier coefficients or co-occurrence matrices (Haralick *et al.* 73). The stacked vector approach has an inherent problem of excessive dimensionality of augmented feature vectors and poor performances at the object boundaries since the *texturemeasures are based on a multipixel sized region. Due to these shortcomings of the stacked vector approach, simultaneous utilization of contextual information is accomplished often by setting up a probabilistic model such as the spatial stochastic model (Yu and Fu 83) which can effectively incorporate contextual information in the resulting classifier. Classifiers in this category usually assume a local dependency of a pixel on its neighbor:; and the classification results are obtained in a recursive way. The procedure of the contextual classification proposed in this report falls into this category.

Other well known procedures in this category are those based on relaxation (Rosenfeld *et a/.* 76), which is an iterative procedure making fuzzy or probabilislic decisions at each iteration and then successively updating those decisions, according to a selected compatibility function and previous decisions (Eklundh *et al.* 80, Richards *eta/.* 81, Kalayeh and Landgrebe 82).

There are several reports on comparative tests of various spatial classifiers. G. Palubinskas (Palubinskas 88) compared performances of various object classifiers with images modeled by a second order causal autoregressive model ancl observed that the performance of the object classifiers was much better than per-pixel classifier. In a Monte Carlo simulation study in (Mohn *et al.* 87), E. Mohn *eta/.* observed that, compared to non-contextual rules, contextual

methods usually reduced error rates considerably and the performance increase was particularly significant in homogeneous areas and on borders with simple structures. Except for the case with very high spatial correlation, however, they found generally no gain in using contextual methods on such scenes with little or no structure at all.

Although there have been many spatial classifiers which can utilize class label dependency contextual information, only a few researchers investigated seriously the problem of estimating 'the class label dependency contexts. J. C. Tilton (Tilton et *al.* 82) and G. R. Dattatreya (Dattatreya 91) investigated unbiased estimation algorithms for evaluating the class label dependency context from the unlabelled samples.

2.2.2 Related Works in Temporal Contextual Classification

In the case of the temporal contextual classification problem, there have been a stacked vector approach (Fleming and Hoffer 77), the so-called, cascade classifier (Swain 78a), a stochastic model based approach (Kalayeh and Landgrebe 86), and an approach based on a mathematical model for spectral development such as a regression model or growth profile (Crist and Malia 80). The stacked vector approach has the same problem as in the spatial classification case. Compared to the cascade classifier, which assumes class-conditional independence of feature vectors of different temporal data sets, the stochastic model based approach (Kalayeh and Landgrebe 86) considers the ground cover types as a stochastic system with a non-stationary Gaussian process as an input and temporal variations of feature vectors as an output under the assumption that the class doesn't change over time; it utilizes the temporal interpixel correlation context in the classification. Since it assumes same set of classes for each temporal data set and requires classes not to change over time, in the training stage, all given temporal data sets must be processed together to define spectral classes. This simultaneous treatment of all given temporal data sets in the training stage increases the total number of necessary spectral classes. This problem is avoided in the cascade classifier by allowing class changes over time.

2.2.3 Related Works in Spatial-Temporal Contextual Classification

Compared to the spatial and temporal contextual classifier cases, there have been only a few reports on spatial-temporal contextual classification. N. Khazenie and M. M. Crawford (Khazenie and Crawford 90) reported a procedure based on an extended version of the autocorrelation model proposed by N. L. Hjort, E. Mohn and G. Strovik (Hjort *et al.* 85, Hjort and Mohn 85) to account for both spatial and temporal correlation structures. This is based on the assumption that the observed process is a sum of two independent processes, one having a class dependent structure and the other, being an autocorrelated noise process. The noise process accounts for both spatial and temporal correlation. Under the assumption of a certain form of the noise covariance matrix, the conditional joint probability of spatial and temporal neighbors; are computed. This approach is very expensive from a computational standpoint.

Although it is almost impossible to compile a comprehensive and exhaustive list of all previous works related to the spatial, temporal and spatial-temporal contextual classifiers, some of the previous works are summarized in Table 2.1 ~ 2.3. Depending on how the contextual information is incorporated into the classifiers;, temporal and spatial-temporal classifiers are also categorized into the same three types as the spatial contextual classifiers.

Table 2.1 Classifiers with Class Label Dependency Context.

Classifier	Category	References
with local frequency distribution of class labels	SP : post-proc.	(Warton 82) (Zhang *et al.* 88)
Majority filtering of pixel-wise classification map	SP : post-proc.	(Drake *et al.* 87)
with template histogram matching, iterative majority filtering, small class merging and class boundary detection	SP : post-proc.	(Guo and Moore 87)
Homogeneous region extraction based on conjunctive, object-seeking method	SP : pre-proc.	(Kettig & Landgrebe 76) (Landgrebe 80)
Nearest neighbor based classifier (spatial class transition probabilities estimated from training samples)	SP : simul.	(Welch and Salter 73)
Probabilistic relaxation algorithm with compatibility function between class labels	SP : simul.	(Eklundh *et al.* 80) (Kalayeh & Landgrebe 82)
Probabilistic and fuzzy relaxation	SP : simul.	(Zenzo *et al.* 87a,b)
Stochastic relaxation based on Markov random field	SP : simul.	(Zhang & Haralick 90)
Geometric model for class label joint probability	SP : simul.	(Owen 84)
with two dimensional Markovian model	SP : simul.	(Haslett 85)
Recursive estimation of joint probability of class labels	SP : simul.	(Haralick *et al.* 84) (Haralick & Joo 86)
Iterative Conditional Models (ICM)	SP : simul.	(Besag 86)
Based on compound decision theory with p-context array	SP : simul.	(Swain et *al.* 81)
Cascade classifier (require temporal class transition probabilities)	TP : simul.	(Swain 78a)

SP : With spatial contextual information only TP : With temporal contextual information only

Table 2.2 Classifiers with Interpixel Correlation Context.

Classifier	Category	References
Object classifier (objects assumed to be extracted)	SP : simul.	(Sclove 81) (Kalayeh & Landgrebe 86)
Region classification using spatial features after edge-based segmentation.	SP : pre-proc.	(Kusaka et *al.* 89) (Kusaka & Kawata 91)
Based on spatial stochastic model	SP : simul.	(Yu and Fu 83)
Based on temporal stochastic model	TP : simul.	(Kalayeh & Landgrebe 86)

SP : With spatial contextual information only TP : With temporal contextual information only

Table 2.3 Classifiers with Both Class Label Dependency and Interpixel Correlation Context.

Classifier	Category	References
Recursive classifier	SP : simul.	(Kiffler & Föglein 84) (Kittler & Pairman 85)
Autocorrelation model for spatial correlation between pixels. Markov random field model for class label dependency	SP : simul.	(Hjort et *al.* 85) (Hjort & Mohn 87)
Autocorrelation model for spatial/temporal correlation between pixels. Markov random field model for spatial class label dependency	SPTP: simul.	(Khazenie & Crawford 90)

SP : With spatial contextual information only SPTP : With spatial-temporal contextual information

2.3 Design of the Spatial-Temporal Contextual Classifier

2.3.1 Introduction

In this section, a general contextual classification framework under which both spatial and temporal contextual information can be utilized is investigated. After spatial and temporal neighbors are defined, a general form of a ***maximum a*** posteriori spatial-temporal contextual classifier is derived. This contextual classifier is simplified under several assumptions.

Noting that the spatial-temporal contextual classification can be thought as a specific application of the more general problem of how to effectively make the most of all available information sources to attain the "best" result. The meaning for "best" might differ problem to problem, and in a classification problem, classification accuracy can be one of the criterion to claim for being "best." The problem of spatial-temporal contextual classification will be considered as a special example of multisource classification (Benediktsson etal. 90, Lee 87) in which the spatial, temporal and/or spatial-temporal contextual information is considered as each being a separate information source. Among many possibilities in simultaneously dealing with various information sources, the decision fusion approach will be investigated, and it will be addressed in detail in Chapter 4.

2.3.2 Spatial-Temporal Contextual Classification

Suppose there are p multitemporal remotely sensed data sets $\{X(1), X(2), ---, X(p)\}$ taken over the same location. These multitemporal data sets are assumed to be registered to each other.

$X(k)$, k = 1, ---, p, denotes the k^{th} temporal data set. The size of each data set is I by J and defined on the lattice $L \equiv \{r = (i, j) \mid 1 \le i \le I, 1 \le j \le J\}$. $x_k(r)$ refers to the feature vector of a pixel at spatial location (or site) r, $r \in L$, on the given k^{th} data set $X(k)$. Therefore, $X(k)$ can be written as $X(k) = \{x_k(r) \mid r \in L\}$, the set of all feature vectors of $x_k(r)$ on L. The class corresponding to $x_k(r)$ is denoted by $c_k(r)$. $c_k(r)$ takes one of the classes in $\Omega_k = \{\omega_{k,1}, ---, \omega_{k\,M_k}\}$ which is the set of all distinguishable classes in the k^{th} data set. M_k is the total number of elements in Ω_k.

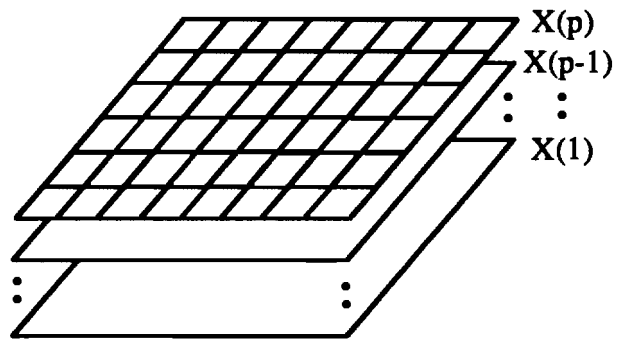


Figure 2.2 p Multitemporal Data Sets.

Since each temporal data set is separately analyzed in the training stage, the Ω_k's and the M_k's are not necessarily the same for different k's. C(k) is defined similarly as the set of class labels of all the pixels in X(k), *i.e.*, $C(k) = \{c_k(r) \mid r \in L\}$.

Let N_S denote a spatial neighborhood. Examples of N_S are given in Fig. 2.3. At the boundary, a spatial neighborhood has a fewer number of pixels.

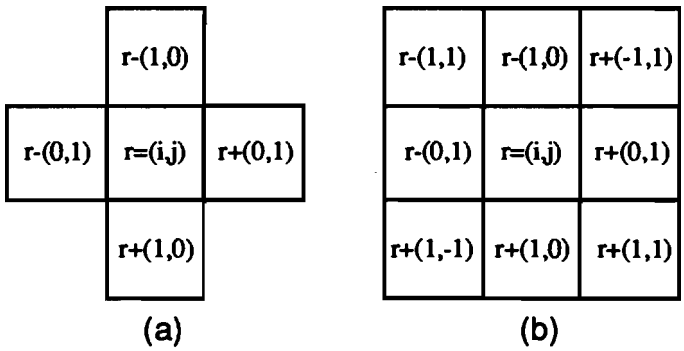


Figure 2.3 Examples of Spatial Neighborhood Systems. (a) First order spatial neighborhood system : $N_S = ((f1, 0), (0, \pm 1))$. (b) Second order spatial neighborhood system : $N_S = \{(\pm 1, 0), (0, \pm 1), (1, \pm 1), (-1, \pm 1)\}$.

Although it is also possible to use different spatial neighborhoods for each X(k), k = 1, ---, p, in this report, the same N_S is used for each temporal data set for simplicity's sake. Define $X_{S,k}(r)$, the set of spatial neighbors of $x_k(r)$, as,

$$X_{S,k}(r) \equiv \{ x_k(r+v) \mid v \in N_S \},\ r \in L \text{ and } k = 1, \text{---}, p$$

It consists of the pixels in the spatial vicinity of $x_k(r)$, (that is, under the first order spatial neighborhood system, it consists of the adjacent pixels of $x_k(r)$ in the north, south, east and west). Since $X_{S,k}(r)$ doesn't contain the pixel $x_k(r)$, another notation, $X'_{S,k}(r)$, is introduced to address the set of the pixel $x_k(r)$ itself and its spatial neighbors as,

$$X'_{S,k}(r) \equiv X_{S,k}(r) \cup \{ x_k(r) \}$$

If N'_S is defined as $N_S \cup \{(0, 0)\}$, then $X'_{S,k}(r)$ can be written as $\{x_k(r+v) \mid v \in N'_S\}$. Similarly, $C_{S,k}(r)$ and $C'_{S,k}(r)$, the set of classes corresponding to $X_{S,k}(r)$ and $X'_{S,k}(r)$, respectively, $k = 1, \text{---}, p$, are defined as,

$$C_{S,k}(r) \equiv \{ c_k(r+v) \mid v \in N_S \} \in \Omega_k^4$$
$$C'_{S,k}(r) \equiv C_{S,k}(r) \cup \{ c_k(r) \}$$

Here, Ω_k^4 denotes the set of all distinguishable classes that $X_{S,k}(r)$ can have. In the same way, notation related to the temporal neighbors are introduced. $X_{T,k}(r)$, the set of temporal neighbors of $x_k(r)$ and $C_{T,k}(r)$, the set of corresponding classes to $X_{T,k}(r)$ are defined as,

$$X_{T,k}(r) \equiv \bigcup_{t=1}^{k-1} X'_{S,k-t}(r)$$

$$C_{T,k}(r) \equiv \bigcup_{t=1}^{k-1} C'_{S,k-t}(r) \in \prod_{t=1}^{k-1} \Omega_{k-t}^5$$

Ω_k^5 is a set of all distinguishable classes that $X'_{S,k}(r)$ can have. $X_{T,k}(r)$ consists of all the temporally previous pixels of $x_k(r)$ and their spatial neighbors.

The elements in the union of spatial and temporal neighbors of $x_k(r)$, that is, the union of $X_{S,k}(r)$ and $X_{T,k}(r)$, are called the spatial-temporal neighbors of $x_k(r)$. $\xi_{X\,k}(r)$ which is defined as follows, is then the set of $x_k(r)$ and its spatial-temporal neighbor:;.

$$\xi_{X,k}(r) \equiv \{ x_k(r) \} \cup X_{S,k}(r) \cup X_{T,k}(r)$$
$$\xi_{C,k}(r) \equiv \{ c_k(r) \} \cup C_{S,k}(r) \cup C_{T,k}(r)$$

$\xi_{C\,k}(r)$ is Ihe set of classes corresponding to $\xi_{X\,k}(r)$. (see Fig. 2.4 for a graphical illustrationi of spatial and temporal neighbors).

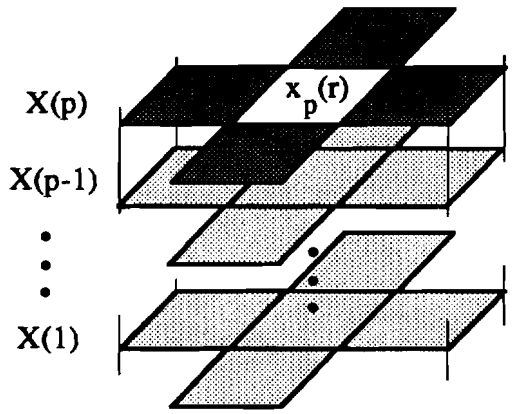


Figure 2.4 **Spatial and Temporal Neighbors of $x_p(r)$ under the First Order Spatial Neighborhood System.; ☐ Temporal neighbors of $x_p(r)$: $X_{T,p}(r)$; ■ Spatial neighbors of $x_p(r)$: $X_{S,p}(r)$; ▩ + ■ Spatial-temporal neighbors of $x_p(r)$.**

From now on, bold faced symbols will be used for random variables and plain symbols will be used for specific realizations of the corresponding random variables whenever there is a need to so differentiate. Also, for notational simplicity, the spatial location argument "(r)" will be dropped in notation where no confusion can result. For example, $\mathbf{x}_p$ means $\mathbf{x}_p(r)$, $r \in L$. Also the realization of the random variables will be omitted in equations whenever there is no confusion by doing so. That is, $P\{\mathbf{x}_k(r)\}$ means $P\{\mathbf{x}_k(r) = x\}$ and so on.

The pixels in the p^{th} temporal data set $\mathbf{X}(p)$, are to be classified to one of the M_p classes using the given multitemporal data sets $\{\mathbf{X}(1), \mathbf{X}(2), ---, \mathbf{X}(p)\}$. The best set of class labels of pixels in $\mathbf{X}(p)$ in the maximum a posterioriprobability (MAP) sense can be obtained using eq. (2.1).

$$\hat{C}_{MAP} = \underset{C}{\text{argmax}} \left[P\{\mathbf{C}(p) = C \mid \mathbf{X}(1) = \mathbf{X}(1), ---, \mathbf{X}(p) = \mathbf{X}(p)\}\right] \qquad (2.1)$$

Even though eq. (2.1) is optimal in the sense of maximum a posteriori *probability*, a direct computation and maximization of $P\{\mathbf{C}(p) = C \mid \mathbf{X}(1), ---, \mathbf{X}(p)\}$ is, in most practical applications, too complex to be useful even for a small sized scene. For example, with M_p classes in $\mathbf{X}(p)$, the total number of possible combinations of the classes amounts to M_p^{IJ}. This easily becomes an explosive number for an even moderate M_p. One of the plausible remedies to avoid this difficulty is to assume that all necessary contextual information can be manifested by its spatial and temporal neighbors. An example of spatial-temporal neighbors of $\mathbf{x}_p(r)$ in case of a first order neighborhood is shown in Fig. 2.3. In many cases, this can be quite reasonable and also a very practical assumption since the interactions between pixels decrease rapidly as the (spatial and temporal) distances between pixels increase.

$$H_{SPTP}(c; r, k) \equiv P\{\mathbf{c}_k = c \mid \mathbf{x}_k = x_k, \mathbf{X}_{S,k} = X_{S,k}, \mathbf{X}_{T,k} = X_{T,k}\} \qquad (2.2)$$

Under this practical assumption, define a spatial-temporal contextual classifier, $H_{SPTP}(c; r, k)$, $r \in L$ and $c \in \Omega_k$, $k = 1, ---, p$ as in eq. (2.2). In the case of $k = 1$, $\mathbf{X}_{T,k}$ is understood as an empty set since there is no temporally previous data. Thus, when $k = 1$, $H_{SPTP}(c; r, k)$ is $P\{\mathbf{c}_k = c \mid \mathbf{x}_k = x_k, \mathbf{X}_{S,k} = X_{S,k}\}$. The spatial-temporal contextual classification can be achieved then by finding the class $c \in \Omega_p$ which maximizes $H_{SPTP}(c; r, p)$. To simplify eq. (2.2) into a computationally more manageable form, several assumptions are made as in eq. (2.3.a,b) and eq. (2.4). The first assumption in eq. (2.3.a,b) is about the classes between spatial and temporal neighbors.

Assumption 1.

For any k, $1 \le k \le p$, and for $\mathbf{C}_A$ and $\mathbf{C}_B$ defined below,

$$P\{\mathbf{c}_{k+1} \mid \mathbf{c}_k, \mathbf{C}_A\} = P\{\mathbf{c}_{k+1} \mid \mathbf{c}_k\} \quad (2.3.a)$$

$$P\{\mathbf{C}_{S,k} \mid \mathbf{c}_k, \mathbf{C}_B\} = P\{\mathbf{C}_{S,k} \mid \mathbf{c}_k\} \quad (2.3.b)$$

where,

$\mathbf{C}_A$ is any non-empty subset of $\xi_{C\,k}$ such that $\mathbf{C}_A \cap \{\mathbf{c}_k\} = \phi$. ϕ is the empty set.

$\mathbf{C}_B$ is any non-empty subset of $\xi_{C,k-1}$.

Equation (2.3.a) assumes that irrespective of the classes of the other spatial-temporal neighbors of $\mathbf{x}_k$, the temporal class dependency context is conveyed to $\mathbf{c}_{k+1}$ from its temporal neighbors only through $\mathbf{c}_k$. This assumption makes it possible to model the temporal class dependency with a simple class transition probability $P\{\mathbf{c}_{k+1} \mid \mathbf{c}_k\}$. Equation (2.3.b) is the spatial counterpart of the eq. (2.3.a), that is, $\mathbf{C}_{S\,k}$, the set of classes of the spatial neighbors of $\mathbf{x}_k$, is assumed to be only dependent on the class $\mathbf{c}_k$, irrespective of the classes of temporal neighbors; of $\mathbf{x}_k$.

Assumption 2.

For any k, $1 \le k \le p$, and for $\mathbf{X}_A$, $\mathbf{C}_A$, $\mathbf{X}_{others}$ and $\mathbf{C}_{others}$ defined below,

$$P\{\mathbf{X}_A \mid \mathbf{C}_A, \mathbf{X}_{others}, \mathbf{C}_{others}\} = P\{\mathbf{X}_A \mid \mathbf{C}_A\} \quad (2.4)$$

where,

$\mathbf{X}_A$ is any non-empty subset of $\mathbf{X}'_{S,k}$.

$\mathbf{C}_A$ is a set of the classes corresponding to $\mathbf{X}_A$.

$\mathbf{X}_{others}$ is any subset of $\xi_{X,p}$ such that $\mathbf{X}_{others} \cap \mathbf{X}'_{S,k} = \phi$.

$\mathbf{C}_{others}$ is any subset of $\xi_{C,p}$ such that $\mathbf{C}_{others} \cap \mathbf{C}'_{S,k} = \phi$.

($\mathbf{C}_{others}$ is not necessarily a set of classes corresponding to $\mathbf{X}_{others}$).

The second assumption is that the pixel values of $\mathbf{X}_A$ (any non-empty subset of $\mathbf{X}_{S,k}$) are affected only by the nature of pixels in $\mathbf{X}_A$, that is, correspondling class identities in $\mathbf{C}_A$, irrespective of the pixels ($\mathbf{X}_{others}$) or the classes ($\mathbf{C}_{others}$) of other temporal data sets. In other words, once the classes of a set of pixels at

one particular time are known, the values or classes of pixels at any other times do not provide any additional knowledge on the pixel values at that particular time. This is a little bit stronger than the conventional class-conditional independence assumption of different temporal data sets given below. Though eq. (2.4) implies the following relation, but, the reverse is not always true.

$$P\{ \mathbf{X}_A , \mathbf{X}_B \mid \mathbf{C}_A, \mathbf{C}_B \} = P\{ \mathbf{X}_A \mid \mathbf{C}_A \} P\{ \mathbf{X}_B \mid \mathbf{C}_B \}$$

where,

$\mathbf{X}_A$ and $\mathbf{X}_B$ are any subsets of pixels in different temporal data sets. $\mathbf{C}_A$ and $\mathbf{C}_B$ are the set of classes corresponding to $\mathbf{X}_A$ and $\mathbf{X}_B$, respectively.

Due to the implication of the class-conditional independence of temporally different data sets, under the assumption in eq. **(2.4)**, temporal correlation context is not counted in the classification. With the assumptions in eq. **(2.3.a,b)** and eq. **(2.4)**, the following theorems and lemmas which are useful in simplifying eq. (2.2) are derived in Appendix A. A direct consequence of the assumptions in eq. **(2.3.a,b)** is the following theorem which relates to the relationship between class labels of temporal neighbors.

***Theorem* 1.**

For any tand usuch that $1 \le t \le u \le p$,

$$P\{\eta_u \mid \eta_t, \mathbf{C}_{others}\} = P\{\eta_u \mid \eta_t\} = P\{\eta_u \mid \mathbf{c}_t\} \qquad (2.5)$$

where,

if u > t, η_u is either $\{\mathbf{c}_u\}$ or $\mathbf{C}'_{S,u}$. η_t is either $\{\mathbf{c}_t\}$ or $\mathbf{C}'_{S,t}$.

if u = t, $\eta_u = \mathbf{C}_{S,u}$ and $\eta_t = \{\mathbf{c}_t\}$.

$\mathbf{C}_{others}$ is any non-empty subset of $\xi_{C,t}$ such that $\mathbf{C}_{others} \cap \eta_u = \mathbf{C}_{others} \cap \eta_t = \phi$.

This theorem states that when u > t, the class $\mathbf{c}_u$ or the set of classes, $\mathbf{C}'_{S,u}$, $1 \le u \le p$, is dependent only on the nearest temporal neighbors $\mathbf{C}'_{S,t}$, or the nearest previous pixel, $\mathbf{c}_t$. If u = t, the probability of $\mathbf{C}_{S,u}$ given $\mathbf{c}_u$ and any non-empty subset of its temporal neighbors, $\xi_{C,t}$, is described as $P\{\mathbf{C}_{S,u} \mid \mathbf{c}_u\}$. Therefore, the set of class identities, $\mathbf{C}_{others}$ doesn't provide any supplementary information on $\mathbf{C}_{S,u}$ once the class identity $\mathbf{c}_u$ is available. Using this theorem,

the first order Markov dependency property of class labels, *i.e.*, $P\{ C'_{S,k} \mid C'_{S,k-1}, \text{---}, C'_{S,1} \} = P\{ C'_{S,k} \mid C'_{S,k-1} \}$ can be easily shown.

Lemma 1.

For C_{others}, η_u and η_t defined as in Theorem 1,

$$P\{C_{others} \mid \eta_u , \eta_t\} = P\{C_{others} \mid \eta_t\} = P\{C_{others} \mid c_t\} \quad (2.6.a)$$

$$P\{C_{T,k} \mid c_k, C_{S,k}\} = P\{C_{T,k} \mid c_k\} \quad (2.6.b)$$

Applying the Bayes theorem to eq. (2.5) results in eq. (2.6.a), which shows a similar relationship as in eq. (2.5) but in the temporally opposite direction; substituting $C_{others} = C_{T,k}$ and $\eta_t = \{c_k\}$, $\eta_u = C_{S,k}$ in eq. (2.6.a) yields eq. (2.6.b), which shows that the probability of $C_{T\,k}$ given c_k and $C_{S\,k}$ will be determined only by $C_{T\,k}$ and c_k. While Theorem 1 and Lemma 1 show the relationship between the class labels of temporal neighbors, the following theorem shows the relationship between feature vectors under the condition of given class labels.

Theorem 2.

For any t and u such that $1 \le t \le u \le p$, and for X_A, η_t and η_u defined as below,

$$P\{X_A \mid \eta_t, \eta_u\} = P\{X_A \mid \eta_t\} \quad (2.7)$$

Especially, if $X_A \cap X'_{S,t} = \phi$,

$$P\{X_A \mid \eta_t\} = P\{X_A \mid c_t\}$$

where,

if $u > t$,

η_t is either $\{c_t\}$ or $C'_{S,t}$. η_u is either $\{c_u\}$ or $C'_{S,u}$

X_A is any non-empty subset of $\xi_{X,t}$ such that $X_A \cap X'_{S,t}$ is either ϕ or $X'_{S,t}$.

if $u = t$,

$\eta_t = \{c_t\}$ and $\eta_u = C_{S,u}$

X_A is any non-empty subset of $\xi_{X,t-1}$

According to Theorem 2 which can be proved by applying the Lemma 1 with the assumption 2, when the class identity, c_t, or a set of class identities, $C'_{S,t}$, at a

certain time t ($1 \le t \le p$) is known, the class identity, c_t or a set of class identities, $C'_{S,u}$, at a later time u ($u > t$) doesn't affect the appearance of the pixels of time t or prior to the time t. In the case of u = t, knowledge of $C_{S,t}$ will be redundant in determining the appearances of the pixels in $\xi_{X\,t-1}$, the pixels observed prior to the time t, if the class identity $\eta_t = \{c_t\}$ is available.

Lemma 2.

$$\begin{aligned} P\{X_{T,k} \mid c_k, c_{k+1}\} &= P\{X_{T,k} \mid c_k\} \\ P\{X_{T,k} \mid c_k, C_{S,k}\} &= P\{X_{T,k} \mid c_k\} \\ P\{X'_{S,k} \mid c_k, c_{k+1}\} &= P\{X'_{S,k} \mid c_k\} \end{aligned} \tag{2.8}$$

Substituting the variables X_A, η_t and η_U with non-abstract quantities in Lemma 2 reveals the meaning of this theorem more clearly. By using assumption 2 and Lemma 1, the following lemma can be derived.

Lemma 3.

For any k, $1 \le k \le p$, and for X_{others} which is any non-empty subset of $\xi_{X,k-1}$,

$$P\{X'_{S,k} \mid c_k, X_{others}\} = P\{X'_{S,k} \mid c_k\} \tag{2.9}$$

This lemma shows that if c_k, the class identity of center pixel in $X'_{S,k}$, is known, X_{others} which is the set of the pixel values of temporally previous data sets, do not provide any additional information on the pixel values $X'_{S,k}$.

Using the results derived in the previous theorems and lemmas, the spatial-temporal contextual classifier in eq. (2.2) is simplified. Applying the result of Lemma 3 and the Bayes theorem, to $H_{SPTP}(\bullet; r, k)$ in eq. (2.2), for k = 2, ---, p, yields,

$$P\{c_k \mid X'_{S,k}, X_{T,k}\} = A_k \frac{P\{c_k \mid X'_{S,k}\}\, P\{c_k \mid X_{T,k}\}}{P\{c_k\}} \tag{2.10}$$

where, $A_k \equiv \dfrac{P\{X'_{S,k}\} P\{X_{T,k}\}}{P\{X_{S,k}, X_{T,k}\}}$.

Since A_k is not dependent on the particular class assigned to the pixel $\mathbf{x}_k(r)$, it doesn't need to be evaluated. Define the spatial contextual classifier $H_{SP}(c; r, k)$, $c \in \Omega_k$, $k = 1, \cdots, p$, as,

$$H_{SP}(c; r, k) \equiv P\{\mathbf{c}_k = c \mid \mathbf{X}'_{S,k} = X'_{S,k}\}, \quad c \in \Omega_k \tag{2.11}$$

This represents how much the spatial contextual information from Ihe pixels $\mathbf{x}_k(r)$ and $\mathbf{X}_{S\,k}(r)$ support the class assignment c to the pixel $\mathbf{x}_k(r)$. In the same way, the temporal contextual classifier $H_{TP}(c; r, k)$, $c \in \Omega_k$, $k = 2, \cdots, p$, is defined as,

$$H_{TP}(c; r, k) \equiv P\{\mathbf{c}_k = c \mid \mathbf{X}_{T,k} = X_{T,k}\}, \quad c \in \Omega_k \tag{2.12}$$

$H_{TP}(c; r, k)$ shows how much the spatial-temporal contextual information from the temporal neighbors $\mathbf{X}_{T,k} = \{\mathbf{X}'_{S,k-1}, \cdots, \mathbf{X}'_{S,1}\}$ advocates the class assignment of c to the pixel $\mathbf{x}_k(r)$. For $k = 1$, $H_{TP}(c; r, k)$ is defined as $P\{\mathbf{c}_k = c\}$. For $c \in \Omega_k$, $k = 2, \cdots, p$, substituting these $H_{SP}(\bullet; r, k)$ and $H_{TP}(\bullet; r, k)$ into eq. (2.10) leads to the following equation. For $c \in \Omega_k$, $k = 2, \cdots, p$,

$$H_{SPTP}(c; r, k) = A_k \frac{H_{SP}(c; r, k)\, H_{TP}(c; r, k)}{P\{\mathbf{c}_k = c\}} \tag{2.13}$$

In the case of $k = 1$, $H_{SPTP}(c; r, k)$ is $H_{SP}(c; r, k)$. Due to the assumptions in eq. (2.3.a,b), the temporal contextual classifier $H_{TP}(c; r, k)$ can be computed using $H_{SPTP}(d; r, k-1)$, $d \in \Omega_{k-1}$, and class transition probabilities between temporal neighbors in the $(k-1)^{th}$ data set and the k^{th} data set. That is, by applying Theorem 2 to eq. (2.12) and Bayes theorem, $H_{TP}(c; r, k)$ can be computed as,

$$H_{TP}(c; r, k) = P\{\mathbf{c}_k = c\} \sum_{d \in \Omega_{k-1}} \frac{H_{SPTP}(d; r, k-1)}{P\{\mathbf{c}_{k-1} = d\}} P\{\mathbf{c}_k = c \mid \mathbf{c}_{k-1} = d\}$$

$$= \sum_{d \in \Omega_{k-1}} H_{SPTP}(d; r, k-1) P\{\mathbf{c}_k = c \mid \mathbf{c}_{k-1} = d\} \tag{2.14}$$

This result is very similar to the case of cascade classifier (Swain 78a). But eq. (2.14) has a quantity reflecting spatial-temporal contexts from the temporal neighbors instead a quantity which reflects only the temporal context from the previous pixel as in (Swain 78a). The temporal contextual classifier $H_{TP}(c; r, k)$ passes the contextual information obtained from the spatial-temporal neighbors of $\mathbf{x}_{k-1}(r)$ to the classifier $H_{SPTP}(c; r, k)$ as a temporal context. This temporal contextual information is then combined with the spatial contextual information coming from spatial neighbors of $\mathbf{x}_k(r)$. The relation in eq. (2.14) is very important from the viewpoint of the actual application of this spatial-temporal contextual classification rule, since it allows a distribution of computational load over different times. In other words, due to the first order Markov property of temporal class labels, this classifier doesn't require one to process all the temporal data sets at one time. At any specific time, $H_{SPTP}(\bullet)$ for that time can be computed using only the current data set and the spatial-temporal classification result of the previous data set. Then, this result of $H_{SPTP}(\bullet)$ can be passed to the next step using eq. (2.14) when the next temporal data set is available. This allows the computational load to be distributed over different times. Spatial-temporal contextual classification with p temporal data sets can be obtained by applying $H_{SPTP}(\bullet; r, p)$ to each pixel in $\mathbf{X}(p)$.

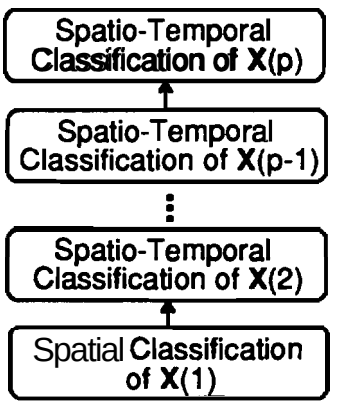


Figure 2.5 Spatial-Temporal Classification with $H_{SPTP}(\bullet)$.

The flowchart of spatial-temporal contextual classification is provided in Fig. 2.5. The result of spatial-temporal classification of the k^{th} temporal data set is fed into the classification process of the $(k+1)^{th}$ temporal data set as spatial-

temporal contextual information. Therefore the classification of a current temporal data set requires only the classification results of previous data set.

This spadial-temporal contextual classifier can be easily generalized to accommodate different spatial neighborhoods for each different temporal data set. This generalization may be quite useful when sensors with different spatial resolutions are used to acquire each temporal data set. In this report, for simplicity's sake, only first order spatial neighborhood system is considered for all the given rnultitemporal data sets.

CHAPTER 3

SPATIAL CONTEXTUAL CLASSIFICATION

3.1 Introduction

In this chapter, the problem of spatial contextual classification with $H_{SP}(\cdot)$ in eq. (2.11) is addressed. Several models and approaches which allow one to compute $H_{SP}(\cdot)$ will be discussed. Since only the spatial contextual classification is considered and the result in this chapter is applicable to any temporal data set $\mathbf{X}(k)$, k = 1, ---, p, the time index will be dropped for a notational simplicity. The spatial location parameter "(r)" will be also dropped whenever possible without causing confusion as in the previous chapter.

Spatial contextual classification can be carried out by applying $H_{SP}(\cdot)$ defined in eq. (2.11) to each pixel in the given data set. $H_{SP}(\cdot)$ can be computed as,

$$H_{SP}(c; r) = \frac{P\{\mathbf{X}_S'(r) = X_S' \mid \mathbf{c}(r) = c\}P\{\mathbf{c}(r) = c\}}{P\{\mathbf{X}_S'(r) = X_S'\}} \tag{3.1}$$

where, $$P\{X_S' \mid c\}P\{c\} = \sum_{C \in \Omega^*} P\{\mathbf{X}_S' = X_S' \mid \mathbf{c} = c, \mathbf{C}_S = C\}P\{\mathbf{c} = c, \mathbf{C}_S = C\}$$

Spatial classifiers rely on the fact that the statistical dependence between spectral responses of adjacent pixels, and subsequently the dependence between their class labels, can provide discriminating information in addition to spectral responses on which pixelwise classifiers depend. As discussed in Chapter 1, there are two different sources of spatial contextual information. One is the contextual information coming from spatial correlation between adjacent pixel feature vectors, and the other is the spatial class label dependency context between adjacent pixels. While the joint probability of class labels, $P\{c, C\}$ in eq.

(3.1), accounts for the spatial class label dependency context, the joint class-conditional probability $P\{X'_S \mid c, C\}$, manifests the spatial interpixel correlation contextual information between feature vectors in X'_S.

3.2 Spatial Interpixel Correlation Context

The interpixel correlation contextual information, in general, is a useful attribute to utilize in the classification and has been successfully used in several cases, for example, see (Yu and Fu 83), but, its inclusion generally requires extensive computation. For this reason, it is often assumed that the feature vectors in X'_S are class-conditionally independent. That is,

$$P\{X'_S \mid C'_S\} = \prod_{v \in N'_S} P\{x(r+v) \mid c(r+v)\} \qquad (3.2)$$

However, as might often be seen in real data, there does exist spatial correlation between adjacent feature vectors, and the spatial correlation coefficients generally vary over the spectral wavelengths and over the classes. It is also dependent on the direction of the spatial lag between pixels. The degree of spatial correlation is also closely related to the spatial resolution of the employed sensor. Spatial correlation coefficients which are class-unconditionally computed have generally higher values and a slower decreasing rate than the class-conditionally computed ones. R. Kettig and D. A. Landgrebe (Kettig and Landgrebe 76) used this fact in the ECHO classifier, which assumes independence of feature vectors in homogeneous; regions since the class-conditionally computed spatial correlation coefficient usually decreases; very quickly as the spatial distance between pixels increases.

Whether the independence assumption in eq. (3.2) is appropriate or not depends on the particular problem under consideration. There are various reasons for spatial correlation to exist between spectral measurements of spatially adjacent pixels. It can arise due to an inherent property of specific ground cover types being observed by the sensor. For example, the spacing of row crops, the plant size in an agricultural scene, or the relative vegetation and soil mixture, etc., could cause spatial variation in spectral responses. This is

generally referred to as the "texture," which can be described as a repeated variation in spectral responses over relatively small areas (Hoffer 78). This textural context would be able to provide valuable information, for example, in identifying forest cover against agricultural crops, but, unfortunately, this textural context may not be so conspicuous in some remotely sensed image data mainly due to a relatively low spatial resolution. Since this textural context is a local spatial characteristic belonging to each different scene cover type and therefore generally spatially variant, its utilization often involves an object extraction step.

Other than the spatial characteristic of scene cover types which causes the texture, there are also other sources such as the so called, "adjacent reflection," - the reflection of spectral energy of adjacent pixels to the sensor, the non-ideal spatial cut-off characteristic of sensor, or the spatial overlaps of pixel elements. Spatial correlation due to these effects seem to be not so directly related to specific cover types in the scene being observed as in textural contexts. Again, whether the spatial correlation should be considered as a property of each different class or not, is solely dependent on the problem at hand and a spatial characteristic of the selected data set. Even though the spatial correlation context may not be a distinguishable characteristics of the classes, its inclusion can help in improving classification performance by allowing more accurate class-conditional joint probability estimates as illustrated in following.

Assume a simple two class problem in one dimensional feature space as,

$$\text{Class } \omega_1 \sim N(m_1, \sigma_1^2) \text{ with prior probability } 0.5$$
$$\text{Class } \omega_2 \sim N(m_2, \sigma_2^2) \text{ with prior probability } 0.5$$

Data are to be classified using the spatial interpixel correlation context. To make the analysis simple, assume only one neighbor, denoted by $x(r+v)$ to $x(r)$. v indicates a spatial displacement of the neighbor $x(r+v)$ from the pixel $x(r)$. Data are spatially correlated as,

$$\text{Cov}\left[x(r),\ x(r+v) \mid \omega_i,\ \omega_j\right] = \sigma_i \sigma_j \rho_{ij}, \qquad 1 \le i, j \le 2$$

Assume that $\sigma_1 = \sigma_2 = a$, and $\rho_{ij} = \rho$, $-1 \le \rho \le 1$. Assuming $\rho_{ij} = \rho$ for all i, j combinations means that the spatial correlation coefficient is independent of classes. Inclusion of this interpixel spatial correlation context will allow more accurate estimate of joint probability of x(r) and x(r+v). An extended feature vector is defined as,

$$X_{ext} = \begin{bmatrix} x(r) \\ x(r+v) \end{bmatrix}$$

With this extended feature vector, X_{ext}, the pixel corresponding to the feature x(r) is to be classified not only using x(r) but also x(r+v). Suppose x(r+v) belongs to ω_j. If x(r) belongs to ω_k where k = 1, 2, then, X_{ext} is distributed as,

$$X_{ext} \sim MVN[M_k, \Sigma]$$

$$\text{where, } M_k = \begin{bmatrix} m_k \\ m_j \end{bmatrix} \text{ and } \Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

The decision rule based on minimum Bayes error with "0-1" loss function with X_{ext} is,

$$(M_2 - M_1)^T \Sigma^{-1} X_{ext} + \frac{M_1^T \Sigma^{-1} M_1 - M_2^T \Sigma^{-1} M_2}{2} \underset{\omega_1}{\overset{\omega_2}{\lessgtr}} 0$$

Suppose $m_1 = -m$, $m_2 = +m$, $m > 0$, then, after algebraic simplification, the decision rule is reduced to following linear classifier.

$$x(r) \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \rho\,[x(r+v) - m_j] \qquad (3.3)$$

This defines a linear decision boundary and its slope is determined by the spatial correlation coefficient p between x(r) and x(r+v).

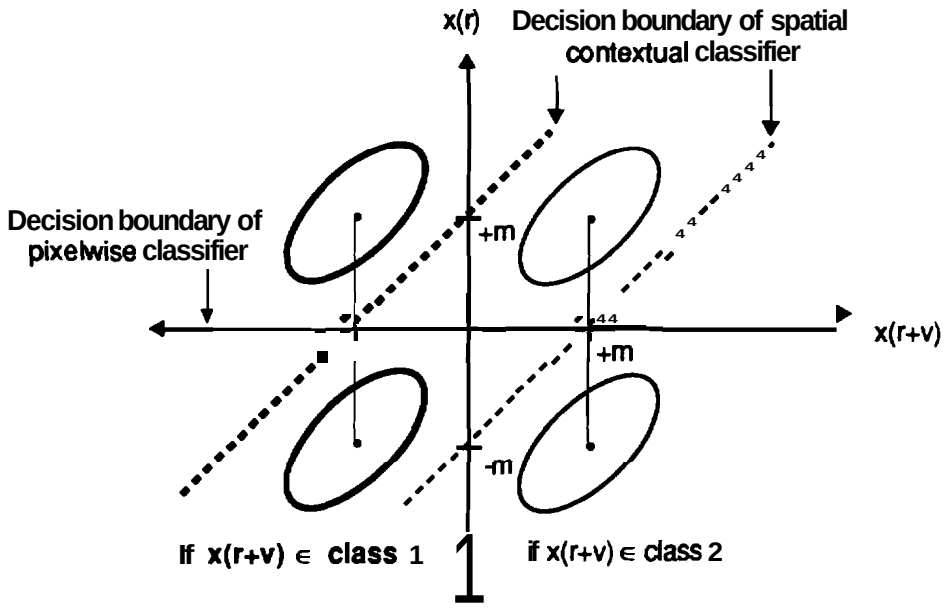


Figure 3.1 Decision Boundary of Spatial Classifier.

A decision rule corresponding to a pixelwise non-contextual classifier can be obtained from eq. (3.3) by setting $\rho = 0$, that is,

$$\text{if } x(r) < 0, \text{ classify } x(r) \text{ to } \omega_1$$
$$\text{if } x(r) > 0, \text{ classify } x(r) \text{ to } \omega_2$$

The decision boundary of the spatial classifier in eq. (3.3) is shown in Fig. 3.1 with that of pixelwise classifier without taking account of the interpixel spatial correlation context for comparison. With $\Phi(x)$, the cumulative distribution function for the standard normal density function defined as,

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\mu^2/2} \, d\mu$$

the Bayes errors corresponding to the spatial contextual classifier and the pixelwise classifier can be written respectively as,

$$E_{Bayes}(\rho) = 1 - \Phi\left(\frac{m}{\sigma\sqrt{1-\rho^2}}\right) \quad (3.4.a)$$

$$E_{Bayes}(\rho=0) = 1 - \Phi\left(\frac{m}{\sigma}\right) \quad (3.4.b)$$

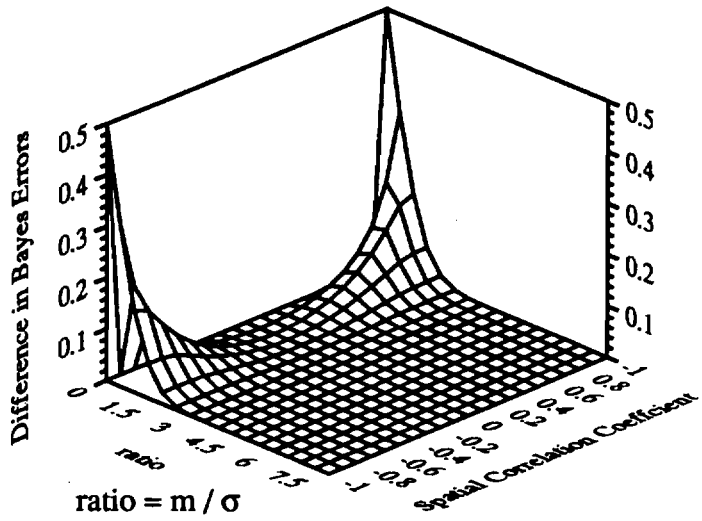


Figure 3.2 Difference in Bayes Errors with and without Spatial Correlation Comtext.

The differonce between these two Bayes errors, denoted as AE, is computed as,

$$\Delta E \equiv E_{Bayes}(\rho=0) - E_{Bayes}(\rho) = \Phi\left(\frac{m}{\sigma\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{m}{\sigma}\right) \quad (3.5)$$

Since AE iis always non-negative for $|\rho| \leq 1$ with minimum value zero at $\rho = 0$, as shown in Fig. 3.2, the classifier designed with spatial correlation context in consideration always reduces the Bayes error compared to the pixelwise classifier. However, the amount of reduction in Bayes error depends on the degree of spatial correlation and also on the separability between the two classes, which is represented by m/σ in this example. If the two classes are well separatedl, that is, if m is large relative to o, then, there are very small differences between the two Bayes errors in eq. (3.4.a,b). Therefore, there would not be so significant an improvement in classification accuracy by using the spatial interpixel correlation context. Note that the individual Bayes errors

are also very small in this case. However, if m/σ is not large enough, there can be significant differences between the two Bayes errors especially when $|\rho|$ is near one. Figure 3.3 shows the Bayes error differences when the ratio m/σ is increased from 0.2 to 1.8.

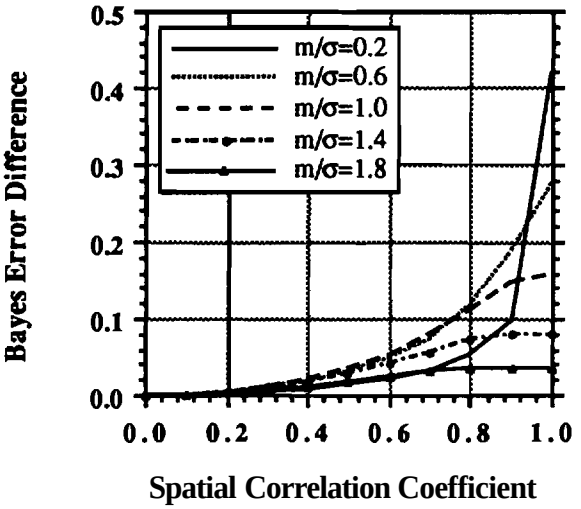


Figure 3.3 Samples of Bayes Errors Differences with and without Spatial Correlation Context.

The value on the vertical axis when the spatial correlation coefficient ρ is one, is the Bayes error in eq. (3.4.b) of the pixelwise classifier. When $m/\sigma = 0.2$, there is a significant Bayes error of about 0.4 for the pixelwise classifier. When $|\rho| \geq 0.8$, this Bayes error can be reduced by 0.05 ~ 0.4 by employing the spatial contextual classifier in eq. (3.3). As the ratio m/σ increases, the amount of possible Bayes error decrease obtainable by using the spatial correlation context becomes less significant. When the Bayes error of the pixelwise classifier is moderate, for example, about 0.15 for the case $m/\sigma = 1.0$, it can be reduced by 0.05 ~ 0.15 when $|\rho| \geq 0.6$ by using spatial interpixel correlation context as in eq. (3.3).

3.3 Modeling of Class-Conditional Joint Probability

According to the property of a jointly Gaussian distribution, non-zero spatial correlation between $x(r)$ and $X_S(r)$ means that they are not statistically independent of each other. Therefore, appropriate modeling of joint conditional

probability can improve classification performance by incorporating spatial correlation into the decision making process. This incorporation of spatial correlation into the classification rule might be expected to become more important as the spatial resolution becomes finer. Since the observations $x(r)$ and $X_S(r)$ are assumed to be jointly Gaussian, one straightforward approach is computing the conditional joint probability using the stacked vector or extended feature vector defined as,

$$X_{ext}(r) \equiv \begin{bmatrix} x(r) \\ X_S(r) \end{bmatrix} \tag{3.6}$$

This stacked vector approach requires estimates of the mean and covariance matrix of the extended feature vector, which requires increased numbers of training samples due to the increased dimensionality of the feature vector. Also the concadenation of feature vectors makes it necessary to define more spectral sub-classes. In most remote sensing applications, it would be very hard to obtain a large enough number of training samples, and this stacked vector approach may be inappropriate in many cases due to this increased dimensionality.

Instead of estimating directly the covariance matrix using feature vectors, model-based approaches can be taken to loosen the requirement of additional training samples by defining and estimating a few parameters which can adequately model the spatial correlation structure. Proper choice of a flexible model which can adequately fit various multispectral images in a given remote sensing application will be very important. One available model is the autocorrellation model proposed by Hjort *etal.* (Hjort *etal. 85).* It is based on the assumption that an observed feature vector, $x(r)$, $r \in L$, is a sum of two independent processes, one being a class dependent spatially independent process and the other being a spatially correlated noise process, *i.e*, for $r \in L$,

$$x(r) \equiv y(r) + \varepsilon(r) \tag{3.7}$$

If $x(r)$ is q dimensionally multivariate Gaussian, that is, $MVN[M(r), \Sigma_0]$, then, $y(r)$ is assumed to be a spatially independent Gaussian process with $MVN[M(r),$ (1-

$\theta)\Sigma_0]$. $M(r)$ denotes a mean vector of the class to which $x(r)$ belongs. Σ_0 is a common covariance matrix of all classes. The noise, $\varepsilon(r)$, is a multivariate Gaussian process with $MVN[0_q, \theta\Sigma_0]$, (0_q is a q by *1* matrix with all zeros), but it is assumed to be spatially correlated as,

$$Cov[\varepsilon(r), \varepsilon(r+v)] = \rho_S^{|v|}\theta\Sigma_0 \qquad (3.8)$$

The $y(r)$'s are considered as bearing information directly about the pixel class label, whereas the noise process, $\varepsilon(r)$'s are assumed to be due to measurement errors and possibly other sources of "extra variations" (Yu and Fu 83) and consequently class-independent. From the relation in eq. (3.7 and 3.8), the covariance matrix between $x(r)$ and $x(r+v)$ is computed as,

$$Cov[x(r), x(r+v)] = Cov[y(r), y(r+v)] + Cov[\varepsilon(r), \varepsilon(r+v)]$$

$$= \rho_S^{|v|}\theta\Sigma_0, \text{ where } v \neq 0_q \qquad (3.9)$$

Spatial correlation parameter ρ_S and common covariance matrix Σ_0 are estimated in the training stage (Hjort *et al.* 85). Using the relation in eq. (3.9), the covariance matrix of $\{x(r), X_S(r)\}$ can be computed as,

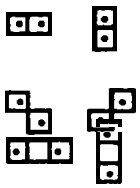
$$\text{Covariance matrix of } \{x(r), X_S(r)\} = \Sigma_0 \otimes \begin{bmatrix} 1 & a & a & a & a \\ \alpha & 1 & \beta & \gamma & \beta \\ \alpha & p & 1 & p & \gamma \\ \alpha & \gamma & \beta & 1 & \beta \\ \alpha & \beta & \gamma & \beta & 1 \end{bmatrix} \qquad (3.10)$$

where, @ is the Kronecker Product

$a = \rho$ Correlation of first order neighbors

$\beta = \rho_S^{\sqrt{2}}\theta$ Correlation of diagonal neighbors

$\gamma = \rho_S^2\theta$ Correlation of second order neighbors



A few comments are deserved here about this model. One of its limitations lies in the fact that it cannot adopt non-identical spatial correlation structure over different spectral wavelengths. It is conceivable to have a different degree of spatial correlation over different spectral wavelengths especially when the spatial resolution is dissimilar for different bands. In this case, this model cannot be easily generalized to the case of non-identical spatial resolution over different wavelengths such as in thermal band of Landsat Thematic: Mapper data which has 120m resolution compared to the others of 30m. Probably the most important notice about this model may be its assumption of the same covariance matrix for all the classes. Note that the second order statistical characteristics which are generally represented by the covariance matrix provide crucial and indispensable information in classification. This limitation occurs since $x(r)$ is decomposed into two different processes and the spatial correlation of the noise process, which is class-independent, is assumed to be directly related to the covariance of $x(r)$.

Before further considering models for the spatial correlation structure, it will be worthwhile to scrutinize a remote sensing system model, especially the scene model, to have a better understanding of spatial correlation. According to the taxonomies of (Kerkes and Landgrebe 89), a remote sensing systenn can be described as a cascade of three components, namely, a scene model, a sensor model and a processing model. The scene model describes the mechanism that input!; spectral radiance to a sensor, and is affected by all spectral and spatial sources and variations of the scene. The sensor model explains the effect of transforming the incident spectral radiance into a both spatially and spectrally sampled discrete image. The processing model accounts for the processing applied to the remotely sensed image data. If the sensor is assumed not to make significant changes in the reflectance values coming from the scene, then, the pixels will vary similarly to the reflectance of the scene in both a spatial and spectral sense. According to the scene model and with this assumption, the formation of multispectral image data can be modeled in the following two steps.

Step 1 : Generation of a spatially correlated but spectrally uncorrelated zero mean signal.

Step 2 : Transformation of this signal to have the appropriate class mean and covariance matrix.

The Markov random field (MRF) model is a good candidate for describing the first step. The second step is the inverse of the so called, the whitening process (Fukunaga 90) or a decorrelation process. The Markov random field model has been well-suited for many problems in statistical image processing, such as restoration and segmentation. It has been also very useful to characterize given spatially correlated or textured images with a few parameters. Therefore in this report, the Markov random field will be used to model the spatial correlation structure. Although many varieties of this model are available (Besag 74, Kashyap 81, Derin and Kelly 89, Derin and Elliot 87), only the conditional Markov (CM) model (Kashyap 81) is considered. This conditional Markov model is used to estimate spatial correlation between neighboring pixels using its parameters which can best fit the given multispectral image data.

Applying the random field model requires the image to be stationary. Stationarity is defined as follows. Feature vectors $x(r)$'s are called covariance stationary if the covariance matrix of $\{x(r), x(r+v)\}$ is dependent only on $|v|$. If $x(r)$ is covariance stationary and additionally satisfies $E[x(r)] = M$ for all r, then, it is called weakly stationary. Note that, in most of images in remote sensing applications, the mean and covariance matrix of each pixel is generally different at each location with respect to its corresponding class. To normalize this effect of class statistics, the normalized feature vector, $y(r)$, is defined as,

$$y(r) = W(r) [x(r) - M(r)] \qquad (3.11.a)$$

$M(r)$ is the mean of the class to which $x(s)$ belongs and $\Sigma(r)$ is the covariance matrix of the class of $x(r)$. The whitening matrix $W(r)$ in eq. (3.11.a) which decorrelates the interband correlation is computed as,

$$W(r) = \mu^{-\frac{1}{2}} \Psi^T \qquad (3.11.b)$$

where, $\Sigma(r)\Psi = \Psi\mu$

Ψ is the eigenvector matrix of $\Sigma(r)$ and μ is the corresponding eigenvalue matrix which has eigenvalues, λ_1, ---, λ_q, at its diagonal. Since $x(r)$ is assumed to follow a multivariate normal distribution with $M(r)$ and $\Sigma(r)$, $y(r)$ has also a multivariate normal distribution as,

$$Y(r) \sim MVN[0_q, I_{qxq}] \tag{3.12}$$

where I_{qxq} is a q by q identity matrix. These normalized feature vectors can be considered as the spatially correlated but spectrally uncorrelated zero mean signal in the step 1. There can be two modes of stationarity. If spatial correlation context is different for each class, modeling with the Markov random field can be performecl for each class separately. This is called "locally" stationary since the stationarity holds for only that class. If the spatial correlation is assumed to be the same for all classes, then, the modeling with the Markov random field is performed over the whole image, and it is called "globally" stationary.

The normalized feature vector $y(r)$'s, $r \in L$, are assumed to be (globally) stationary and follow the conditional Markov(CM) model. Although the following derivation is based on the globally stationary case, the result can be easily modified to the "locally" stationary case. Since there is no interband correlation in $y(r)$, each band is assumed to follow the conditional Markov(CM) model separately with generally different parameters. According to the model, $y(r)$ satisfies,

$$y(r) = \sum_{v \in N_S} \theta_v \, y(r+v) + \Lambda e(r) \tag{3.13}$$

$$\text{where, } \theta_v = \begin{bmatrix} \theta_{v,1} & & 0 \\ & \ddots & \\ 0 & & \theta_{v,q} \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_q} \end{bmatrix}$$

N_S is the spatial neighborhood defining set. Even though any order neighbor system is possible, for simplicity, only the first order neighbor system, $N_S = \{(f1, 0), (0, \pm 1)\}$ is considered. θ_v and A are diagonal matrices. According to the CM

model, θ_v is symmetric, that is, $\theta_v = \theta_{-v}$, and stationary noise field $e(r)$ is distributed as $MVN[0_q, I_{qxq}]$ with following properties.

$$E\{e(r)\, e^T(r+v)\} = \begin{cases} -\theta_v, & \text{if } v \in N_S \\ I_{qxq}, & \text{if } v = (0.0) \\ 0, & \text{otherwise} \end{cases} \qquad (3.14.a)$$

$$Pr\{e(r) \mid \text{all } y(v)\text{'s}, v \neq r\} = Pr\{e(r) \mid y(r+v), v \in N_S\} \qquad (3.14.b)$$

$$E\{e(r)\, y(v)\} = 0, \quad r \neq v \qquad (3.14.c)$$

Unknown parameter matrices θ_v and A are estimated using training samples. Since no interband correlation in $y(r)$ is assumed, the unknown parameters $\theta_{v,i}$'s and λ_i's are estimated separately for each band i, $i = 1, ---, 9$.

There are three different methods of estimating $\theta_{v,i}$ and λ_i, maximum likelihood estimation (MLE), the coding method and the least squared error (LS) method. Although maximum likelihood estimation can give estimates with desirable properties, like asymptotic consistency and efficiency, it is computationally very complex due to a difficulty in deriving an explicit log-likelihood function expression because of an evaluation of the Jacobian of the transforming matrix. Although the coding method (**Besag** 74) succeeds in avoiding this complex calculation by dividing the pixels into disjoint subsets and estimating unknown parameters over each subset, one of its drawback, especially significant in remote sensing application is its low efficiency in data utilization since it can use the data only partially in estimating unknown parameters. A least squared error (LS) approach is computationally simple, asymptotically consistent and also efficient in the utilization of the training data (Chellappa 81). 'Therefore in this report, the least squared error (LS) approach is taken. For each band i, $i = 1, ---, 9$, the i^{th} component (band) of $y(r)$ is written as,

$$y_i(r) = \sum_{v \in N_S} \theta_{v,i}\, y_i(r+v) + \sqrt{\lambda_i}\, e_i(r) \qquad (3.15)$$

Note that θ_v is symmetric, therefore, $\theta_{(1,0),i} = \theta_{(-1,0),i}$ and $\theta_{(0,1),i} = \theta_{(0,-1),i}$. Denote $\Theta_{v,i}$ as a matrix of unknowns as,

$$\Theta_{v,i} \equiv \begin{bmatrix} \theta_{(1,0),i} \\ \theta_{(0,1),i} \end{bmatrix} \tag{3.16.a}$$

and, $\zeta_i(r)$ as,

$$\zeta_i(r) \equiv \begin{bmatrix} y_i(r + (1,0)) + y_i(r - (1,0)) \\ y_i(r + (0,1)) + y_i(r - (0,1)) \end{bmatrix} \tag{3.16.b}$$

then, the estimate based on the least square approach is obtained as,

$$\hat{\Theta}_{v,i} = \left[\sum \zeta_i(r)\, \zeta_i^T(r)\right]^{-1} \left[\sum \zeta_i(r)\, y_i(r)\right] \tag{3.17}$$

The summation in eq. (3.17) is performed over all training samples. If isotropy is assumed for the spatial correlation, that is, if spatial correlation is assurned to be independent of the direction of the spatial lag between pixels, then., $\theta_{(1,0),i} = \theta_{(1,0),i} = \theta_{(0,1),i} = \theta_{(0,-1),i}$. Therefore, it is sufficient to estimate only one parameter $\theta_{(1,0),i}$ for each band by using eq. (3.17) with,

$$\Theta_{v,i} \equiv \left[\theta_{(1,0),i}\right] \tag{3.18.a}$$

$$\zeta_i(r) \quad [\, y_i(r + (10\ ) + y_i(r - (1,0)) + y_i(r + (0,1)) + \Upsilon \quad (0,1))\, \rfloor \tag{3.18.b}$$

Using the properties given in eq. (3.14.b,c) and the estimated parameters $\theta_{v,i}$'s, spectral density function of $y_i(r)$'s can be derived as,

$$S_{y_i}(u_1, u_2) = \frac{\lambda_j}{1 - 2\left\{\theta_{(1,0),i} \cos u_1 + \theta_{(0,1),i} \cos u_2\right\}} \tag{3.19}$$

The covariance of $\{y_i(r), Y_i(r+v)\}$ is then obtained by inverse Fourier transforming the spectral density function in eq. (3.19) as in,

$$\text{Cov}\{y_i(r), y_i(r+v)\} = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} S_{y_i}(u_1, u_2)e^{\sqrt{-1}\,u\cdot v}\,du_1\,du_2 \qquad (3.20)$$

where, $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$, $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ and $u \cdot v = u_1v_1 + u_2v_2$

Using eq. (3.20), the covariance matrix of $\{y(r), Y_S(r)\}$, which is denoted as Σ_Y, can be computed. For each band i, $i = 1, \cdots, q$, define the following covariances which comprise the 5q by 5q symmetric joint covariance matrix of Σ_Y.

$$\begin{aligned}
\alpha_{1,i} &= \text{Cov}\,[\,y_i(r), y_i(r+(1,0))\,] \\
\alpha_{2,i} &= \text{Cov}\,[\,y_i(r), y_i(r+(0,1))\,] \\
\beta_{1,i} &= \text{Cov}\,[\,y_i(r), y_i(r+(1,1))\,] \\
\gamma_{1,i} &= \text{Cov}\,[\,y_i(r), y_i(r+(2,0))\,] \\
\gamma_{2,i} &= \text{Cov}\,[\,y_i(r), y_i(r+(0,2))\,]
\end{aligned} \qquad (3.21)$$

Using these components, the covariance matrix Σ_Y is written as,

$$S_Y = \begin{bmatrix} I_{qxq} & A_1 & A_2 & A_1 & A_2 \\ A_1 & I_{qxq} & B_1 & C_1 & B_1 \\ A_2 & B_1 & I_{qxq} & B_1 & C_2 \\ A_1 & C_1 & B_1 & I_{qxq} & B_1 \\ A_2 & B_1 & C_2 & B_1 & I_{qxq} \end{bmatrix} \qquad (3.22)$$

where, for k = 1, 2,

$$A_k = \begin{bmatrix} \alpha_{k,1} & & 0 \\ & \ddots & \\ 0 & & \alpha_{k,q} \end{bmatrix} \qquad (3.23.a)$$

$$C_k = \begin{bmatrix} \gamma_{k,1} & & 0 \\ & \ddots & \\ 0 & & \gamma_{k,q} \end{bmatrix} \qquad (3.23.b)$$

and,

$$B_1 = \begin{bmatrix} \beta_{1,1} & & 0 \\ & \ddots & \\ 0 & & \beta_{1,q} \end{bmatrix} \qquad (3.23.c)$$

In an isotropic case, $A_1 = A_2$ and $C_1 = C_2$. If the spatial correlation is independent of wavelength, then with $\alpha_k = \alpha_{k,1} = \cdots = \alpha_{k,q}$, $\beta_1 = \beta_{1,1} = \cdots = \beta_{1,q}$ and $\gamma_k = \gamma_{k,1} = \cdots = \gamma_{k,q}$, the matrices, A_k, C_k, and B_1 can be further simplified as,

$$\begin{aligned} A_k &= \alpha_k \, I_{qxq} \\ B_k &= \beta_k \, I_{qxq} \\ C_k &= \gamma_k \, I_{qxq} \end{aligned} \qquad (3.24)$$

Since $y(r)$ is obtained from $x(r)$ by performing the linear transformation of eq. (3.11.a), the joint covariance matrix of $\{x(r), X_S(r)\}$ given their classes $\{c(r), C_S(r)\}$ can be computed by using the transformation matrix $W_{ext}(r)$ as,

$$\Sigma_{ext}(r) \equiv Cov\{x(r), X_S(r) \mid c(r), C_S(r)\} = W_{ext}(r) \, \Sigma_Y \, W_{ext}^T(r) \qquad (3.25)$$

where, $W_{ext}(r) \equiv \begin{bmatrix} W(r) & 0 & 0 & 0 & 0 \\ 0 & W(r+(0,1)) & 0 & 0 & 0 \\ 0 & 0 & W(r-(1,0)) & 0 & 0 \\ 0 & 0 & 0 & W(r-(0,1)) & 0 \\ 0 & 0 & 0 & 0 & W(r+(1,0)) \end{bmatrix}$

Notice that the joint covariance matrix in the form of eq. (3.22) and consequently the covariance matrix in the form of eq. (3.25) is not limited only to the Markov random field model but, in fact, is quite general. For example, the joint covariance matrix in eq. (3.10) which is derived under the autocorrelation model in eq. (3.7) can be written in the form of eq. (3.25) with appropriate values of α_k, β_1 and γ_k, and assuming the covariance matrices are the same for all classes. More generally, the form of eq. (3.22) and eq. (3.25) can be assumed to be valid, and the constituent unknown parameters can be directly estimated From the available training samples without explicit modeling of the given image with such models as the conditional Markov model, or the autocorrelation model.

Since $\{x(r), X_S(r)\}$ given their classes is assumed to be multivariate Gaussian, its joint class-conditional probability is computed by using $M_{ext}(r)$ defined as,

$$\mathbf{M}_{ext}(r) \equiv \begin{bmatrix} M(r) \\ M(r+(0,1)) \\ M(r-(1,0)) \\ M(r-(0,1)) \\ M(r+(1,0)) \end{bmatrix} \tag{3.26}$$

and the covariance matrix $\Sigma_{ext}(r)$ in eq. (3.25). Classification is then, performed by finding a class, $c \in \Omega$ which maximizes,

$$H_{SP}(c; r) = P\{\mathbf{c}(r) = c \mid \mathbf{x}(r) = x(r), \mathbf{X}_S(r) = X_S(r)\} \tag{3.27}$$

Note that evaluation of $H_{SP}(\cdot)$ requires summations over all possible combinations of $C \in \mathbf{R}^4$ as shown in eq. (3.1). The number of these class combinations would be very large since it grows exponentially with respect to the number of classes. This can be avoided by taking a recursive scheme as a sub-optimal approach, instead of its direct maximization over all combinations in one pass. Under the recursive scheme, $H_{SP}(\cdot)$ reduces to the following equation, which needs only the knowledge of the class identities of spatial neighbors.

$$H_{SP}(c; r \mid C) \equiv P\{\mathbf{c}(r) = c \mid \mathbf{x}(r) = x(r), \mathbf{X}_S(r) = X_S(r), \mathbf{C}_S(r) = C\}$$

$$= \frac{P\{\mathbf{x}(r) = x(r), \mathbf{X}_S(r) = X_S(r) \mid \mathbf{c}(r) = c, \mathbf{C}_S(r) = C\} P\{\mathbf{c}(r) = c \mid \mathbf{C}_S(r) = C\}}{P\{\mathbf{x}(r) = x(r), \mathbf{X}_S(r) = X_S(r) \mid \mathbf{C}_S(r) = C\}} \tag{3.28}$$

The denominator of eq. (3.28) does not depends on the class c and it need not be evaluated. Since the class identities of spatial neighbors are not available, intermediate classification results are used instead as estimates. This process is recursively applied to the pixels over all x-sites and •-sites in Fig. 3.4 at each recursion until negligible changes of class assignments are attained.

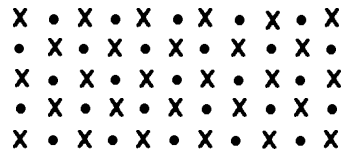


Figure 3.4 x-Sites and .-Sites of First Order Spatial Neighborhood System.

This recursive approach precludes not only the necessity of considering all the combinations of classes but also the need of evaluating exponential function to obtain probabilities from log-likelihood values.

In many multispectral images, however, especially in such scenes of agricultural areas, there are many homogeneous fields which are relatively large compared to the pixel size. For those pixels in homogeneous regions, it will be unnecessary to check all the possible combinations of $\{C_S(r) = C\}$. Therefore, if M classes are present, it will be sufficient to check only those M cases assuming all $X_S(r)$ have the same class as $x(r)$. This will save computation time significantly. Furthermore, pixels $\{x(r), X_S(r)\}$ are all classified simultaneously to one of the M classes. This simultaneous classification of all pixels in $\{x(r), X_S(r)\}$ will remove any isolated errors in the classification map which may be present otherwise. To avoid any blurring of the classification map near field boundaries, a careful choice of homogeneity test would be very important,

There are many measures of homogeneity of a set of pixels. The log-likelihood value of the conditional joint probability is one of the possible homogeneity measures. The ECHO classifier (Kettig and Landgrebe 76), for example, used this log-likelihood value to check the homogeneity of cells.

$$\text{Log}[\, P\{x(r), X_S(r) \mid c(r) \text{ and } C_S(r) \text{ are of same class})] > \text{THD} \qquad (3.29)$$

If the log-likelihood value is greater than a certain pre-defined threshold value, denoted by THD, then the pixels $\{x(r), X_S(r)\}$ are considered as homogeneous. In this case, instead of checking all combinations of $\{C_S(r) = C\}$, pixels $\{x(r), X_S(r)\}$ are classified simultaneously to one of the M classes. If the pixels $\{x(r), X_S(r)\}$ do not pass the homogeneity test of eq. (3.29), then those pixels are not

homogeneous. Therefore those pixels are subject to either the usual recursive operation or simple pixel-wise classification.

3.4 Modeling of Prior Probability

In eq. (3.1), spatial interpixel class dependency context is conveyed by the conditional prior probability $P\{\mathbf{c}(\mathbf{r}) = c \mid \mathbf{C}_S(\mathbf{r}) = C\}$. This conditional prior probability can impose a constraint reflecting the global property on each class assignment of the pixels and can produce a more homogeneous classification map. On the other hand, the class-conditional joint probability of the feature vectors $P\{\mathbf{x}(\mathbf{r}), \mathbf{X}_S(\mathbf{r}) \mid \mathbf{c}(\mathbf{r}), \mathbf{C}(\mathbf{r})\}$ may only reflect the local characteristics of the pixels in terms of appropriate distance measures, such as the Euclidean or the Mahalanobis distance measures. In many applications of remotely sensed data, the pixel size is much less than that of the real object or field on the ground.; therefore, classification results consisting of homogeneous regions with less isolated spots would be expected. This inherent tendency of coherent class labels of spatially adjacent pixels can be accounted for in classification by using conditional prior probabilities which can impose some constraints on the configurations of the class labels of spatially adjacent pixels.

The most straightforward way of obtaining conditional prior probabilities is to estimate 'the probabilities directly from training samples, or the class map after each iteration in case of a recursive approach. Although simple in concept and computation, this has several drawbacks in practice. First of all, there may not be enough training samples, in many real situations, to adequately estimate the prior probabilities of so many different configurations of classes. Furthermore, it will be very hard, and in some times, almost impossible, to gather a representative set of training samples containing all the meaningful configurations of classes in adequate amounts. Estimation using the intermediate result of classification is also another possibility, but this method is known to result in biased estimates (Tilton et *al.* 82, Dattatreya 91). Instead of estimating the prior probability itself, transition probabilities can be estimated under a simplifying assumption such as,

$$P\{\mathbf{c}(\mathbf{r}+\mathbf{v}) \mid \mathbf{c}(\mathbf{r}), \mathbf{c}(\mathbf{r}+\mathbf{t}), \mathbf{v} \text{ and } \mathbf{t} \in N_S, \mathbf{t} \neq \mathbf{v}) = P\{\mathbf{c}(\mathbf{r}+\mathbf{v}) \mid \mathbf{c}(\mathbf{r})\} \qquad (3.30)$$

This assumption implies that the probability of a class of a certain neighboring pixel of x(r) given all the other neighboring classes and the class of x(r) depends only on the class of the center pixel x(r). With this assumption, the conditional joint prior probability is simplified as,

$$P\{\mathbf{c}(r), \mathbf{C}_S(r)\} = P\{\mathbf{c}(r)\} \prod_{v \in N_S} P\{\mathbf{c}(r+v) \mid \mathbf{c}(r)\} \qquad (3.31)$$

If enough prior information is available, all the transition probabilities can be set up using prior information in advance. Otherwise, the transition probabilities can be estimated and iteratively updated after each iteration using the class map in a recursive way.

It is also possible to assume a certain form of distribution function for the joint prior probabilities and deduce conditional prior probabilities from the distribution function. This is analogous to assuming a multivariate normal distribution for feature vectors to compute the class-conditional joint probability of the feature vectors. The justification for assuming a certain form of distribution function for the joint prior probability might not be easily attainable. Nevertheless, among other possibilities, (for example, the geometric probability model (Owen 84), the Gibbs random field model (GRF) (Derin and Kelly 87, Besag 86, Derin and Elliot 87) is taken as a model of priors in this report, since it can exploit the inherent tendency of coherent class labels of spatially adjacent pixels in a very straightforward and efficient way. Also the Markov property of Gibbs random field model allows the conditional prior probability $P\{\mathbf{c}(r)|\mathbf{C}_S(r)\}$'s relatively in a very simple form.

The class labels $\mathbf{c}(r)$'s, $r \in L$, are assumed to be modeled by the Gibbs random field, then the conditional prior probabilities are given as, for $c \in \Omega$ and $C \in R^4$,

$$P\{\mathbf{c}(r) = c \mid \mathbf{C}_S(r) = C\} = \frac{1}{Z} e^{-U(c, C)} \qquad (3.32)$$

where, $U(c, C) = \sum_{W \in N_\alpha} V_W(c, C)$, $\quad Z = \sum_{c \in W} e^{-U(c, C)}$

$V_W(c, C)$ is a potential function of a class configuration $\{c(r) = c, C_S(r) = C\}$ on a clique $w \in N_{CL}$, and $U(c, C)$ is an energy function. A clique is a set of sites (including single sites) such that any two elements in the set are neighbors of each other. Types of cliques of first order spatial neighborhood system are shown in Fig. 3.5.

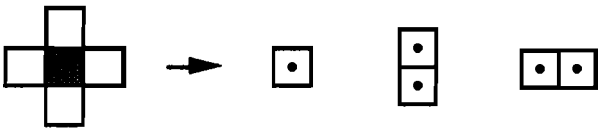


Figure 3.5 Clique Types of First Order Spatial Neighborhood System.

N_{CL} in eq. (3.32) denotes a set of cliques consisting of only the sites of pixels in $\{x(r), X_S(r)\}$. Z is a normalizing factor and called as partition function. Since this is not dependent on the particular realization of $\{c(r) = c, C_S(r) = C\}$, it needs not be evaluated. It is very important to have a proper potential function to be able to exploit the class label coherence in a classification. The clique potential function $V_W(c, C)$ is defined as $V_W(c, C) = a_C$, $c \in \Omega$ in the case of type 0. For type i, $i = 1, 2$, $V_W(c, C)$ is defined as,

$$V_W(c, C) \equiv \begin{cases} -b_i, & \text{if all classes of pixels in the clique w are the same} \\ +b_i, & \text{otherwise} \end{cases}$$

While $\{a_C \mid c \in R\}$ determines the relative likelihood of each class c, $c \in R$, $\{b_1, b_2\}$ determines the emphasis of interactions between classes of adjacent pixels. As the b's become large, more homogeneous regions will be favored in the classification. If the b's are set to zero, this model of priors will be converted back to a classification with no interaction between class labels of adjacent pixels. Therefore, it is possible to control the preferred sizes and directions of homogeneous regions with appropriate values of $\{b_1, b_2\}$. Assuming $a_C = a$, for all c in Ω, *i.e.*, equal prior probabilities of the classes and $b = b_1 = b_2$, then the class-conditional probability is given as,

$$P\{c(r) = c \mid C_S(r) = C\} \propto e^{2b(m-2)} \tag{3.33}$$

where, m_i is the number of occurrences of the class c in C, the classes of neighbors. This simplifies the log likelihood value computation of eq. (3.28) as in,

$$Log\,[H_{SP}(c;\, r \mid \mathbf{C}_S(r) = C)]$$

$$= Log\,[P\{x(r), X_S(r) \mid c(r) = c, \mathbf{C}_S(r) = C\}] + 2b(m-2) + const. \qquad (3.34)$$

where "const." doesn't depend on the particular class assignment c to the pixel x(r). If the pixels $\{x(r), X_S(r)\}$ are assumed to be independent of each other given their classes, then, eq. (3.34) is modified to much simpler form,

$$Log\,[H_{SP}(c;\, r \mid \mathbf{C}_S(r) = C)] = Log\,[P\{x(r) \mid c(r) = c\}] + 2b(m-2) + const. \qquad (3.35)$$

3.5 Experiments of Spatial Contextual Classification

3.5.1 Description of Experiments

To test the spatial contextual classification rule $H_{SP}(\cdot)$ proposed in this chapter, two Landsat Thematic Mapper (TM) data sets, which were acquired over the same agricultural areas in Tippecanoe County, Indiana during July and September, 1986, were used in an experiment.

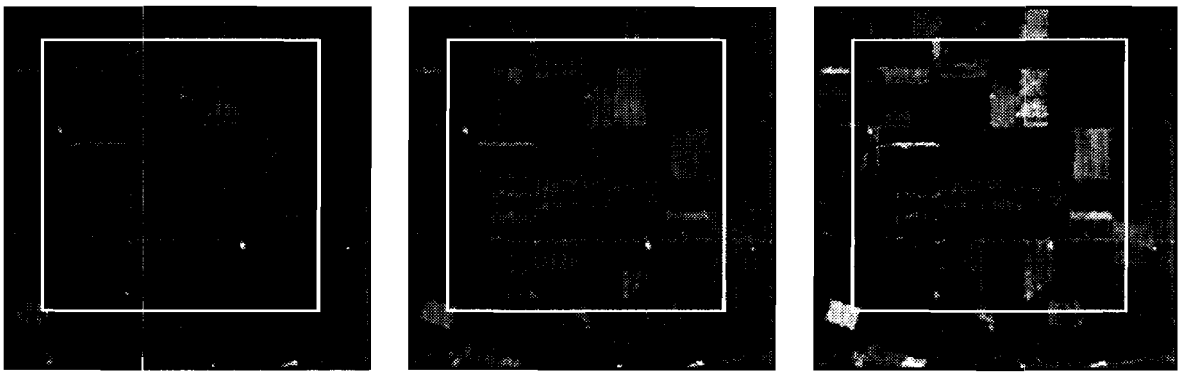
All 7 bands were used in the classification. Four information classes were determined from the available ground truth data and several spectral sub-classes were defined for each information class separately for each diata set to satisfy the multivariate normal assumption. Training and test samples were chosen as in Table 3.1.

Table 3.1 Training and Test Samples.

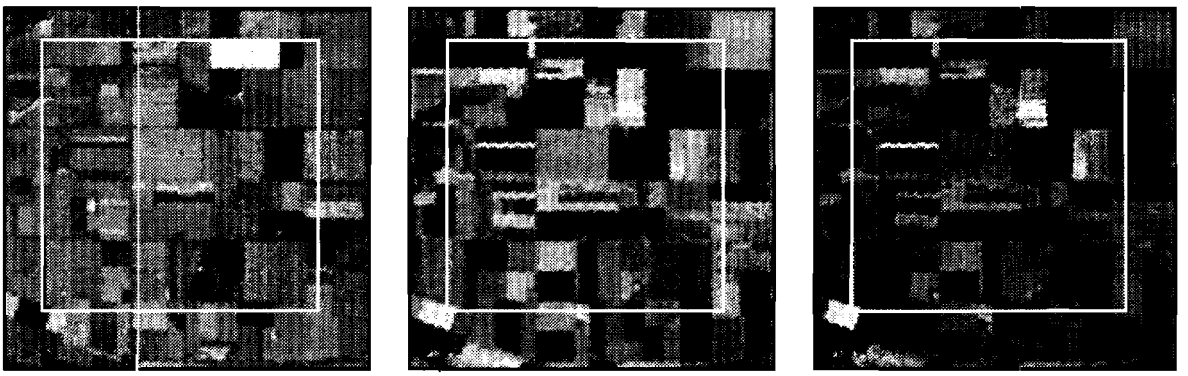
Information Class	July Data			September Data		
	Sub-classes	number of Training	Test	Sub-classes	number of Training	Test
Corn	1	286	5559	2	376	5559
Soybeans	3	495	4773	1	408	4773
Wheat	2	208	1215	2	330	1215
Alfalfa/Oat	3	321	1366	2	219	1366
Total	9	1310	12998	7	1333	12998

A portion of image, 128 by 128 pixels, is chosen as a test data set. Since there are pixels with unknown identities in the test portion of the data, only 12998 pixels which have known ground truth labels, were counted when assessing classification performance. Classification performance is computed in terms of ***overall classification accuracy*** **(OVA)** and ***class averaged classification accuracy*** **(CAG)** and compared with that of a pixelwise maximum likelihood classifier.

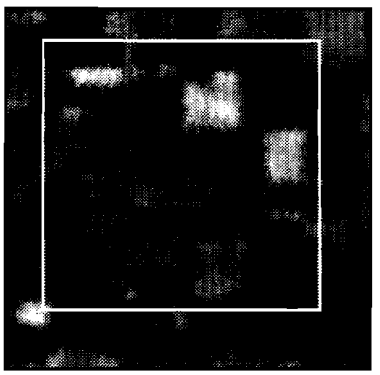
To see the effectiveness of spatial contextual information, three different experiments were carried out. The first experiment was performed only with the spatial interpixel correlation context. Secondly, only the spatial class label dependency context was used and in the last experiment, both were used simultaneously in the classification. Figure **3.6** and **3.7** show the July and September data sets and Fig. **3.8** is the associated ground truth map.



Band 1 (0.45 - 0.52 µm) Band 2 (0.52 - 0.60 pm) Band 3 (0.63 - 0.69 pm)

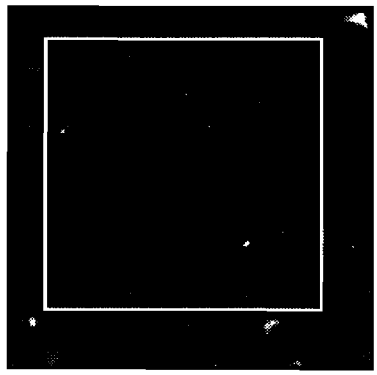


Band 4 (0.76 - 0.90 pm) Band 5 (1.55 - 1.75 pm) Band 6 (2.08 - 2.35 µm)

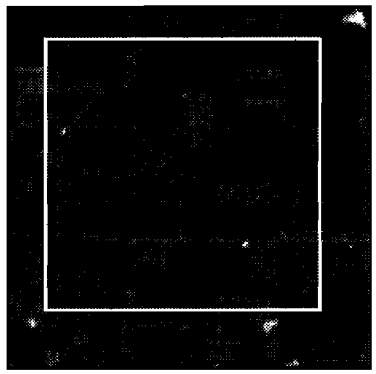


Band 7 (10.4 - 12.5 pm)

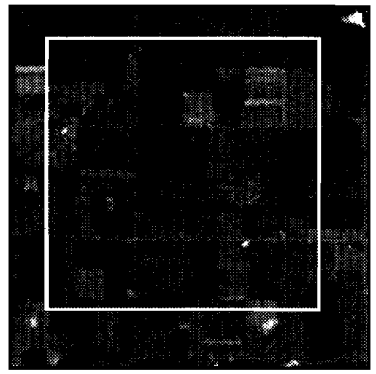
Figure 3.6 July Thematic Mapper (TM) Data Set.; The white box shows the 128 by 128 pixel portion of selected test fiekd.



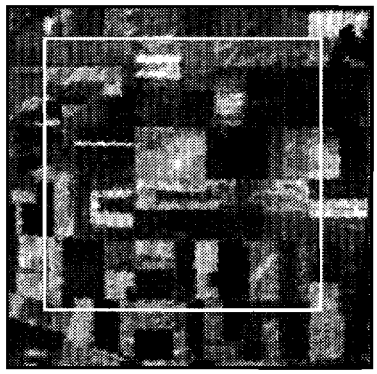
Band 1 (0.45 - 0.52 pm)



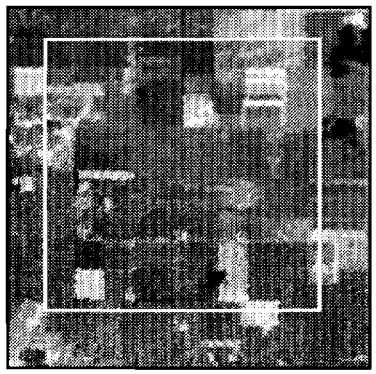
Band 2 (0.52 - 0.60 pm)



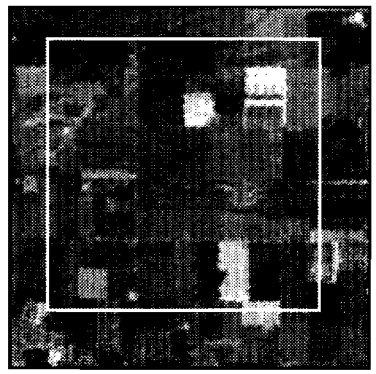
Band 3 (0.63 - 0.69 pm)



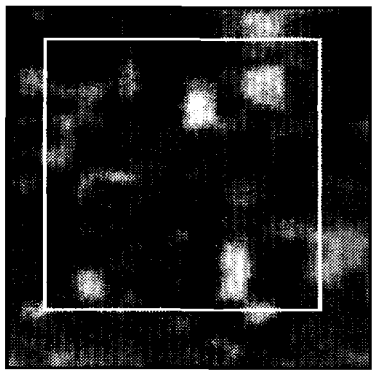
Band 4 (0.76 - 0.90 pm)



Band 5 (1.55 - 1.75 pm)



Band 6 (2.08 - 2.35 pm)



Band 7 (10.4 - 12.5 pm)

Figure 3.7 September Thematic Mapper (TM) Data Set.; The whiie box shows the 128 by 128 portion of selected test field.

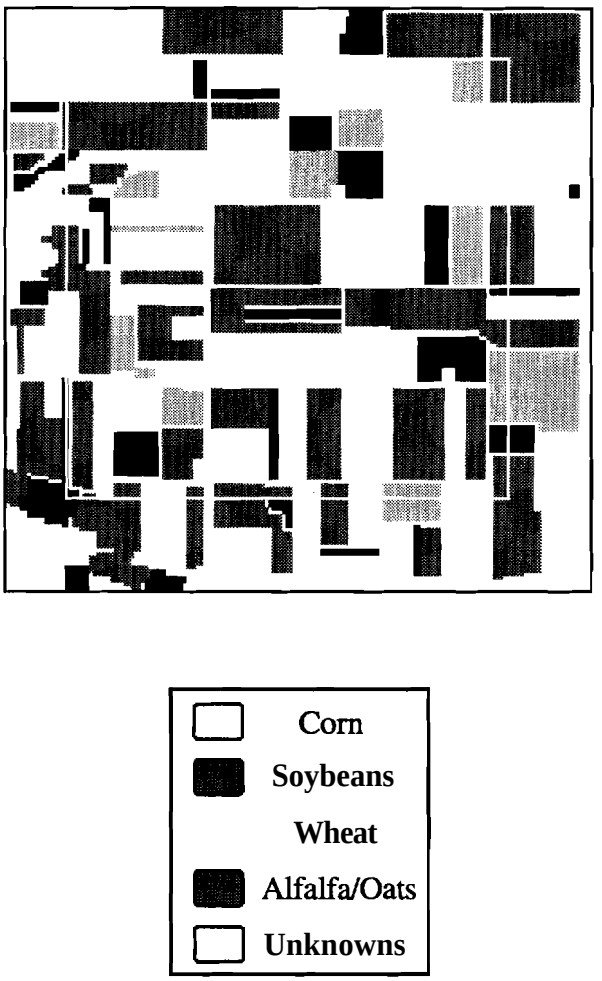


Figure 3.8 Associated Ground Truth Map.; The white box shows the 128 by 128 portion of selected test field.

3.5.2 Spatial Contextual Classification with Interpixel Correlation Context

To test how much increase of classification accuracy can be attained by incorporating spatial correlation context into classification, the following 4 different classifiers in Table 3.2 with only spatial interpixel correlation context, were applied to the July and September data.

Table 3.2 Description of Spatial Contextual Classifiers with Interpixel Correlation Context Only.

Classifier	Description
RECU - 1	Recursive spatial contextual classifier as in *eq.* (3.28) (With isotropy assumption)
RECU - 2	Recursive spatial contextual classifier as in *eq.* (3.28) (Without isotropy assumption)
C M - 1	Spatial correlation context only for homogeneous pixels with eq. (3.29) (With isotropy assumption)
CM - 2	Spatial correlation context only for homogeneous pixels with eq. (3.29) (Without isotropy assumption)

RECU stands for a recursive spatial contextual classifier in eq. (3.28). Since this recursive classifier is very time-consuming, in the classifiers of CM - 1 and 2, the homogeneity test in eq. (3.29) was first carried out to find homogeneous group of pixels. If the pixels $\{x(r), X_S(r)\}$ were homogeneous according to the test in eq. (3.29), then one of the M classes which maximized the joint conditional probability in eq. (3.28) was assigned to all pixels in $\{x(r), X_S(r)\}$ without checking every combinations of $\{c(r), C_S(r)\}$. If the pixels $\{x(r), X_S(r)\}$ failed the homogeneity test, those pixels were classified by a pixelwise maximum likelihood classification. These classifiers are denoted by CM. In both the RECU and CM classifiers, two conditions, that is, with and without isotropy assumptions in eq. (3.16.a) and eq. (3.18.a), were tested.

For comparisons, a pixelwise maximum likelihood classifier (ML) was used to classify the July and September data sets. Classification accuracies are summarized in Table 3.3 and 3.4.

Table 3.3 Percent Classification Accuracy of July Data by Spatial Contextual Classifier with Interpixel Correlation Context Only.

Classifier	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
ML	90.18	57.72	68.72	77.89	73.63	74.37
RECU - 1	93.70	56.00	73.33	80.97	76.00	76.00
RECU - 2	94.10	56.04	72.67	81.26	76.02	76.16
CM - 1	92.98	62.69	72.59	79.87	77.03	77.97
CM - 2	92.95	62.62	72.67	80.09	77.08	77.96

CM-1 : THD = -150 , CM-2 : THD = -150

Table 3.4 Percent Classification Accuracy of September Data by Spatial Contextual Classifier with Interpixel Correlation Context Only.

Classifier	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
ML	82.59	55.06	51.28	47.07	59.00	65.28
RECU - 1	83.07	61.85	57.12	46.93	62.24	68.51
RECU - 2	82.96	62.46	56.38	46.63	62.11	68.59
CM - 1	85.57	62.04	61.32	49.78	64.68	70.34
CM - 2	84.82	66.44	27.08	64.49	60.71	69.98

CM-1 : THD = -150 , CM-2 : THD = -150

In the July data set, for all classifiers tested above, there was considerable confusion of soybeans and wheat into alfalfa/oats. The spatial classifiers increased the classification accuracies compared to the maximum likelihood classifier (ML) as much as 5% for all classes except soybeans. The recursive classifiers (RECU-1, 2) were better than the pixelwise ML classifier by 1.63% and 1.79% in overall classification accuracy. CM-1 and CM-2 were better than the ML classifier by 3.6% and 3.59% in overall classification accuracy. The CM classifiers gave better results than the recursive (RECU) classifier for the class

soybeans, and the overall and class average classification accuracies were both better than the totally recursive cases (RECU-1, 2). Note that the CM classifiers are implicitly relying on the spatial class dependency context since a homogeneous group of pixels, $\{x(r), X_S(r)\}$, are assigned the same class simultaneously. When there are many homogeneous fields, the approach of first testing homogeneity and classifying homogeneous pixels, would give better performance than a totally recursive approach. It can also reduce the computational time. There were not significant differences between the isotropic and non-isotropic cases. The estimated values of $\theta_{(1,0),i}$ and $\theta_{(0,1),i}$ in eq. (3.16.a) were very similar to each other and this caused similar classification results.

In the case of the September data, soybeans and wheat were also confused mostly as **alfalfa/oats**. Generally, the spatial classifiers increased the classification accuracy as much as 10% for all classes but **alfalfa/oats**. There were 3.23% ~ 5.06% of increase in overall classification accuracy compared to the ML classifier. Again, there were not noticeable differences between isotropic and non-isotropic assumptions.

Several threshold values for the homogeneity criterion in eq. (3.29) were tested in the CM classifiers. Classification accuracy of CM -1 and CM -2 were observed not so sensitive to a small change of threshold value as seen in Fig. 3.9.

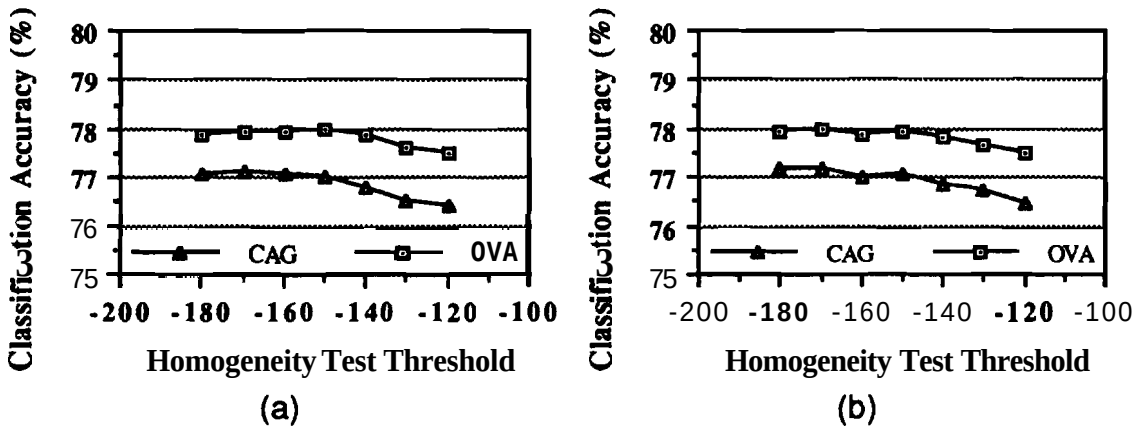


Figure 3.9 **Classification Accuracies of July Data with Various Thresholds for the Homogeneity Test. (a) CM-1 Classifier. (b) CM-2 Classifier.**

In the Ckl-classifiers, if the threshold value is set up too high, then only a small number of pixels will be classified with spatial correlation contextual information. But if it is too small, then many pixels will be considered as homogeneous with their neighbors and will result in undesirable blurs near field boundaries. In the experiments with July and September data, threshold values between -140 ~ -170 were generally satisfactory.

For visual assessment of classification performance, the classification maps of the spatial classifiers considered here and the maximum likelihood classifier are also shown in Fig. 3.10 ~ 3.12. The classification maps of pixelwise maximum likelihood classifiers shown in Fig. 3.10 have many isolated spots and most of them were erroneous classifications (For the locations of errors, see Fig. 3.16). The spatial classifiers provided much cleaner class maps as shown in Fig. 3.11 and 3.12. The totally recursive classifiers (RECU-1, 2) resulted in less spatially small isolated classes than the classifiers CM-1, 2 which checked homogeneity first.

Since the recursive classifiers utilize spatial correlation context for all pixels in the given image, classification errors tend to be blocky. That is, when a small group of pixels are incorrectly classified, this error region tends to grow by encroaching into its neighbors through the recursive process. In the CM classifiers, this growing of error regions is prevented by the homogeneity test. This gives a slightly higher classification accuracy for CM classifiers although the class maps of the totally recursive classifiers look cleaner.

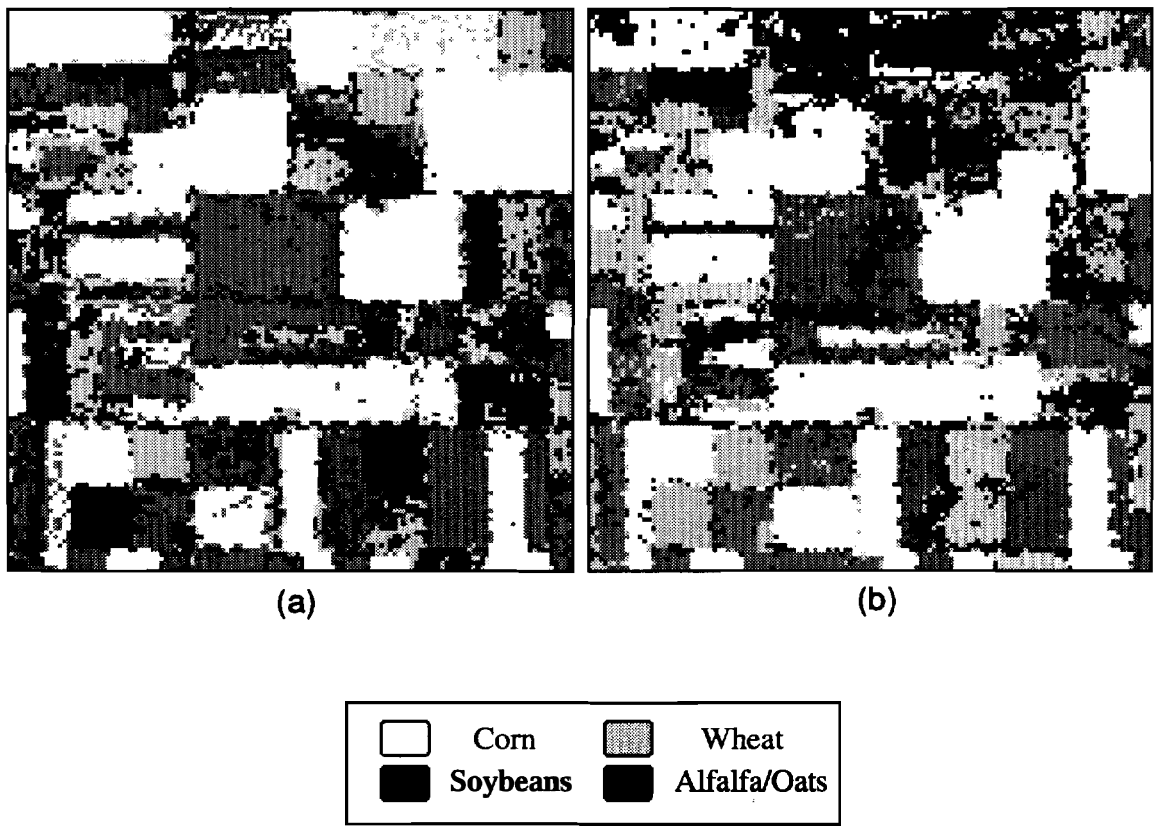


Figure 3.10 **Classification Maps Obtained by the Pixelwise Maximum Likelihood Classifier. (a) July Data. (b) September Data.**

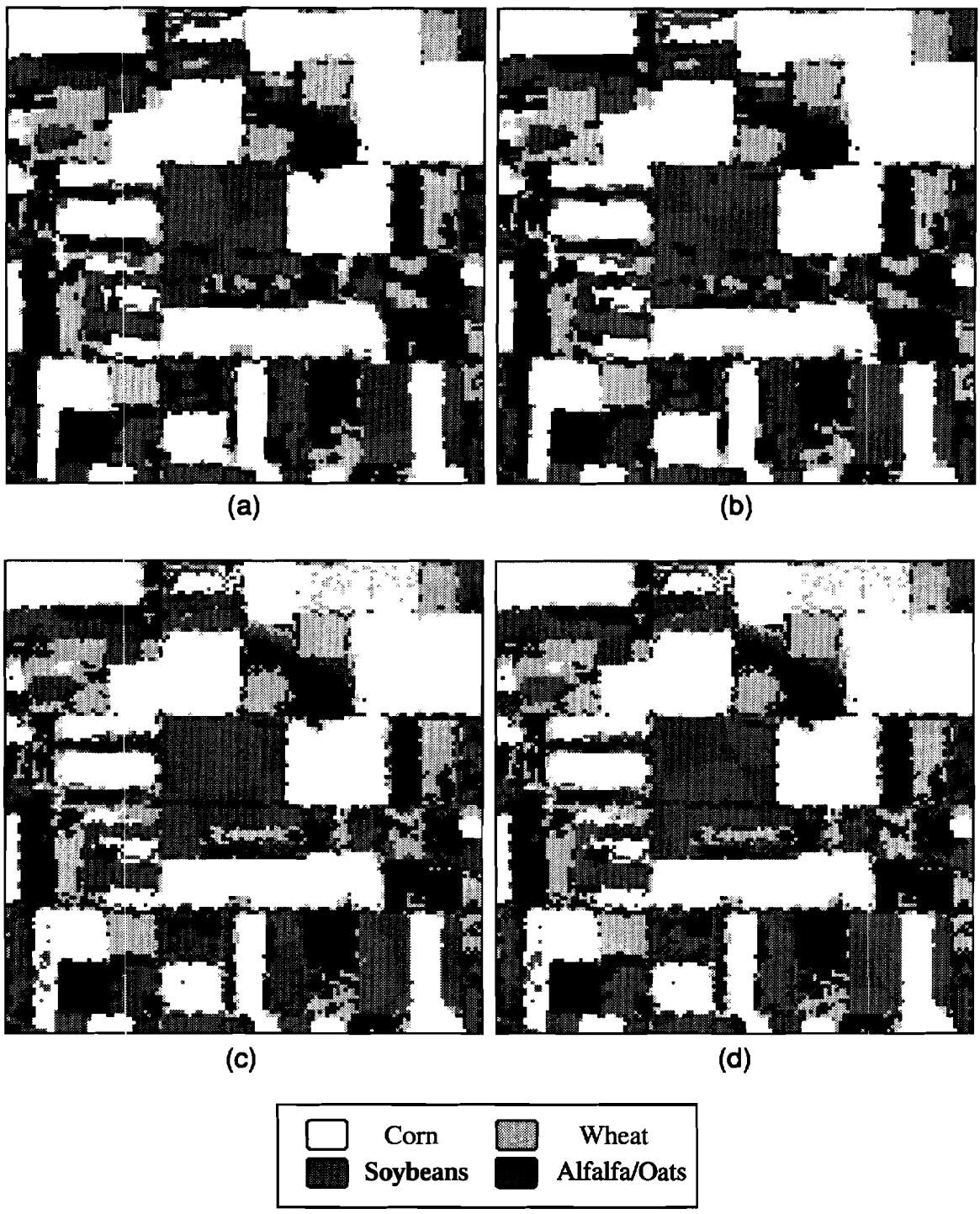


Figure 3.11 Classification Maps of July Data Obtained by the Spatial Classifier with Interpixel Correlation Context Only. (a) Isotropic Recursive (RECU-1) Classifier. (b) Non-isotropic Recursive (RECU-2) Classifier. (c) Isotropic CM (CM-1) Classifier. (d) Non-isotropic CM (CM-2) Classifier.

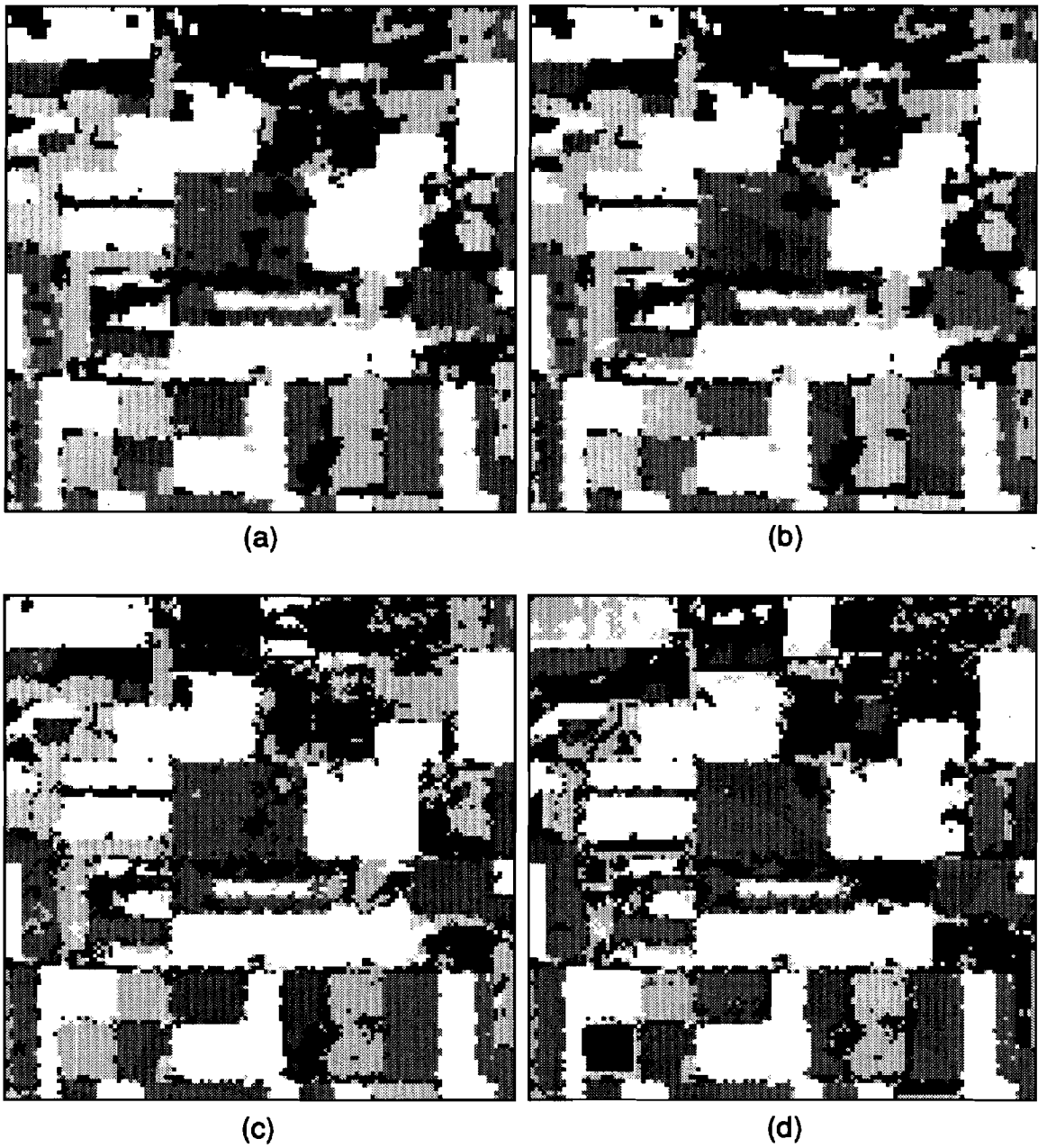


Figure 3.12 Classification Maps of September Data Obtained by the Spatial Classifier with Interpixel Correlation Context Only. (a) Isotropic Recursive (RECU-1) Classifier. (b) Non-isotropic Recursive (RECU-2) Classifier. (c) Isotropic CM (CM-1) Classifier. (d) Non-isotropic CM (CM-2) Classifier.

3.5.3 Spatial Contextual Classification with Class Label Dependency Context

In the second experiment, only the class label dependency context, $Pr\{C_S(r), c(r)\}$, was used in spatial classification. That is, the feature vectors were assumed to be class-conditionally independent as in eq. (3.2) and the classifier in eq. (3.35) was used. Classification results are shown in Table 3.5. In this experiment, only the GRF prior model was considered.

Table 3.5 **Percent Classification Accuracy of Spatial Contextual Classifier with Class Label Dependency Context Only.**

Data Set	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
July Data	94.51	57.28	73.50	80.82	76.53	76.82
September Data	83.79	59.52	52.51	50.07	61.47	67.86

b = 30

When the spatial class dependency context was used via H_{SP} with the prior probabilities based on the Gibbs random field, the classification accuracy was increased by 2 ~ 3% over the pixelwise maximum likelihood classification. In the case of the July data, there were significant increases in classification accuracy of the classes corn and wheat. In September data, there was about a 4% increase of classification accuracy for the class soybeans and the other classes also showed classification accuracy increases.

Although some algorithms which can estimate parameter b in eq. (3.35) are available (Dattatreya 91), in modeling the conditional prior probabilities in eq. (3.33), several different values of b were tested as shown in Fig. 3.13 to see its effect on the GRF prior model. The classification performance was observed to increase as b became larger to a certain value and to level off thereafter. When b was small, there were no significant changes of classification accuracy compared to the case without spatial priors. As b increased, classification accuracy was observed to increase to a certain extent and then start to decrease. Larger values of b means more emphasis given to spatial class homogeneity. The result in Table 3.5 were obtained with the best result for various b's as shown in Fig. 3.13. The classification result was not so sensitive to the value of b if it was large enough.

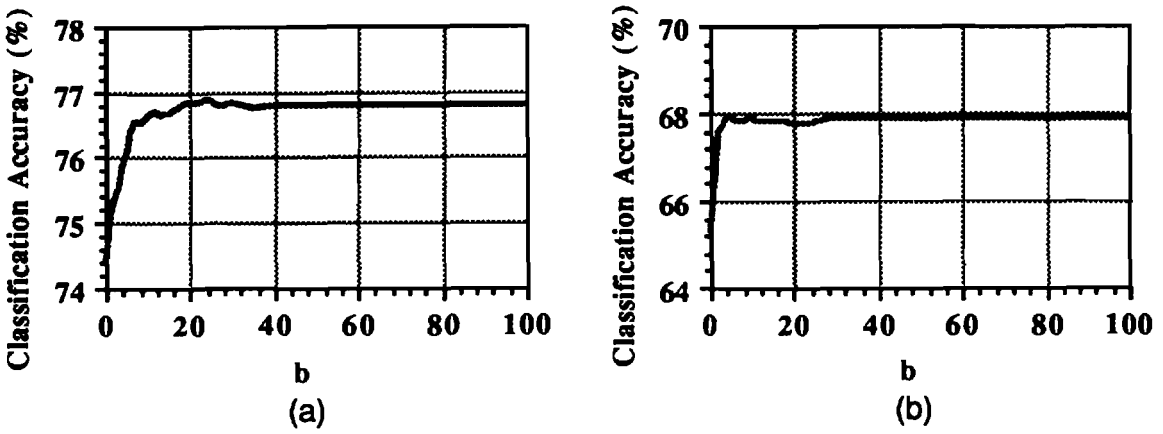


Figure 3.13 Percent Overall Classification Accuracy (OVA) Versus b in the Spatial Contextual Classification Using H_{SP} in Eq. (3.35). (a) July Data Set. (b) September Data Set.

Figure **3.14** shows a histogram of the differences of the first and second largest log-likelihood values of each pixel in July data set. Approximately **45%** of the pixels have differences of first and second largest log-likelihood values larger than 20.

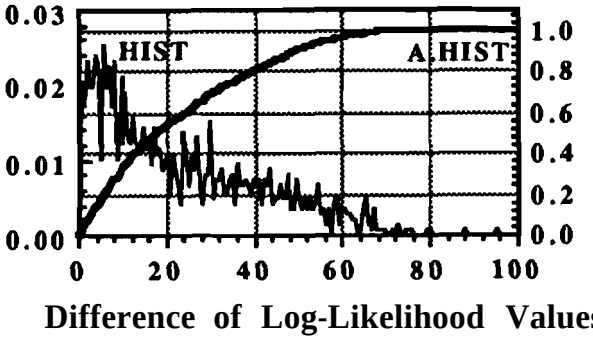


Figure 3.14 Histogram (Hist) and Accumulated Histogram (A.Hist) of Differences Between First Largest And Second Largest Log-Likelihood Values Of Pixels in July Data Set.

Therefore if the spatial prior term in eq. **(3.35)** is too small compared to the differences of log-likelihood values, the inclusion of spatial priors won't change the classification performance much.

Figure **3.15** shows classification maps of July and September data when only the **class** label dependency context was used in spatial classification. Imposing class label coherence for spatially neighboring pixels with the Gibbs random field in eq. **(3.35)** was very effective in removing scattered and isolated classification errors as shown in Fig. **3.15.** Field boundaries were more regular. Figure **3.16** and **3.17** show error maps which identify the locations **where** error occur.

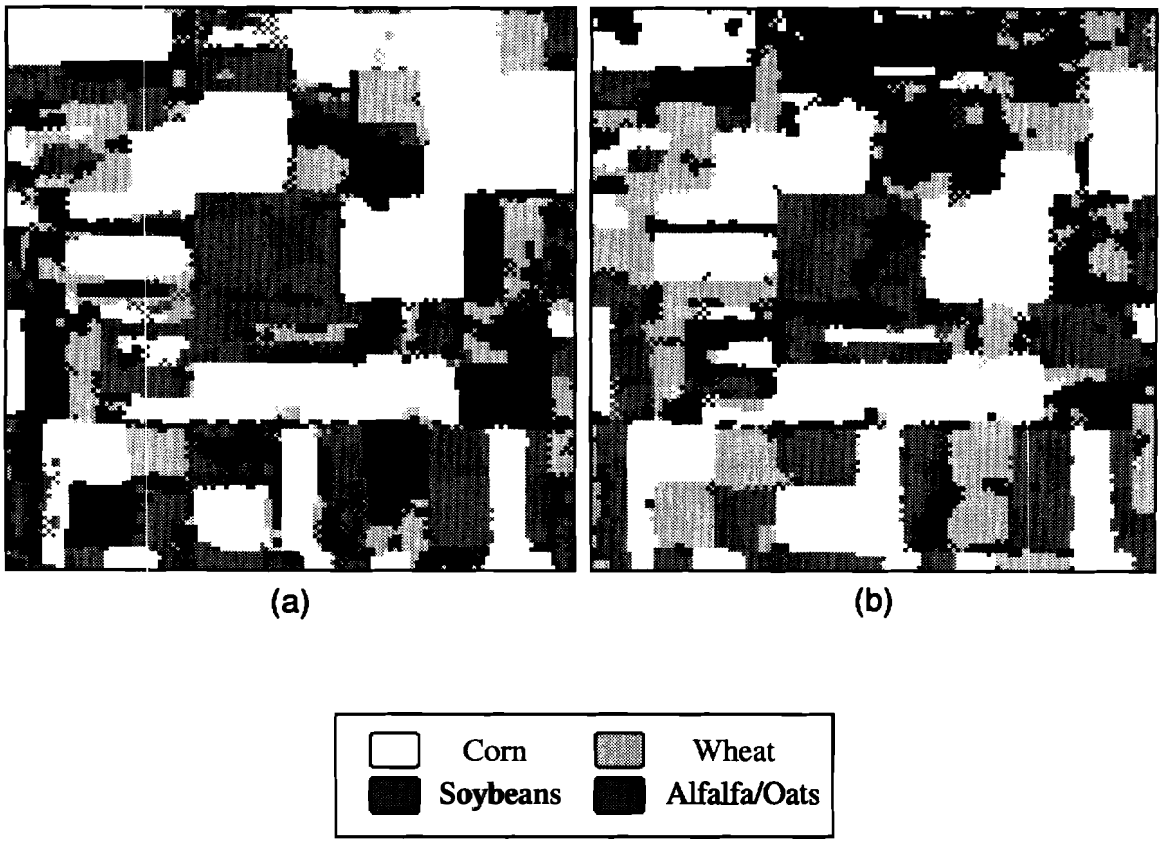


Figure 3.15 **Classification Maps Obtained by the Spatial Classifier with Class Label Dependency Context Only. (a) July Data Set. (b) September Data Set.**

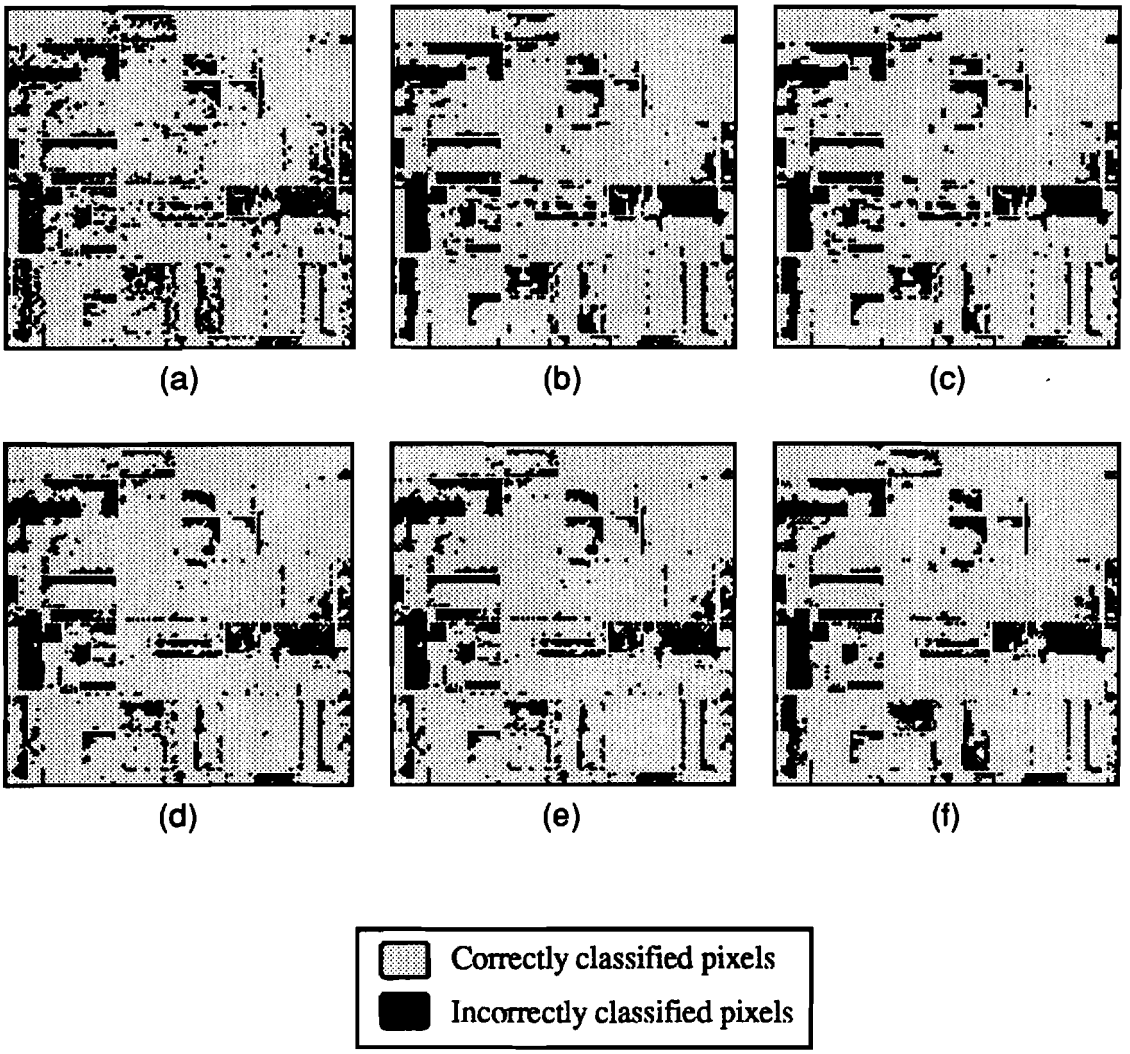


Figure 3.16 Error Maps of the July Data Set without Spatial Class Dependency Context. (a) Pixelwise maximum likelihood classifier. (b) Isotropic Recursive (RECU-1) Classifier. (c) Non-isotropic Recursive (RECU-2) Classifier. (d) Isotropic CM (CM-1) Classifier. (e) Non-isotropic CM (CM-2) Classifier. (f) with Only Spatial Class Dependency Context in eq. (3.35).

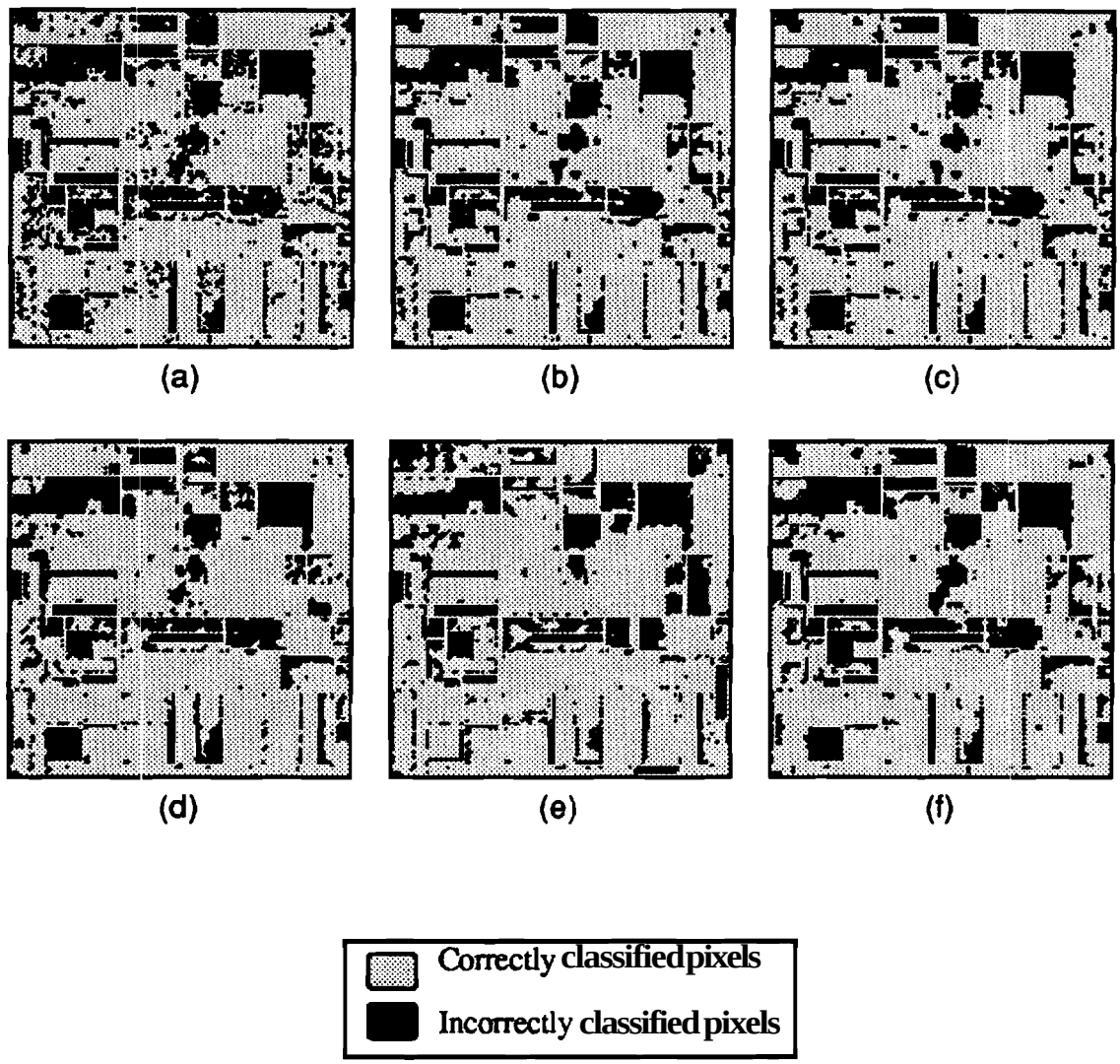


Figure 3.17 Error Maps of the September Data Set without Spatial Class Dependency Context. (a) Pixelwise maximum likelihood classifier. (b) Isotropic Recursive (RECU-1) Classifier. (c) Non-isotropic Recursive (RECU-2) Classifier. (d) Isotropic CM (CM-1) Classifier. (e) Non-isotropic CM (CM-2) Classifier. (f) with Only Spatial Class Dependency Context in eq. (3.35).

3.5.4 Spatial Contextual Classification with Both Interpixel Correlation Context and Class Label Dependency Context

In the third experiment, spatial class dependency context in the form of spatial prior probability, $Pr\{C_S(r), c(r)\}$, is also included in the classification addition to the spatial interpixel correlation context. The same value of b in the GRF prior model is used as with the case of spatial class dependency context only in previous section.

When pixels are classified with CM classifiers which check the homogeneity of a given group of pixels $\{x(r), X_S(r)\}$, the class of homogeneous groups of pixels is not changed by additionally incorporating spatial prior probability since the pixels are classified simultaneously to the same class. However, those pixels which fail the homogeneity test are subjected to pixelwise maximum likelihood classification with the spatial priors. Therefore, only those inhomogeneous pixels undergo the recursive procedure for class label dependency context. This procedure is capable of utilizing the spatial interpixel correlation context where it is most suitable, namely in homogeneous regions and to use only the class label dependency context for inhomogeneous pixels. This will significantly reduce erroneous pixels near homogenous pixels or field boundaries.

Table 3.6 and 3.7 summarize classification accuracies obtained by using both of the spatial contexts simultaneously. CM classifiers which first test homogeneity of the pixels and then selectively apply the spatial contextual rule were observed to perform better than the totally recursive cases (RECU-1, 2).

Table 3.6 **Percent Classification Accuracy of the July Data using the Spatial Contextual Classifier with Both Interpixel Correlation and Class Label Dependency Context.**

Classifier	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
RECU - 1	94.84	57.95	74.49	81.92	77.30	77.41
RECU - 2	95.02	57.97	74.16	81.84	77.25	77.46
CM - 1	94.53	63.06	73.83	80.97	78.10	79.00
CM - 2	94.50	63.10	73.91	81.33	78.21	79.04

CM-1 : THD = -150, CM-2 : THD = -150, b : 30

Table 3.7' Percent Classification Accuracy of the September Data using the Spatial Contextual Classifier with Both Interpixel Correlation and Class Label Dependency Context.

Classifier	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
RECU - 1	83.74	61.05	55.72	49.63	62.54	68.66
RECU - 2	83.92	61.41	56.13	49.85	62.83	68.93
CM - 1	85.77	62.39	61.81	49.71	64.92	70.60
CM - 2	85.18	66.94	25.76	65.30	60.79	70.28

CM-1 : THD = -150, CM-2 : THD = -150, b : 30

Compared with the pixelwise maximum likelihood classifier results in Table 3.3, all classes were classified much better except the class wheat in September data with CM-2. As with the results in Table 3.3 and 3.4 which were obtained by using only the interpixel correlation context, the results in Table 3.6 an 3.7 were generally better. As for July data, simultaneous incorporation of both spatial contexts increased classification accuracies for all classes comparedl with the previous cases of using only one type of spatial contextual information. However, in the September data, although both the overall and class (averaged classificalion accuracies showed improvement over the previous results in Table 3.4 and 3.5, some classes such as soybeans and wheat, had slightly worse classification accuracy.

Figure 3.18 and 3.19 show the classification maps obtained with the spatial classifiers with both spatial contexts. The corresponding error maps are presented in Fig. 3.20 and 3.21. For a comparison, the error maps of the pixelwise maximum likelihood classifier and the spatial classifier in eq. (3.35) with only spatial class dependency context are also included in Fig. 3.20 and 3.21.

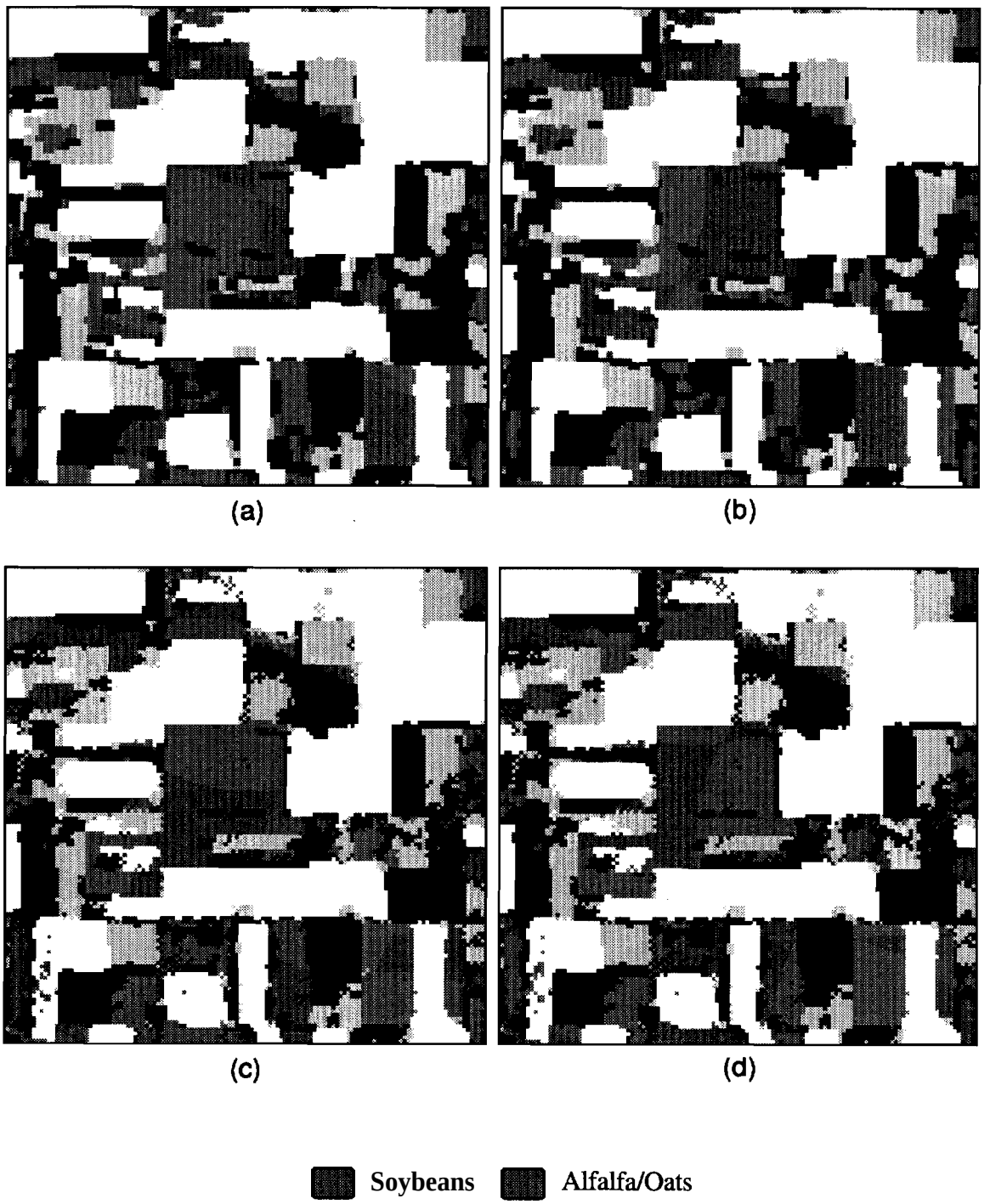


Figure 3.18 Classification Maps of the July Data Obtained by Spatial Contextual Classifier with Both Interpixel Correlation and Class Label Dependency Contexts. (a) Isotropic Recursive (RECU-1) Classifier. (b) Non-isotropic Recursive (RECU-2) Classifier. (c) Isotropic CM (CM-1) Classifier. (d) Non-isotropic CM (CM-2) Classifier.

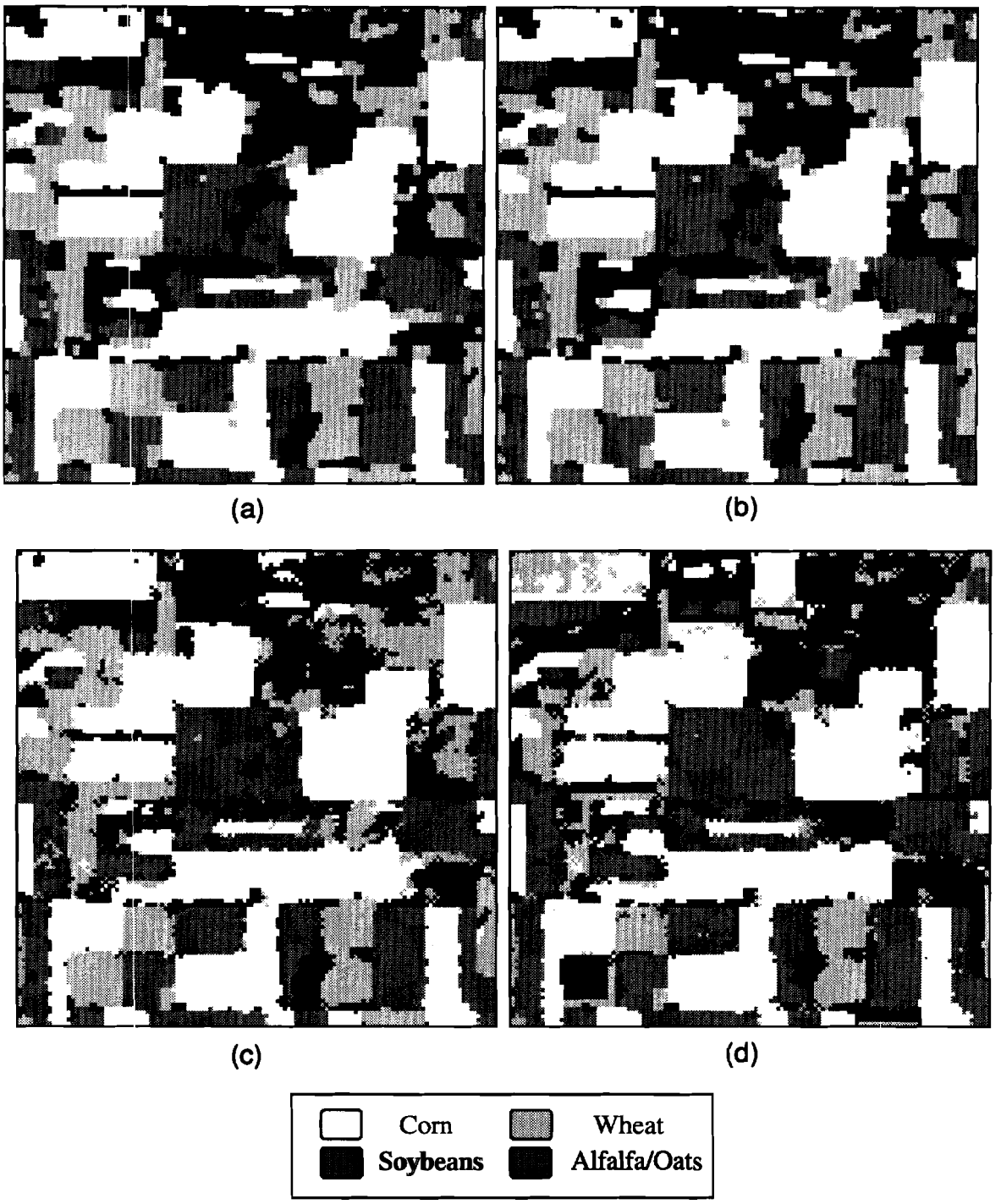


Figure 3.1'9 Classification Maps of the September Data Obtained by Spatial Contextual Classifier with Both Interpixel Correlation and Class Label Dependency Contexts. (a) Isotropic Recursive (RECU-1) Classifier. (b) Non-isotropic Recursive (RECU-2) Classifier. (c) Isotropic CM (CM-1) Classifier. (d) Non-isotropic CM (CM-2) Classifier.

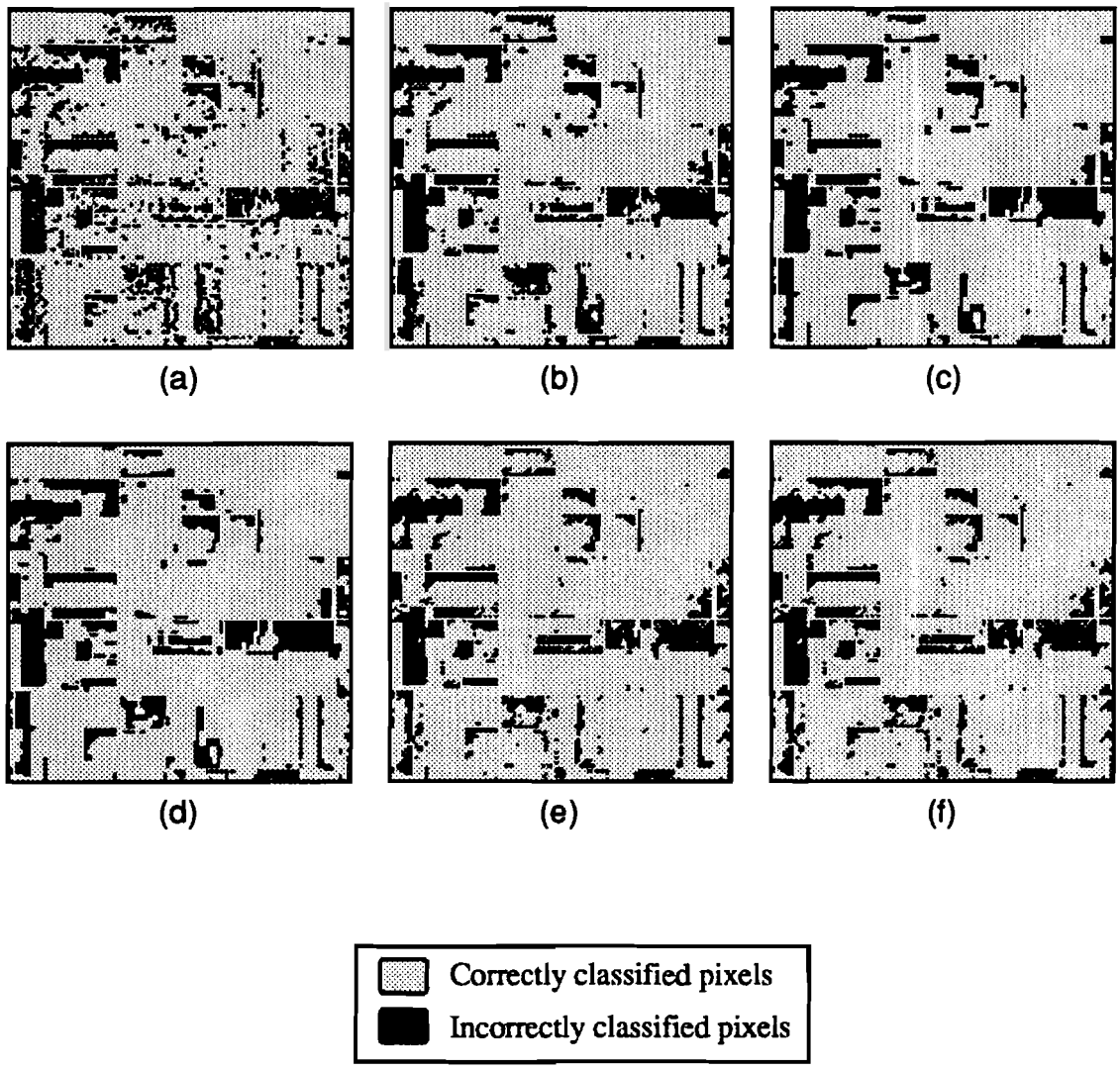


Figure 3.20 Error Maps of the July Data Set with Spatial Class Dependency Context. (a) Pixelwise Maximum likelihood classifier (without Spatial Class Dependency context). (b) with Only Spatial Class Dependency context. (c) Isotropic Recursive (RECU-1) Classifier. (d) Non-isotropic Recursive (RECU-2) **Classifier.** (e) Isotropic CM (CM-1) Classifier. **(f)** Non-isotropic CM (CM-2) Classifier.

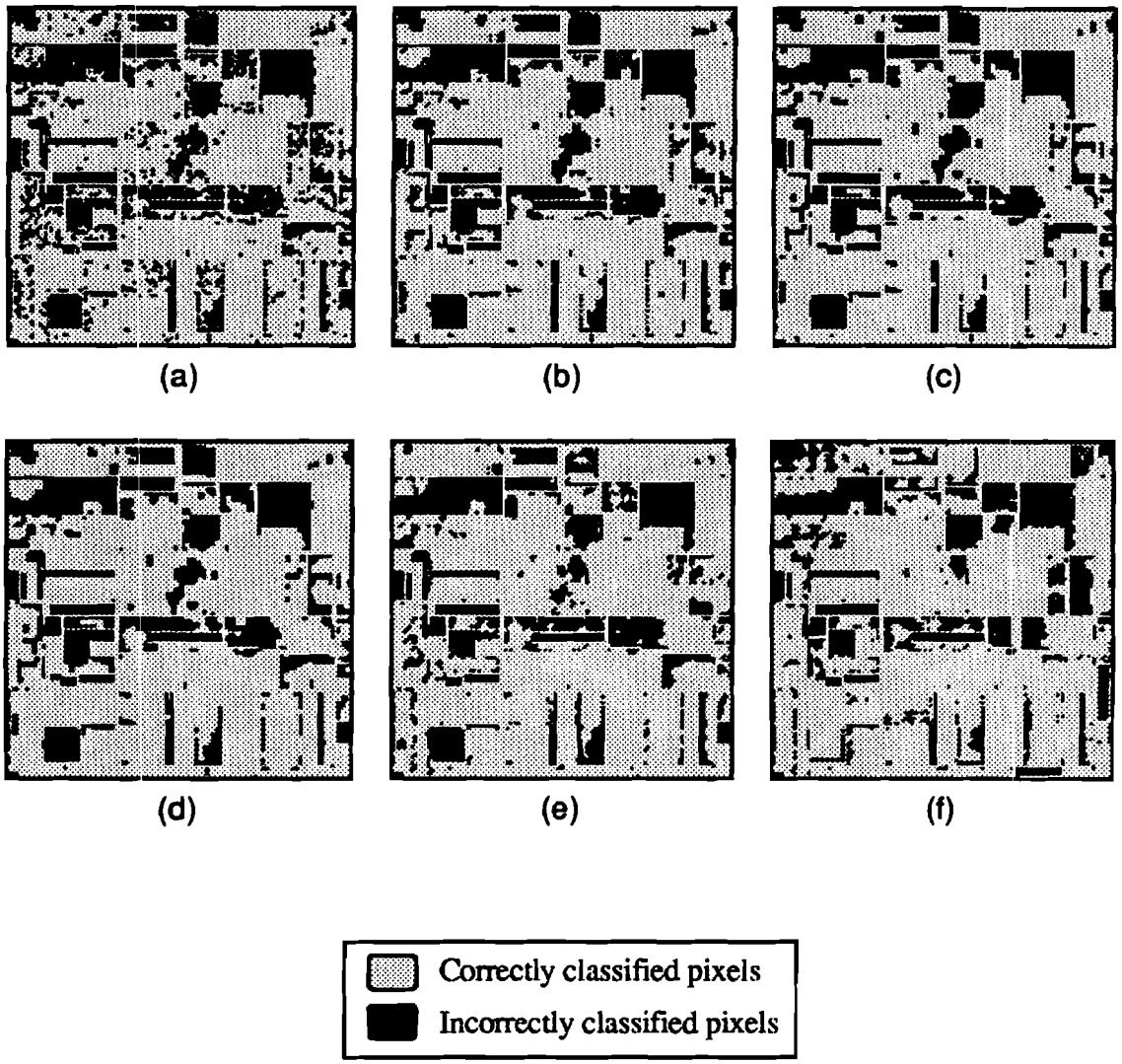
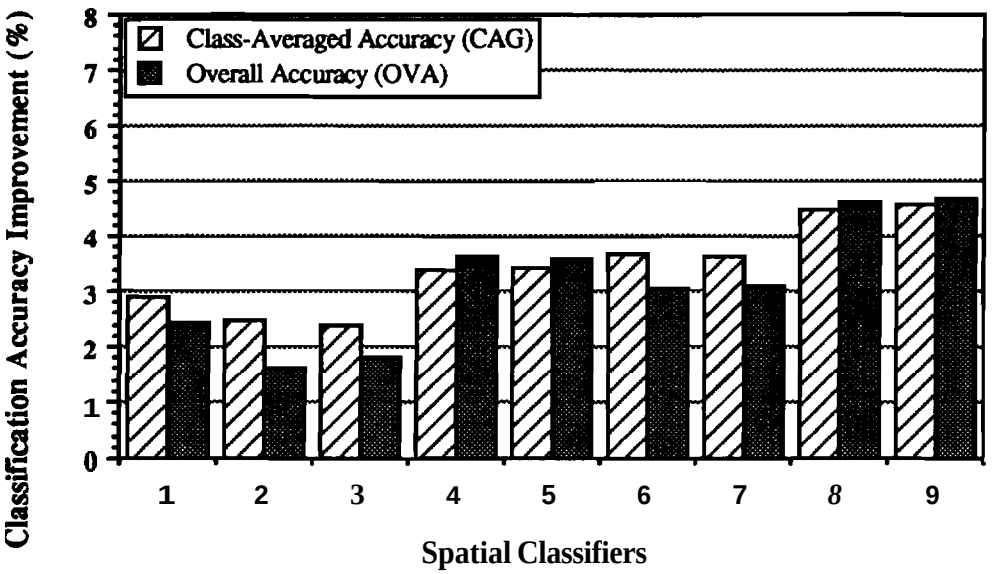


Figure 3.21 Error Maps of the September Data Set with Spatial Class Dependency Context. (a) Pixelwise maximum likelihood classifier (without Spatial Class Dependency context). (b) with Only Spatial Class Dependency context. (c) Isotropic Recursive (RECU-1) Classifier. (d) Non-isotropic Recursive (RECU-2) Classifier. (e) Isotropic CM (CM-1) Classifier. (f) Non-isotropic CM (CM-2) Classifier.

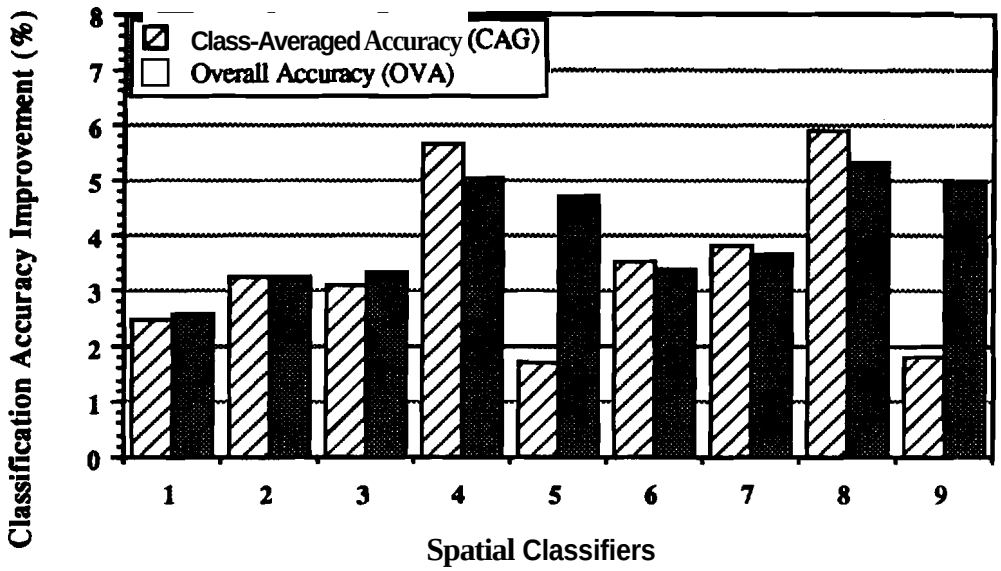
As shown in Figure 3.18 and 3.19, the spatial classifiers accounting for both spatial contexts simultaneously resulted classification maps with many fewer isolated errors. These clean class maps would be very useful in applying the classification results since they show field boundaries very clearly.

In summary, Fig. 3.22 and 3.23 show the classification accuracy increases over the pixelwise maximum likelihood classifier obtained by incorporating spatial contextual information.



Class Label dependency Context only - classifier 1	
Interpixel Correlation Context Only	Both of Spatial Contexts
RECU - 1 : classifier 2	RECU - 1 : classifier 6
RECU - 2 : classifier 3	RECU - 2 : classifier 7
CM - 1 : classifier 4	CM - 1 : classifier 8
CM - 2 : classifier 5	CM - 2 : classifier 9

Figure 3.22 **Classification Accuracy Improvement by Spatial Contextual Information over Pixelwise Maximum Likelihood Classifier (July Data).**



Class Label dependency Context only - classifier 1	
Interpixel Correlation Context Only	Both of Spatial Contexts
RECU - 1 : classifier 2	RECU - 1 : classifier 6
RECU - 2 : classifier 3	RECU - 2 : classifier 7
CM - 1 : classifier 4	CM - 1 : classifier 8
CM - 2 : classifier 5	CM - 2 : classifier 9

Figure 3.23 **Classification Accuracy Improvement by Spatial Contextual Information over Pixelwise Maximum Likelihood Classifier (September Data).**

As seen in Fig. 3.22 and 3.23, there were about 2 ~ 6% classification accuracy increases. It is very difficult to judge which contextual information is more useful in improving classification results. While the answer should be dependent on each particular data set classified, from a computational viewpoint, however, the interpixel correlation context is much more complex to properly account for due to calculating joint class conditional probabilities of augmented feature vectors. An incorporation of the class label dependency context is relatively siimple. In the case of the Gibbs random field (GRF) model, it only requires to count the number of neighboring pixels which have the same class as the center pixel under consideration. The parameter b in eq. (3.33) decides the relative importance of homogeneity of the class labels compared to the differences in

likelihood values. Since the degree of this class label coherence represented by the value b in eq. (3.33) is not necessarily all the same over the given image, - for example, larger weights on this class label coherence might be beneficial inside homogeneous fields and lesser weights in inhomogeneous fields or near boundaries-, spatially different values of b could be useful in classifying given image data. However, this is left unsolved for future research.

3.6 Conclusion

In this chapter, the spatial contextual part $H_{SP}(\cdot)$ in the spatio-temporal contextual classifier derived in Chapter 2, was applied to a problem of spatial classification. Spatial interpixel correlation context was represented by the joint class-conditional probabilities and, class label dependency context was accounted for by the joint prior probability which was modeled by a Gibbs random field.

Experiments were carried out with two Landsat Thematic Mapper data sets with these two spatial contexts. In incorporating spatial interpixel correlation context, the procedure of first testing homogeneity of a given group of pixels and then selectively utilizing contextual information was very useful both in improving classification performance and reducing computational complexity. The class label dependency context was relatively simple and computationally less demanding, and it was very useful in improving classification accuracy by removing spatially small isolated classes. The Gibbs random field model was effective in implementing class dependency context.

The spatial classifiers discussed in this chapter not only increased classification accuracies over the pixelwise maximum likelihood classifier and but also resulted in classification maps with far less isolated errors and clean field boundaries.

CHAPTER 4

'TEMPORAL CONTEXTUAL CLASSIFICATION: A DECISION FUSION APPROACH

4.1 Introduction

This chapter addresses the temporal contextual classification problem. A few desirable properties of the temporal classifier to be developed are as follows.

1. Since there are usually only a limited number of training samples available for each temporal data set, employing a temporal contextual classifier should not require extra training samples additional to those already available for pixelwise non-temporal contextual classification.

2. Thus, it should be possible to train a temporal classifier separately for each temporal data set. For this requirement, it is quite common to assume class-conditional independence of features of different temporal data sets.

3. It will be also very desirable if a temporal contextual classifier can facilitate distribution of computation required for classification over different times. In other words, as new temporal data sets becomes available, the intermediate results already computed with previous temporal data sets should be able to be updated so that they can be again used when the next temporal data set becomes available.

4. Different temporal data sets can have distinct properties and varying discriminating power, therefore, one should be able to associate a "reliability factor." A temporal classifier would be very useful if it can accommodate different reliability factors associated with each temporal data sets.

Noting that a temporal contextual classification can be thought as a special example of the multisource classification problem with temporal data sets being considered as separate information sources, focus is brought on a more general problem of multisource classification and the term "temporal contextual" classification and "multisource classification" will be interchangeably used in this chapter.

Unlike those customary data combination approaches in multisource classification, this chapter is addressing a totally new multisource classifier which is based on a fusion of "class decisions" of each separate data set. Each data set is separately fed into a local classifier and a final classification is performed by summarizing these local class decisions. An optimum decision fusion rule based on a minimum expected cost is derived. This new decision fusion rule is developed to be capable of handling not only data set reliabilities but also classwise reliabilities of each data set.

The temporal contextual classification algorithms discussed in this chapter will be used in spatial-temporal classification in Chapter 5 in conjunction with the spatial contextual parts developed in Chapter 3. When they are complete as is, experimental results with multiternporal data are included in this chapter.

4.2 Multisource Data Classification

With remarkable advances in sensor technology in many application fields, it becomes quite common to employ several sensors and to extract desirable information from the amassed set of all available data sets. This approach allows more reliable and improved results. One application of this, for example, can be found in an analysis of multisource data which deals with data sets obtained by multiple sensors possibly with different characteristics. Other than digital image data sets, non-image data sets such as geophysical measurernenl data or, cartographic data sets, etc., are also often available for analysis, for example, in a geographic information system (GIS). These disparate information sources are utilized simultaneously to improve the results of data analysis. There have been many efforts to effectively employ multisource data sets; for example, see (Benediktsson and Swain 92, Benediktsson et *al.* 90, Lee *et al.* 87). These can be categorized in terms of how the disparate information sources are combined

to attain a desired objective. In this chapter, we address a multisource classification algorithm based on an optimal fusion of decisions of each data set.

The idea of the decision fusion is to let each local classifier make a (local) decision based only on its own data set and forward the decision to the central classifier which finalizes a decision based on a set of local decisions and any available prior knowledge, such as the reliabilities of the respective local decisions. Since each data set is separately fed into its own classifier and only the decision of the local classifier is required by the fusion rule, this approach can significantly ease a requirement at the training stage and subsequently computational complexity.

Different information sources can have different degrees of reliability. One data set might be more reliable than another data set in the analysis of specific data. This is to be expected since the conditions or characteristics of the various sensors or data sets are not necessarily all the same. In classification problems, for example, one set of data might be able to provide more discriminating power than another data set if the classes are more separable from each other using that data set. It is also very likely that a certain class or a subset of classes may be discriminated more successfully than others, and it is clear that a less reliable data set should have less effect on the overall classification procedure, thus a classification algorithm should be able to appropriately deal with data set reliabilities. It will ,therefore be very useful to associate a reliability factor not only to the data set but also to the classes which the local classifier defines.

In this chapter, the reliability factor associated with each class will be called the "classwise reliability." A simplest use of this "classwise reliability" can be found where decision fusion is based on selecting the particular local decision which has the largest classwise reliability among other local decisions. Although data set reliabilities have been successfully utilized in combining disparate information sources, for example, in the approaches proposed by (Benediktsson and Swain 92, Benediktsson et *at.* 90, Lee *et at.* 87), few examples of considering classwise reliability can be found. One of the objectives of this research is the effective utilization of classwise reliabilities as well as data set reliabilities in classification.

The fusion of the decisions of different data sets can be formulated in a manner similar to M-ary distributed hypothesis testing problems, which have been a

subject of considerable research attention (Tenny and Sandell 86, Chair and Varshney 86, Reibman and Nolte 87, Hoballa and Varshney 89, Tang et *al.* 89) in such fields as radar systems and military surveillance systems. We apply the maximum likelihood decision fusion rule as in (Tang *et al.* 89) in a multisource classification problem and extend the result in (Tang *etal.* 89) by adopting a modified cost function to find the optimum fusion rule which can handle both data set and classwise reliabilities.

Applying a decision fusion approach in multisource classification has several advantages over conventional algorithms based on the conventional delta fusion. Since the algorithm based on a decision fusion requires only classes assigned using each data sets and doesn't need to keep the class-conditional probabilities, it. is very simple not only from the computational viewpoint but also from a memory requirement. It can also deal with several disparate data sets which have significantly different underlying distributions. For example, there can be a data set which cannot be successfully modeled by a set of statistical distribution functions on which the conventional data fusion multisource classification algorithm!; are formulated. However, forwarding only the classes assigned with each data set forfeits information of data fusion, although the prior information required in decision fusion can supplement the loss to some extent. Much simpler computation and reduced memory requirements would be able to reduce the performance degradation due to the loss. In many practical applications in which the information carried by posterior probabilities is inaccurate to some extent, however, if properly estimated, the prior information, supplied for the decision fusion process, can surpasses the information loss and result in better performance.

The organization of this chapter is as follows. In section 4.3, a brief review of multisource classification and its previous research is presented. In section 4.4, an optimal decision fusion algorithm based on minimum expected cost is derived. The problem of selecting data set and classwise reliabilities is addressed in section 4.5. Some comments on information combination structures in multisource classification are given in section 4.6. Experimental results with three remotely sensed multitemporal Thematic Mapper (TM) data sets are presented in section 4.7. Finally, section 4.8 concludes this chapter.

4.3 Review of Previous Works

Suppose there are p different data (or information) sources which produce a set of features (or, feature vectors) $\{x_1, \cdots, x_p\}$ where x_k, k = 1, ---, p, is a q-dimensional feature vector of the k^{th} information source (note that the dimension ***need not*** be the same for different data sets). The objective with these amassed multisource data sets is to make the best decision on the nature of the object observed as $\{x_1, \cdots, x_p\}$.

A decision is to be made among the classes in Ω_0, a set of user-defined information classes. In temporal contextual classification with these p multitemporal data sets, the set Ω_0, is typically the set of classes in the p^{th} data set, that is, Ω_p. However, in this chapter, Ω_0 can be any set of user-defined information classes. M_0 is the number of information classes in Ω_0. The term "information class" means a class which is directly of informational value to user according to the specific purpose of data analysis. If the purpose is for finding classes of objects on the ground via remotely sensed data, then, the list of information classes might include the names of objects on the ground, e.g., specific plant species. If it is for detecting a particular target, then, the information classes could be {target, non-target}. Since information classes are defined ordinarily based solely on the user's interests, they may not be separable in the feature space. Thus, in the classifier training stage, the given data sets might be analyzed, for example, through a clustering, to find a mutually exclusive and exhaustive set of sub-classes or "data classes" each of which can be modeled by an appropriate probability density function. Due to the computational complexity and a practical limitation on the requirement for training samples etc., the data sets are assumed, in general, to be class-conditionally independent of each other (see (Lee ***et al.*** 87) for a discussions of this assumption), and each data set is separately analyzed in the training stage. Therefore different data sets can have generally a distinct set of data classes with a different number of data classes. As defined in Chapter 2, Ω_k, k = 1, ---, p, is a set of data classes in the k^{th} data set with M_k elements.

The problem of multisource classification is to determine the optimum decision rule given the multisource data $\{x_1, \cdots, x_p\}$, a priori information such as Ω_0, Ω_k and the estimated probability density functions of the data classes in each Ω_k, k =

1, ---, p. The optimal decision rule in a Bayesian approach, is to find the class $\omega_j \in \Omega_0$ which maximizes the probability $P\{\omega_j \mid x_1, ---, x_p\}$. Under the independence assumption mentioned above, this is equivalent to finding a class ω_j maximizing $P(\omega_j)P(x_1 \mid \omega_j) \cdots P(x_p \mid \omega_j)$. Note that, in this Bayesian approach, each data set has the same effect on the final decision of Ω.

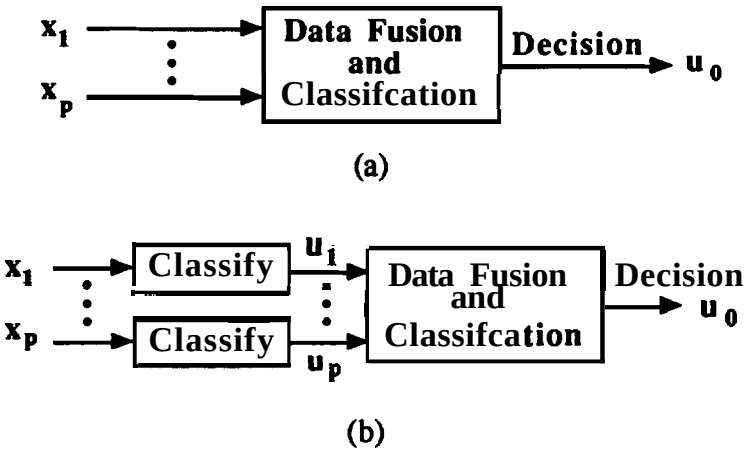


Figure 4.1 **Multisource Classification Structures. (a) Fusion of Features. (b) Fusion of Decisions.**

Generally, there are two different approaches to multisource data set classificalion as shown in Fig. 4.1.

The feature vectors of each data source (or sensor) can be fed into a central decision procedure as in Fig. 4.1.(a) to draw a final decision which is denoted by $u_0 \in \Omega_0$. For a detailed review on works in this category, refer to (Benediktsson *et al.* 90). In this decision procedure, a subset of original features might be used by applying a suitable feature extraction algorithm. A simple straightforward example of this approach might be the so called, extended vector approach in which all given feature vectors are used simultaneously to form a single compound feature vector for classification. The cascade classifier (Swain 78a) belongs also to this category. Another example might be the layer-based classifier where a different subset of features are used in each different level (Hoffer *et al.* 79). Lee *et al.* (Lee *et al.* 87) developed a statistical multisource classifier which was later extended by Benediktsson *et al.* (Benediktsson *et al.* 90) to be capable of having

reliability factors associated with data sets. In (Benediktsson et *al.* 90), a global membership function $F_j(\cdot)$ which is defined for $\omega_j \in \Omega_0, j = 1, ---, M_0$, as,

$$F_j(x_1, ---, x_p) = P(\omega_j) \prod_{k=1}^{p} \left[\frac{P(\omega_j \mid x_k)}{P(\omega_j)} \right]^{\alpha_k} \tag{4.1}$$

is used to perform data fusion and classification. Equation (4.1) shows how the individual weighted posterior probabilities affect the global membership function. α_k is a reliability factor associated with the k^{th} data set. The decision is made by selecting a class among Ω_0 which gives a maximum membership function value.

The reliabilities associated with the data sets, $\{\alpha_1, ---, \alpha_p\}$ in eq. (4.1) are set considering such factors as class separabilities, classification accuracies, and/or equivocation in such a way that the percentage change in the posterior probability of one data set is proportional to the percentage change in the global membership function multiplied by the reliability factor of that data set as,

$$\frac{\partial F_j(X)}{F_j(X)} = \alpha_k \frac{\partial P(\omega_j \mid x_k) / P(\omega_j)}{P(\omega_j \mid x_k) / P(\omega_j)}$$

The evidential reasoning approach based on interval-valued probabilities has been also used to perform multisource classification with data set reliabilities (Kim and Swain 90). However, neither of these approaches handle classwise reliabilities.

If the conditional probability of feature vector x_k given both the data class and information class is assumed to be the same as that of feature vector x_k given only the data class, then, the source-specific posterior probability is computed as,

$$P(\omega_j \mid x_k) = \frac{1}{P(x_k)} \sum_{u_k \in \Omega_k} P(x_k \mid u_k)\, P(\omega_j \mid u_k)\, P(u_k) \tag{4.2}$$

The probability $P(\omega_j \mid u_k)$, called the "class transition probability," shows a relationship between the data class and the information class. As discussed in Chapter *i!*, in multitemporal classification, this class transition probability provides temporal contextual information between temporal data sets.

Define a class transition matrix, $T(\Omega_0|\Omega_k)$ which consists of class transition probabilities, $P(\omega_j \mid u_k)$, $\omega_j \in \Omega_0$ and $u_k \in \Omega_k$, as,

$$T(\Omega_0 \mid \Omega_k) \equiv \begin{bmatrix} P(\omega_1 \mid u_1) & \cdots & P(\omega_1 \mid u_{M_k}) \\ \vdots & P(\omega_i \mid u_j) & \vdots \\ P(\omega_{M_0} \mid u_1) & \cdots & P(\omega_{M_0} \mid u_{M_k}) \end{bmatrix} \qquad (4.3.a)$$

Since there are M_0 and M_k elements respectively in Ω_0 and Ω_k, $T(\Omega_0 \mid \Omega_k)$ is M_0 by M_k. In the case of the cascade classifier (Swain 78a) which was developed for bi-temporal contextual classification, final classification is performed among $\Omega_0 = \Omega_p$, $p = 2$, and the matrix, $T(\Omega_2 \mid \Omega_1)$ provides information about class transitions between two temporal data sets. With this class transition matrix, the relation in eq. (4.2) can be equivalently written in a vector form as,

$$\vec{P}(x_k, \Omega_0) = T(\Omega_0 \mid \Omega_k)\, \vec{P}(x_k, \Omega_k) \qquad (4.3.b)$$

$$\text{where,} \quad \vec{P}(x_k, \Omega_0) \equiv \begin{bmatrix} P(x_k \mid \Omega_1)P(\Omega_1) \\ \vdots \\ P(x_k \mid \Omega_{M_0})P(\Omega_{M_0}) \end{bmatrix} \quad \text{and} \quad \vec{P}(x_k, \Omega_k) \equiv \begin{bmatrix} P(x_k \mid u_1)P(u_1) \\ \vdots \\ P(x_k \mid u_{M_k})P(u_{M_k}) \end{bmatrix}$$

Only one component is may be dominant over the others in $\vec{P}(x_k, \text{Ilk})$; for example, suppose $P(x_k \mid u_j)P(u_j)$ is dominantly larger than other components, then, eq. (4.3.b) is approximated by,

$$\vec{P}(x_k, \Omega_0) \approx \begin{bmatrix} P(\omega_1 \mid u_j) \\ \vdots \\ P(\omega_{M_0} \mid u_j) \end{bmatrix} P(x_k \mid u_j)\, P(u_j)$$

Since the term $P(x_k|u_j)P(u_j)$ is common for all classes in Ω_0, the relative differences between terms in $\{P(\omega_1|u_j), \cdots, P(\omega_{M_0}|u_j)\}$ determine an actual contribution of temporal contextual information. If these relative differences are much smaller than those between terms in $\{P(x_k|\omega_1)P(\omega_1), \cdots, P(x_k|\omega_{M_0})P(\omega_{M_0})\}$, then the information of $\vec{P}(x_k, \Omega_k)$ won't contribute much discriminating information in the global membership function in eq. (4.1).

Frequently, a certain class in one data set strongly indicates a particular class in Ω_0 and only one component in each row of $T(\Omega_0 \mid \Omega_k)$ is dominant over others in that row. Suppose comparing two classes, ω_i and ω_j in Ω_0 and class u_m and u_n in the k^{th} data set, strongly indicates those ω_i and ω_j classes, respectively. In this case, the ratio of information class-conditional probabilities is,

$$\frac{P(x_k \mid \omega_i)\, P(\omega_i)}{P(x_k \mid \omega_j)\, P(\omega_j)} \approx \frac{P(\omega_i \mid u_m)}{P(\omega_i \mid u_n)} \cdot \frac{P(x_k \mid u_m)\, P(u_m)}{P(x_k \mid u_n)\, P(u_n)}$$

The ratio of data class-conditional probabilities is shown to be directly related to the ratio of the corresponding information class-conditional probabilities. Note that the relative values of the data class-conditional probabilities are generally widely variant for different data sets. If any data set happens to have data class-conditional probabilities which have very large differences among them, the information class-conditional probabilities corresponding to this data set are very likely to dominate the global membership function in eq. (4.1) unless its data set reliability factor is very small.

In the approach of Fig. 4.1.(b), a final class decision is made by summarizing only the classification result of each data source. J. Tubbs and W. Alltop (Tubbs and Alltop 91) considered a problem of integrating classification results from multiple sensors and suggested a decision process based on a ranked lists of

class decisions. The local decision of each data source is denoted by u_k, k = 1, ---, p. The decision rule of each data source is assumed to be already determined. In general, the key problem of the approach in Fig. 4.1.(b) is how to determine an optimum decision $u_0 \in \Omega_0$ given the local decisions $\{u_1, ---, u_p\}$. This problem is very similar to that of M-ary distributed hypothesis testing.

There are two issues in distributed hypothesis testing, or the distributed detection problem, one being the design of the local classifiers and the other being the fusion rule of local decisions. R. R. Tenny and N. R. Sandell (Tenny arid Sandell 86) proposed first a distributed detection algorithm in the case of two sensors. Z. Chair and P. K. Varshney (Chair and Varshney 86) derived an optimum fusion rule when binary local decisions were given in a multiple sensor detection problem. Later, A. R. Reibman and L. W. Nolte (Reibman and Nolte 87) reported a system-wide optimum solution for a restricted case when the statistics and thresholds of the local detectors are assumed to be identical. Z. Tang et *al.* (Tang 89) presented a solution of the more general case of a distributed M-ary detection problem with multiple sensors which will be extended in this chapter by adopting a modified cost function to find an optimum fusion of local decisions.

4.4 Decision Fusion Approach in Multisource Classification

Suppose we have the problem of finding an optimum decision $u_0 \in \Omega_0$ given the local decisions $\{u_1, ---, u_p\}$. The decision rule of each data source is assumed to be already determined. This problem of decision fusion is analogous to the decision-making of a main expert to whom the decisions of local experts are forwarded. The main expert has a *prior* information about the reliabilities of the decisions which the local experts make. We denote the classwise reliability, $rel(k,u_k)$, $u_k \in \Omega_k$, k = 1, ---, p, as a reliability of a decision on u_k using the k^{th} data set (or by k^{th} expert). In the same way, REL(k), k = 1, ---, p, is denoted as the reliability of the k^{th} data set.

To find an optimal decision fusion rule based on the Bayesian minimum cost approach, a cost function is defined as follows. A cost $J(u_0, \omega_j)$ is given to the decision u_0, $u_0 \in \Omega_0$ when the true class is ω_j, $\omega_j \in \Omega_0$. Then, given a set of local decisions $\{u_1, ---, u_p\}$, an optimal fusion rule in the sense of minimum expected cost can be derived (Tang et *al.* 89) as in,

$$\text{choose } u_0 = u \in \Omega_0 \qquad (4.4.a)$$

$$\text{where, } u = \arg \min_{c \in \Omega_0} \sum_{\omega_j} J(u_0 = c, \omega_j)\, P\{u_1, --, u_p, \omega_j\}$$

Consider the "0-1" cost function given as,

$$J(u_0, \omega_j) \equiv [\, 1 - \delta(u_0, \omega_j)\,] \qquad (4.4.b)$$

$$\text{where, } \delta(x, y) = 1, \text{ if } x = y, \text{ and, } \delta(x, y) = 0, \text{ otherwise}$$

With this "0-1" cost function, an optimal fusion rule in eq. (4.4.a) will choose a class $u_0 = u \in \Omega_0$ having a maximum joint probability of $P\{u_1, ---, u_p, \omega_j = u\}$ which shows the likelihood of joint occurrence of $\{u_1, ---, u_p\}$ and $\{a, = u)$.

$$\text{choose } u_0 = u \in \Omega_0 \qquad (4.5.a)$$

$$\text{where, } u = \arg \max_{c \in \Omega_0} P\{u_1, --, u_p, \omega_j = c\}$$

In other words, this fusion rule will find a class $u_0 = u$ which is most likely to occur jointly with the local decisions $\{u_1, ---, u_p\}$. If the conditional independence of u_k's given u_0 is assumed as,

$$P\{u_k \mid u_{k-1}, ---, u_1, u_0\} = P\{u_k \mid u_0\} \qquad (4.5.b)$$

then, the joint probability $P\{u_1, ---, u_p, u_0\}$ is simplified as,

$$P\{u_1, ---, u_p, u_0\} = P(u_0) \prod_{k=1}^{p} P(u_k \mid u_0) \qquad (4.5.c)$$

For each data set, a set of conditional probabilities, $\{P(u_k \mid u_0) \mid u_k \in \Omega_0, u_0 \in \Omega_0\}$, is required. Note that this straightforward result cannot support a disparate

degree of data set reliabilities nor classwise reliabilities. This is because the cost function in eq. (4.4.b) is determined only on a basis of (u_0, ω_j).

Among the local decisions $\{u_1, ---, u_p\}$, some of the decisions could be more dependable in terms of data set and classwise reliabilities than others. In this case, it would be more desirable to have a final decision as consistent as possible with those reliable local decisions. This consistency, or consensus over the local decisions will also be as important as the maximum likelihood of joint occurrence which is pursued by eq. (4.5.a).

To acconimodate this idea, a slightly modified cost function is considered so that an optimum decision fusion algorithm selects a decision u_0 which is not only most likely to occur jointly with local decisions but also as consistent as possible to the *reliable* local decisions among $\{u_1, ---, u_p\}$.

Let's consider a new cost function which is dependent not only on (u_0, ω_j) but also on $(u_1, ---, u_p)$. This cost function will be designed to allow a final decision maximally consistent with the local decisions $(u_1, ---, u_p)$. The degree of consistency to each local decision will be based on the classwise and data set reliability. Specifically, a cost function $J(u_0, u_1, ---, u_p, \omega_j)$ in the following form is examined.

$$J(u_0, u_1, ---, u_p, \omega_j) = \sum_{k=1}^{p} J(u_0, u_k, \omega_j) \qquad (4.6)$$

$J(u_0, u_k, \omega_j)$, called a local cost function associated with k^{th} data set, determines a cost given to an action of selecting u_0 based on the k^{th} local decision u_k. A summed value of all the local costs is then, the actual cost assigned to the action of selecting u_0 based on $\{u_1, ---, u_p\}$ and $J(u_0, u_1, ---, u_p, \omega_j)$ is called a global cost function.

To determine the proper local cost function, consider assigning costs to the following five possible actions in Table 4.1.

Table 4.1 Cost Assignments to Courses of Actions.

Conditions	$J(u_0, u_k, \omega_j)$	cases
1. $u_k = \omega_j$ and $u_0 = u_k$	0	$u_0 = \omega_j$
2. $u_k = \omega_j$ and $u_0 \neq u_k$	1	$u_0 \neq \omega_j$
3. $u_k \neq \omega_j$ and $u_0 = u_k$	A'	$u_0 \neq \omega_j$
4. $u_k \neq \omega_j$,	B'	$u_0 = \omega_j$
5. $u_k \neq \omega_j$, $u_0 \neq u_k$ and $u_0 \neq \omega_j$	1	$u_0 \neq \omega_j$

$0 \leq A', B' \leq 1$; A', B' are not both 1

The idea in assigning cost values to the courses of action is to give lower cost to those desirable actions and higher costs to less desirable actions. As in case 1 in Table 4.1, if a decision u_0 matches the local decision u_k and if it is a correct decision (*i.e.*, $u_0 = \omega_j$), then, the lowest cost, which is selected as zero in this case, is assigned. On the other hand, if a decision u_0 doesn't match the local decision u_k and if the selected decision is also wrong (*i.e.*, $u_0 \neq \omega_j$), then the highest cost, which is selected as one, is assigned.

Since it is desirable for a fusion rule to choose a decision u_0 which is as consistent as possible to the local decisions $\{u_1, ---, u_p\}$, a cost A' which is not necessarily the largest cost of one is assigned to the case 3 in Table 4.1. Even if the decision u_0 is erroneous, the cost value A' can be less than the largest cost, since the decision of u_0 follows the decision of the k^{th} data set, u_k. Similarly, a cost value B' which can be larger than the smallest cost value of zero is given to case 4 since the decision u_0 doesn't follow u_k, even if the decision of u_0 may be true. When u_0 doesn't follow u_k and u_k is not correct either, the largest cost value is assigned.

If the costs A' and B' are not both 1, then the cost assignments in Table 4.1 can be expressed in terms of two separate components, one being a function of (u_0, ω_j) and the other, a function of (u_0, u_k), in two different ways as,

$$J(u_0, u_k, \omega_j) = K_1 J(u_0, u_k) \cdot J(u_0, \omega_j) + (1 - K_1) \qquad (4.7.a)$$

$$J(u_0, u_k, \omega_j) = K_2 J(u_0, u_k) + (1 - K_2) J(u_0, \omega_j) \qquad (4.7.b)$$

K_1 and K_2 are constants which are independent of the class decisions. The cost function component $J(u_0, u_k)$ imposes a consistent relationship with the local decision u_k on the decision u_0. On the other hand, the component $J(u_0, q)$ imposes a constraint of maximum likelihood of co-occurrence as in eq. (4.4.b). These two cost function components can be expressed as,

$$J(u_0, u_k) \equiv 1 - A\,\delta(u_0, u_k), \quad 0 \le A \le 1 \qquad (4.8.a)$$

$$J(u_0, \omega_j) \equiv 1 - B\,\delta(u_0, \omega_j), \quad 0 \le B \le 1 \qquad (4.8.b)$$

where, A and B are not both 0.

With a cost function in a form of eq. (4.7.a), the parameters A, B, and K_1 are related to the costs A' and B' in Table 4.1 as,

$$A' = 1 - K_1 A \text{ and } B' = 1 - K_1 B \qquad (4.9.a)$$

where, $K_1 = 1 / [1 - (1-A)(1-B)]$

Since constant K_1 in eq. (4.7.a) does not affect a selection of u_0, without loss of generality, the cost function in eq. (4.7.a) can be redefined as,

$$J(u_0, u_k, \omega_j) = J(u_0, \omega_j) \cdot J(u_0, u_k) \qquad (4.9.b)$$

In the case of a cost function in an additive form in eq. (4.7.b), the constant K_2 is not related to the value of A and B, and it can be set arbitrarily since it doesn't affect the global decision. For simplicity, it is set to 1/2. The relationship between A, B and A', B' is,

$$A' = 1 - A \text{ and } B' = 1 - B \qquad (4.10)$$

With appropriate values of A and B (or equivalently, A' and B'), it is possible to control the relative importance between selecting a decision maintaining maximal consistency with the local decisions $\{u_1, ---, u_p\}$ and selecting a class of highest joint occurrence likelihood with the local decisions. The cost function defined in eq. (4.6) with eq. (4.7.a,b) is quite general in its scope of application since it can

define various cases of cost function by choosing different values of A and B. For example, the cost function $J(u_0, \omega_j)$ in eq. (4.4.b) which is based on maximum likelihood fusion, is achieved with A' = 1 and B' = 0. If parameters of A' = 0 and B' = 1 are used, the resulting fusion rule will be selecting a majority decision among the local decisions.

Employing this new cost function in eq. (4.6), an expected cost of choosing u_0 given $\{u_1, ---, u_p\}$ is computed as,

$$\text{Expected Cost} = E\{J(u_0, u_1, ---, u_p, \omega_j)\}$$

$$= \sum_{u_1, --, Up} \sum_{\omega_j \in \Omega_0} J(u_0, u_1, ---, u_p, \omega_j)\, P(u_1, ---, Up, \omega_j)$$

Define the inner summation term in the above equation as a function H, as,

$$H(u_0 \mid u_1, ---, u_p) \equiv \sum_{\omega_j} J(u_0, u_1, ---, u_p, \omega_j)\, P\{u_1, ---, u_p, \omega_j\}$$

An optimum decision which minimizes the expected cost can be found by minimizing H with respect to $u_0 \in \Omega_0$. Note that a choice of A in the cost function in (4.8.a) controls the relative importance of consistency between u_0 and local decision u_k, therefore it should be dependent on the particular data set employed and a local decision u_k, according to the data set and classwise reliabilities. Thus, the notation "$A_k(u_k)$" would be more appropriate to explicitly show the dependence of "A" on the particular data set and the local decision u_k. That is, according to the data set and classwise reliabilities, REL(k)'s and rel(k,u_k)'s, appropriate values of $A_k(u_k)$'s can be determined in such a way that a less reliable local decision has less effect on making a final decision u_0 through a selected fusion rule. Substituting the cost function in eq. (4.9.b) into H results in,

$$H(u_0 \mid u_1, ---, u_p) = \left[\sum_{k=1}^{p} J(u_0, u_k) \right] \left[\sum_{\omega_j} J(u_0, \omega_j)\, P(u_1, ---, u_p, \omega_j) \right] \quad (4.11)$$

The first term in eq. (4.11) which is rewritten as in eq. (4.12.a) accounts for a consistency constraint between the local decisions $\{u_1, ---, u_p\}$ and u_0.

$$p - \sum_{k=1}^{p} A_k(u_k)\, \delta(u_0, u_k) \qquad (4.12.a)$$

To understand the role of this term a bit more clearly, suppose weight factors, $A_k(\cdot)$'s are all the same. Then the cost function with only this term would choose the class u_0 which is a majority class among u_k's. Therefore it is a majority rule. With the distinct reliabilities associated with u_k's, the "vote" of each local decision is weighted according to $A_k(\cdot)$'s. Then the fusion rule in eq. (4.12.a) will select a class u_0 attaining most of weights. For this reason, this fusion rule will be called a "weighted majority decision fusion rule." On the other hand, the second component of eq. (4.11), which may be re-written as,

$$P\{u_1, ---, u_p\} \left[1 - B\, P\{u_0 \mid u_1, ---, u_p\} \right] \qquad (4.12.b)$$

is dependent only on u_0 and ω_j. If the cost function $J(u_0, \omega_j)$ was employed for itself alone (that is, if all $A_k(u_k)$'s are zero), it would choose u_0 based on a relative likelihood of the joint occurrence of $\{u_1, ---, u_p, u_0\}$ as in eq. (4.5.a). The relative magnitude of the $A_k(u_k)$'s and B will determine the actual degree of balance between emphasizing the importance of the term in eq. (4.12.a) and lhat in eq. (4.12.b) in deciding a class u_0 under the eq. (4.11).

In the same manner, if the cost function in eq. (4.7.b) is employed, then the following H is obtained.

$$H(u_0 \mid u_1, ---, u_p) =$$

$$\frac{P(u_1, ---, u_p)}{2} \sum_{k=1}^{p} J(u_0, u_k) + \frac{1}{2} \sum_{\omega_j} P(u_1, ---, u_p, \omega_j) \sum_{k=1}^{p} J(u_0, \omega_j) \qquad (4.13)$$

As before, the first term in eq. (4.13), which is rewritten as,

$$P(u_1, ---, u_p)\left[\frac{p}{2} - \sum_{k=1}^{p} A_k(u_k)\,\delta(u_0, u_k)\right] \qquad (4.14.a)$$

is related to emphasizing the consistency of a global decision with the local decisions. The second term in eq. (4.13) which is simplified as,

$$P(u_1, ---, u_p)\left[\frac{p}{2} - B_{ML}\,P(u_0 \mid u_1, ---, u_p)\right] \qquad (4.14.b)$$

$$\text{where,} \quad B_{ML} = \sum_{k=1}^{p} 1 - A_k(u_k)$$

is based on the maximum likelihood decision fusion and this is equivalent to eq. (4.5.a). The parameter B_{ML} accounts for the total weight given to the maximum likelihood based fusion. The posterior probability $P(u_0|u_1,---, u_p)$ can be computed as in eq. (4.5.c).

4.5 Data Set and Classwise Reliability

In eq. (4.12.a and 4.14.a), the data set and classwise reliability factors are reflected in the $A_k(u_k)$'s. It would be very logical to assign a large cost to the case when the fusion rule fails to follow a local decision which has high reliability. In the report in (Benediktsson et *al.* 90), several different measures of data set reliability were introduced. Statistical separabilities between classes are a possible candidate for assessing data set reliability. The computation involved in evaluating separabilities could be non-trivial if the multivariate normality assumption about the data set cannot not be satisfied. Furthermore, in the case of a data set where the data values are not changing enough, e.g., in digital elevation data, the covariance matrix may be ill-conditioned. Another measure of reliability based on equivocation is introduced in (Benediktsson et *al.* 90), and in this approach, the data set reliability is related to the degree that the data classes indicate specific information classes. If the data classes in one data set strongly indicate the corresponding information classes, then this data set is considered to

be reliable. Since the purpose of multisource data analysis in this chapter lies in a classification, classification accuracy could be a logical choice for the reliability measure. Any data set which has higher classification accuracy may be assumed more reliable than the others. Note that classification accuracy can be easily obtained irrespective of assumptions about underlining probability density functions. Data set reliability REL(K) can be determined similarly to (Benediktsson *et al.* 90) based on these criteria.

However, these measures are not directly applicable to the classwise reliability which is a measure of reliability of a particular local decision selected based on a given (local) data source. Two different measures based on classification accuracy can be examined as follows. For $u_k \in \Omega_K$ and $\omega_j \in \Omega_0$, $k = 1, \cdots, p$, and $j = 1, \cdots M_0$,

$$\text{rel}(k, u_k = \omega_j) = P(u_k = \omega_j \mid x_k = \omega_j) \tag{4.15}$$

Equation (4.15) is the probability of correctly classifying x_k as belonging to a class ω_j, and it is the detection probability of class ω_j. Any class with high classification accuracy should be associated with large classwise reliability. However, there can be a problem in using this measure as manifested in following hypothetical example. Suppose a local classifier is very poorly designed or, feature vectors of a certain data set are of very bad quality, and it assigns a particular class to all pixels. In this case, the measure of eq. (4.15) will assign the highest reliability of 1 to that particular class, although the decision to this class is meaningless.

$$\text{rel}(k, u_k = \omega_j) = P(x_k = \omega_j \mid u_k = \omega_j) \tag{4.16}$$

On the other hand, the measure in (4.16) doesn't have this problem, and it is the probability that a pixel x_k is truly from the same class as the local decision u_k. Since this reliability measure is one minus the probability that the local decision is incorrect, if the probability of eq. (4.16) is high, then, statistically speaking, the knowledgo of a local decision u_k will be able to indicate the class of x_k with a high probability. These classwise reliabilities can be estimated from the classification results of representative training samples.

$$A_k(u_k) = REL(K) \cdot rel(k,u_k) \qquad (4.17)$$

There still remains a problem in associating the data set and classwise reliability measures to actual values of weights $A_k(\cdot)$'s. Since it appears difficult to do optimally, at least for now, the seemingly simple way of eq. (4.17) is used.

4.6 Information Combination Structures in Multisource and Temporal Contextual Classification

The multisource classifiers discussed so far can be straightforwardly used for temporal contextual classification. One difference between these two applications may be distinguishing the order of data sets in temporal classifiers. Generally, there can be two different structures in combining multiple data sets as shown in Fig. 4.2.

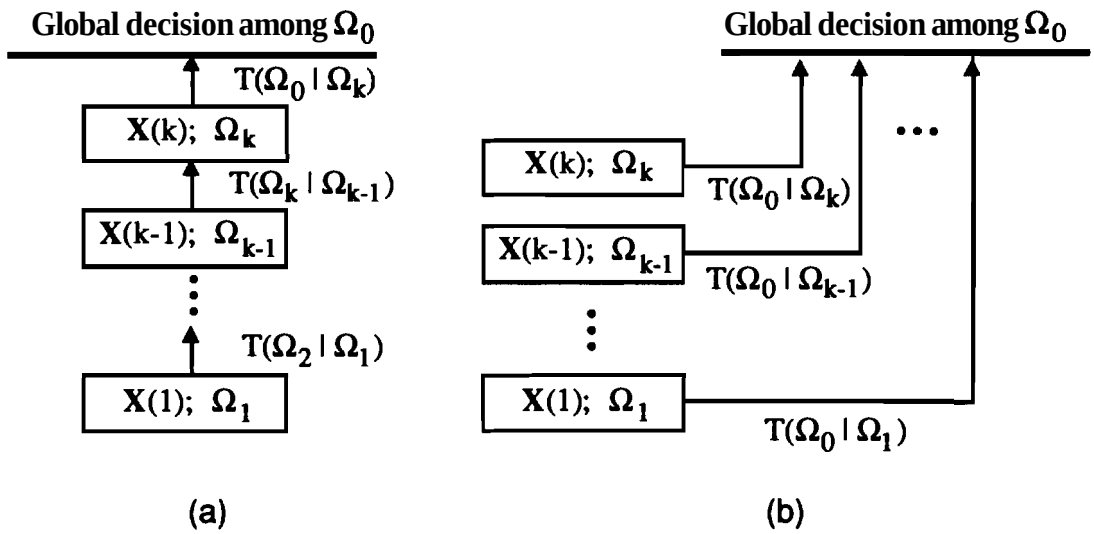


Figure 4.2 Information Combination Structures. (a) Serial Structure. (b) Parallel Structure.

The serial structure of Fig. 4.2.(a) may best fit the temporal contextual classification since the temporal information up to t = k can be conveyed as temporal information to the next classification process of the $(k+1)^{th}$ data set. As a new temporal data set becomes available, likelihood values are updated to be

used with next temporal data set. An intermediate classification result can be obtained at each temporal stage. In this serial structure, the different order of data sets;, which is mostly likely to be chronological, generally give, different results.

The parallel structure shown in Fig. 4.2.(b) is what the rnultisource classifier in eq. (4.1) is based on, and there is no distinction in the order of data sets, since information from each data set is independently fed into the (global) classifier which makes decisions among Ω_0. There are no intermediate classification results. This structure is more straightforward to accommodate data set reliabilities than the serial one. The decision fusion algorithms previously discussed are based on this structure. Note that this parallel combination structure is based on an assumption of class-conditional independence! between data sets. Note that since the class transition matrix to Ω_0 is required only at the final temporal data set in the serial structure, selecting a different class set, Ω_0, affects only the last temporal stage. However, in the parallel structure, this flexibility cannot be attained.

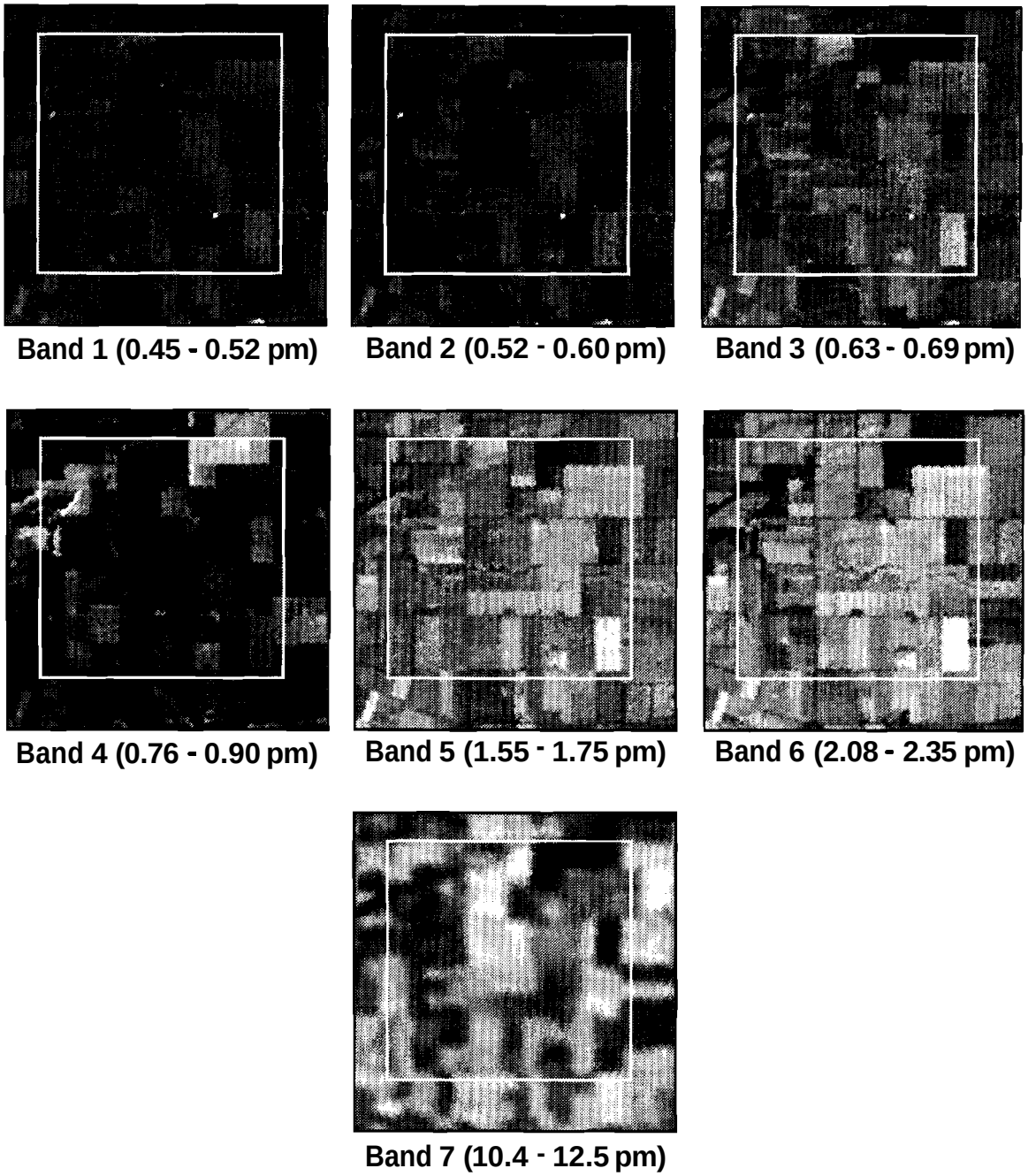
4.7 Experiments and Discussion on Temporal Contextual Classification

4.7.1 Description of Experiment

To test the multisource (or, temporal contextual) classification algorithms discussecl in this chapter, three Landsat Thematic Mapper (TM) data :sets were used. In additional to the July and September data sets which were introduced in previous chapter, Thematic Mapper data acquired in April was used for temporal classification among the same four information classes {corn, soybeans, wheat, alfalfa/oats} as in previous chapter.

In the April data set, there was not much difference between these four information classes except wheat. In fact, the information class "wheat" was the only green crop type which could be observed in the given agricultural fields at that time, and the information class, wheat, was identifiable from the others with high accuracy using this April image. The April data set is shown in Fig. 4.3, in which the darker regions in band 6 matches well with the location of the class wheat in the truth map shown in Fig. 3.8. Note that in band 6, green vegetation

has relatively lower spectral reflectance than soil (Swain 78b). Thus, the regions corresponding to green vegetation would look darker than those corresponding to soils.



Band 1 (0.45 - 0.52 pm) Band 2 (0.52 - 0.60 pm) Band 3 (0.63 - 0.69 pm)

Band 4 (0.76 - 0.90 pm) Band 5 (1.55 - 1.75 pm) Band 6 (2.08 - 2.35 pm)

Band 7 (10.4 - 12.5 pm)

Figure 4.3 April Thematic Mapper (TM) Data Set.; The white box shows the 128 by 128 portion of selected test field.

Only two information classes (wheat and "others") were defined in the April data set, and several sub-classes of these two information classes were defined to meet the multivariate normality assumption. The number of samples in the April data selected for training and test are summarized in Table 4.2.

Table 4.2. Training and Test Samples of April Thematic Mapper Data.

Information Class	April Data		
	Subclasses	number of Training Samples	Test Samples
Others	2	960	11698
Wheat	1	376	1215
Total	3	1336	12998

Experiments were carried out with the several classifiers discussed in this chapter. Final (global) decisions were made among the user-defined information classes of Ω_0 = {corn, soybean, wheat, alfalfa/oat} and classification performances were compared in terms of the overall classification accuracy (OVA) and the class-averaged classification accuracy (CAG).

Applying fusion rules requires a class transition matrix in eq. (4.3.a). In the experiments, class transition probabilities were selected heuristically in such a way that a transition between the same information class had a higher probability than other cases. To implement this idea, following relationship was used. For $u_k \in \Omega_k$, $\omega_j \in \Omega_0$, k = 1, ---, p,

$$P\{u_k \mid \omega_j\} = P_0 \,/\, n, \quad \text{if } u_k \text{ and } \omega_j \text{ belong to the same information class}$$
$$P\{u_k \mid \omega_j\} = (1 - P_0) \,/\, (M_k - n), \quad \text{otherwise} \tag{4.18}$$

n is the number of total sub-classes of the information class to which sub-class u_k belongs. P_0 is a user defined number between zero and one. P_0 being one means no allowance of class transition to another information class. If P_0 is zero, class transition is permitted only to different information classes. Several different values of P_0 were tested as in Table 4.3 and the values of P_0 which gave the best performance was chosen for comparison with other classifiers. (P_0 = 0.99 for

class corn and soybean, $P_0 = 0.8$ for class wheat, $P_0 = 1$ for class alfalfa/oats). In the case of classifying July data with April data or September data with April data, P_0 was set to one.

Table 4.3 Bi-temporal Classification of July Data with September Data with Different Class Transition Probabilities. (Equal Data set reliability).

(%)

P_0 in eq.(4.18)	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
0.80	91.02	59.42	66.34	78.55	73.83	75.20
0.99	92.03	60.28	64.12	78.92	73.84	75.78
1.00	89.84	60.03	61.40	79.43	72.68	74.55
Best[1]	91.80	63.08	69.88	73.35	74.53	76.67

[1] $P_0 = 0.99$ for corn and soybeans. $P_0 = 0.8$ for wheat, $P_0 = 1$ for alfalfa/oats

4.7.2 Temporal Classification with Data Fusion

The multisource classifier based on data fusion in eq. (4.1) was applied to the classification of July data with April and September data, and classification results are shown in Table 4.4 ~ 4.6. Since the ground truth was gathered in July and it matches best with July data, all comparisons were made with respect to the July data set. Non-contextual maximum likelihood classification results of each temporal data set separately are also included in the tables. In the classification of single data sets, the July data set gave better classification performance than the September data set for all classes. But some of the classes (soybeans, wheat in July data or soybeans, wheat and alfalfa/oats in September data) had very poor classification accuracy. Note that the class, wheat and others in the April data set were discriminated very successfully from each other.

Several different data set reliability factors were tested to see their effect on classification accuracies. As seen in the Table 4.4 ~ 4.6, temporal contextual classification based on data fusion with eq. (4.1) generally gave better results than any of the single pixelwise maximum likelihood classification.

Inclusion of April data improved the classification accuracy of wheat and alfalfa./oats significantly. Although this improvement couldn't increase the overall classification accuracy (OVA) much due to a relatively small portion of sample

numbers belonging to those classes, the class-averaged classification accuracies (CAG) were increased by as much as 5% (July data) or 14% (September data). The September data set was helpful in classifying the class soybeans in July data as seen in Table 4.5, but there was a slight degradation in classification accuracy for the class, alfalfa/oats. The classification accuracies in September data were generally very low except corn and the improvement due to including September data in classifying July data was not significant.

Table 4.4 Classification Accuracy Comparison of the Statistical Multisource Classifier with Different Data Set Reliabilities (Classification of July and September Data with April Data).

		Percent Classification Accuracy					
Data Set Weights		Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
Separate Maximum Likelihood Classification of Each Data Set							
April		89.59[1]		90.29		89.94	89.65
July		90.18	57.72	68.72	77.89	73.63	74.37
September		82.59	55.06	51.28	47.07	59.00	65.28
April	July	Classification of July Data with April Data					
0.6	1	90.30	56.69	85.27	83.02	78.82	76.13
0.7	1	90.29	56.44	85.51	83.24	78.87	76.08
0.8	1	90.23	56.42	86.26	83.31	79.05	76.13
0.9	1	90.29	56.42	86.50	83.16	79.09	76.16
1	1	90.32	56.42	86.58	82.94	79.07	76.16
1	0.9	90.27	56.44	86.91	82.80	79.11	76.16
1	0.8	90.23	56.36	87.41	82.72	79.18	76.15
1	0.7	90.12	56.30	87.49	82.72	79.16	76.09
1	0.6	90.12	56.25	87.49	82.65	79.13	76.07
April	Sept.	Classification of September Data with April Data					
0.6	1	82.64	56.27	87.65	57.61	71.05	70.26
0.7	1	82.59	56.27	88.15	59.88	71.72	70.52
0.8	1	82.53	56.23	88.89	61.93	72.40	70.76
0.9	1	82.51	56.21	89.47	62.96	72.79	70.91
1	1	82.50	56.19	89.63	64.35	73.17	71.06
1	0.9	82.44	56.15	89.79	65.45	73.46	71.15
1	0.8	82.41	56.15	89.71	66.25	73.63	71.21
1	0.7	82.30	56.15	89.63	66.54	73.66	71.19
1	0.6	82.19	56.13	89.79	66.84	73.74	71.18

[1] In classifying April data with a maximum likelihood classifier, there were only 2 information classes {wheat, others}. Classification accuracy of "others" is given under corn. (see Table 4.2).

Table 4.5 Classification Accuracy comparison of the Statistical Multisource Classifier with Different Data Set Reliabilities (Classification of July Data with September Data and vice versa).

		Percent Classification Accuracy					
Data Set Weights		Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
Separate Maximum Likelihood Classification of Each Data Set							
July		90.18	57.72	68.72	77.89	73.63	74.37
September		82.59	55.06	51.28	47.07	59.00	65.28
Sept.	July	Classification of July Data with September Data					
0.6	1	91.31	61.70	69.79	75.92	74.68	76.21
0.7	1	91.53	61.95	69.71	75.26	74.61	76.32
0.8	1	91.65	62.37	69.47	75.11	74.65	76.49
0.9	1	91.71	62.81	69.63	74.67	74.70	76.64
1	1	91.80	63.08	69.88	73.35	74.53	76.67
1	0.9	91.87	63.40	69.71	72.40	74.35	76.70
1	0.8	91.99	63.86	69.47	71.01	74.08	76.75
1	0.7	92.08	64.42	69.14	69.11	73.69	76.77
1	0.6	92.30	64.63	69.22	67.72	73.47	76.80

Table 4.6 Classification Accuracy Comparison of the Statistical **Multisource** Classifier with Different Data Set **Reliabilities** (Classification **of** July and September Data with April Data).

Data Set Weights			Percent Classification Accuracy					
			Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
Separate Maximum Likelihood Classification of Each Data Set								
April			89.59[1]		90.29		89.94	89.65
July			90.18	57.72	68.72	77.89	73.63	74.37
Sepltember			82.59	55.06	51.28	47.07	59.00	65.28
April	Sept.	July	Classification of July Data with September and April Data					
0.60	0.60	1.00	91.53	61.60	86.01	81.41	80.13	78.36
0.70	0.70	1.00	91.78	61.95	86.75	81.41	80.47	78.67
0.80	0.80	1.00	91.92	62.54	86.83	81.48	80.69	78.96
0.90	0.90	1.00	91.99	63.19	86.91	81.63	80.93	79.25
1.00	1.00	1.00	92.08	63.63	87.16	81.63	81.12	79.47
0.90	1.00	0.90	92.14	64.05	87.41	81.41	81.25	79.65
0.80	1.00	0.80	92.26	64.61	87.57	80.89	81.34	79.87
0.70	1.00	0.70	92.41	65.24	87.74	80.16	81.39	80.10
0.60	1.00	0.60	92.52	65.56	88.07	79.50	81.41	80.23
1.00	0.90	0.90	92.01	63.61	87.33	81.70	81.16	79.46
1.00	0.80	0.80	91.98	63.63	87.65	82.06	81.33	79.52
1.00	0.70	0.70	91.96	63.67	87.90	81.63	81.29	79.50
1.00	0.60	0.60	91.87	63.61	87.98	81.77	81.31	79.47

[1] In classifying April data with a maximum likelihood classifier, there were only **2 information** classes {wheat, others). Classification accuracy of "others" is given under corn. (see Table **4.2).**

When all three temporal data sets were used all together, classification results were much improved for all **4** information classes as seen in Table **4.6.** Especially the classes, soybeans and wheat had major classification accuracy improvements. Notice that improvement for both the class wheat in Table **4.4** and for the class soybeans in Table **4.5** were achieved in the results in Table **4.6.** Classification error maps of the best multisource classification results in Table 4.4 **~ 4.6** are shown in Fig. **4.4.**

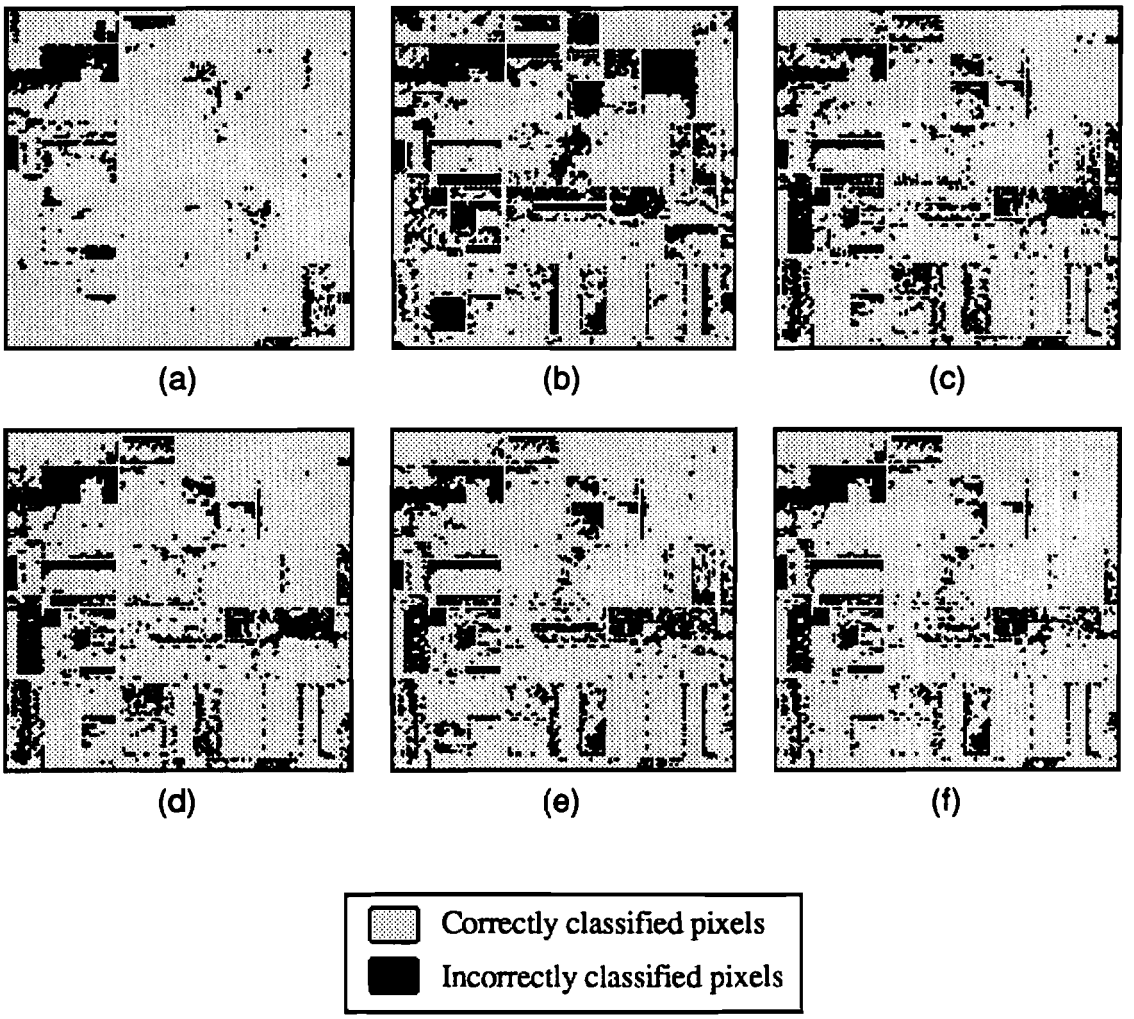


Figure 4.4 **Classification Error Maps of Multisource Classifier. (a) April Data set. (b) September Data set. (c) July Data set. (d) July Data with April Data. (e) July Data with September Data. (f) July Data with September and April Data.**

4.7.3 Temporal Classification with Decision Fusion

The multisource classifiers based on decision fusion were applied in classifying July data with April and September data sets. The maximum likelihood decision fusion rule in eq. (4.5.a) requires prior knowledge of the joint probability, $P\{u_1, ---, u_p, u_0\}$. Under the conditional independence assumption of eq. (4.5.b), the amount of prior information required can be reduced substantially by using the relationship in eq. (4.5.c). If there are M_k classes in the k^{th} data set and M_0 different u_0's, then, M_k times M_0 conditional probabilities of $P\{u_k \mid u_0\}$ are required. This a priori information would not be necessary in data fusion-based multisource classification since information is combined in terms of posterior probabilities. In the decision fusion-based approaches where information is combined in a level of decisions, only limited information (that is, that of decisions) is transferred for global decision making. However, the additional a priori information about conditional probabilities of $P\{u_k \mid u_0\}$'s provides information needed in making a global decision. As for the weighted majority fusion rule in eq. (4.12.a), note that only M_k different classwise reliability factors are sufficient for decision fusion.

Table 4.7 shows classification accuracy comparisons between the data fusion-based algorithm with eq. (4.1) and the maximum likelihood decision fusion scheme in eq. (4.5.a). The best results in terms of overall classification accuracy (OVA) in Table 4.4 ~ 4.6 are also included in Table 4.7 for easy comparison.

The maximum likelihood decision fusion rule in eq. (4.5.a) compares very favorably to the data fusion based multisource classifier in eq. (4.1), although only limited information of local class decisions were combined. A priori information about conditional probabilities, $P(u_k|u_0)$'s were found to be very effective in combining information for classification. Especially in classifying July data with September data, the maximum likelihood decision fusion rule resulted in about 5.4% of overall classification accuracy increase over the best data fusion multisource classification result in Table 4.5. The classification performance increase was significant for the classes corn and soybeans. Compared to the maximum likelihood classification of July data only (that is, non-contextual), this amounts to a 7.9% of increase.

Table 4.7 Classification Accuracy Comparison of the Maximum Likelihood Decision Fusion.

	Percent Classification Accuracy					
Data Set Weights	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
Separate Maximum Likelihood Classification of Each Data Set						
April	89.59[1]		90.29		89.94	89.65
September	82.59	55.06	51.28	47.07	59.00	65.28
July	90.18	57.72	68.72	77.89	73.63	74.37
Data Fusion Based Classifier in eq. (4.1)[2]						
JUL+APR	90.29	56.42	86.50	83.16	79.09	76.16
JUL+SEP	92.30	64.63	69.22	67.72	73.47	76.80
JUL+APR+SEP	92.52	65.56	88.07	79.50	81.41	80.23
Maximum Likelihood Decision Fusion Rule in eq. (4.5.a)						
JUL+APR	90.18	57.72	89.96	80.82	79.67	76.67
JUL+SEP	94.19	75.63	68.72	73.79	78.08	82.24
JUL+APR+SEP	95.79	77.08	88.89	71.52	83.32	85.10

[1] In classifying April data with a maximum likelihood classifier, there were only 2 information classes {wheat, others). Classification accuracy of "others" is given under corn. (see Table 4.2).

[2] These are the best results in Table 4.4 ~ 4.6.

Classification error maps of this maximum likelihood decision fusion rule are shown in Fig. 4.5. Combining local decisions of September data to those of July were very effective as seen in Fig. 4.5 where many corn and soybeans pixels were correctly classified as in Fig. 4.5.(c). The April data set was effective in improving classification accuracy of wheat. Notice that both of the improvements in Fig. 4.5.(b) and (c) are visible in Fig. 4.5.(d), which shows the error rnap when all three data sets are used.

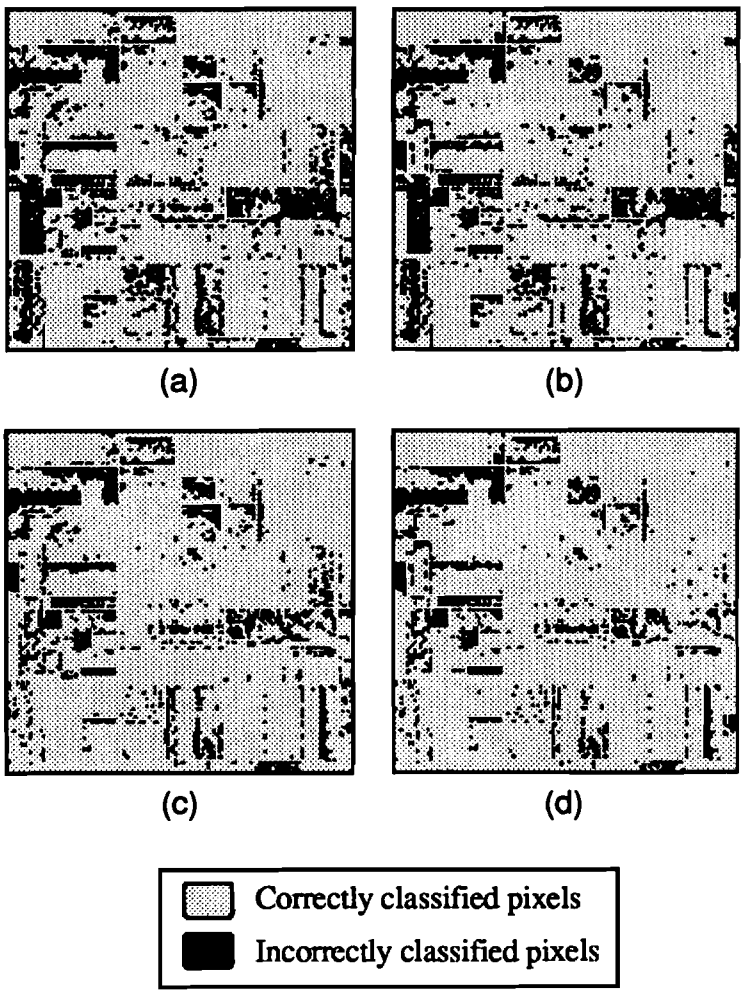


Figure 4.5 **Classification Error Maps of July Data with the Maxiimum Likelihood Decision Fusion. (a) July Data only (non-temporal). (b) July Data with April Data. (c) July Data with September Data. (d) July Data with April and September Data.**

Classification results with the weighted majority decision fusion are presented in Table 4.8 ~ 4.10. Under the weighted majority decision fusion, different data set reliability factors can be assigned to each data set. In the experiment, several different data set reliabilities were tested as seen in the tables. Both of the two different classwise reliabilities in eq. (4.15) and (4.16) were tested to see their effectiveness. Note that the weighted majority decision fusion requires much less prior information than the maximum likelihood decision fusion. For comparison purposes, the results of data fusion and maximum likelihood decision fusion are also included in Table 4.8 ~ 4.10.

Table 4.8 Classification Accuracy Comparison for Weighted Majority Decision Fusion (Classification of July Data with April Data).

		Percent Classification Accuracy					
Data Set Weights		Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
see below[1]		90.29	56.42	86.50	83.16	79.09	76.16
see below[2]		90.18	57.72	89.96	80.82	79.67	76.67
April	July	with Classwise reliability in eq. (4.15)					
0.60	1.00	90.18	57.72	68.72	77.89	73.63	74.37
0.70	1.00	90.18	53.55	69.55	77.89	72.79	72.92
0.80	1.00	92.61	53.55	68.64	77.89	73.17	73.87
0.90	1.00	92.61	53.55	90.78	70.86	76.95	75.20
1.00	1.00	92.61	53.55	90.78	70.86	76.95	75.20
1.00	0.90	86.24	53.55	90.86	70.86	75.38	72.49
1.00	0.80	86.24	53.55	90.86	70.86	75.38	72.49
1.00	0.70	86.24	53.55	90.86	70.86	75.38	72.49
1.00	0.60	86.24	53.55	90.86	70.86	75.38	72.49
April	July	with Classwise reliability in eq. (4.16)					
0.60	1.00	92.61	57.72	69.71	76.50	74.13	75.36
0.70	1.00	92.61	57.72	69.71	76.50	74.13	75.36
0.80	1.00	92.61	57.72	69.71	76.50	74.13	75.36
0.90	1.00	92.61	57.72	69.71	76.50	74.13	75.36
1.00	1.00	92.61	57.72	84.53	73.50	77.09	76.43
1.00	0.90	92.61	57.72	84.53	73.50	77.09	76.43
1.00	0.80	92.61	57.72	84.53	73.50	77.09	76.43
1.00	0.70	92.61	56.53	84.53	73.50	76.79	75.99
1.00	0.60	92.61	56.53	84.53	73.50	76.79	75.99

[1] With data fusion-based multisource classifier of eq. (4.1) and data set reliability April = 0.9, July = 1.

[2] With Maximum Likelihood Fusion Rule in eq. (4.5.a).

Table 4.9 Classification Accuracy Comparison for Weighted Majority Decision Fusion (Classification of July Data with September Data).

		Percent Classification Accuracy					
Data Set Weights		Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
see below[1]		92.30	64.63	69.22	67.72	73.47	76.80
see below[2]		94.19	75.63	68.72	73.79	78.08	82.24
Sept.	July	with Classwise reliability in eq. (4.15)					
0.60	1.00	90.18	57.72	68.72	77.89	73.63	74.37
0.70	1.00	94.21	56.46	68.72	77.89	74.32	75.63
0.80	1.00	94.21	56.46	68.72	77.89	74.32	75.63
0.90	1.00	96.11	56.46	64.86	77.89	73.83	76.09
1.00	1.00	96.83	56.46	64.86	70.64	72.20	75.63
1.00	0.90	96.83	56.46	64.86	70.64	72.20	75.63
1.00	0.80	96.83	38.24	61.81	78.40	68.82	69.47
1.00	0.70	96.83	57.18	61.81	74.30	72.53	76.00
1.00	0.60	96.82	57.41	50.21	46.78	62.80	72.10
Sept.	July	with Classwise reliability in eq. (4.16)					
0.60	1.00	92.80	78.11	67.00	57.69	73.90	80.70
0.70	1.00	96.82	77.23	67.00	57.69	74.68	82.09
0.80	1.00	96.82	77.23	67.00	55.20	74.06	81.83
0.90	1.00	96.82	77.23	67.00	51.46	73.13	81.44
1.00	1.00	96.83	76.89	64.44	51.46	72.41	81.08
1.00	0.90	96.82	77.08	62.63	51.46	72.00	80.97
1.00	0.80	96.82	77.08	62.63	51.46	72.00	80.97
1.00	0.70	96.82	77.08	67.90	44.58	71.59	80.74
1.00	0.60	96.82	77.08	67.90	44.58	71.59	80.74

[1] With data fusion-based multisource classifier of eq. (4.1) and data set reliability Sept. = 1, July = 0.6.

[2] With Maximum Likelihood Fusion Rule in eq. (4.5.a).

Table 4.10 Classification Accuracy Comparison for Weighted Majority Decision Fusion (Classification of July Data with April and September Data).

Data Set Weights			Percent Classification Accuracy					
			Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
see below[1]			92.52	65.56	88.07	79.50	81.41	80.23
see below[2]			95.79	77.08	88.89	71.52	83.32	85.10
April	Sept.	July	with Classwise reliability in eq. (4.15)					
0.60	0.60	1.00	91.78	58.62	78.44	76.43	76.32	76.14
0.70	0.70	1.00	94.93	54.70	79.09	76.43	76.29	76.11
0.80	0.80	1.00	94.93	54.70	79.01	76.43	76.27	76.10
0.90	0.90	1.00	94.93	54.70	81.98	74.96	76.64	76.23
1.00	1.00	1.00	95.59	54.70	81.98	68.16	75.11	75.80
0.90	1.00	0.90	94.59	54.70	81.98	68.16	74.85	75.37
0.80	1.00	0.80	94.59	54.70	81.98	68.16	74.85	75.37
0.70	1.00	0.70	94.93	53.95	90.86	65.74	76.37	75.81
0.60	1.00	0.60	94.93	53.95	90.86	65.74	76.37	75.81
1.00	0.90	0.90	94.93	54.70	81.98	74.96	76.64	76.23
1.00	0.80	0.80	94.93	54.70	79.01	76.43	76.27	76.10
1.00	0.70	0.70	94.93	54.70	79.09	76.43	76.29	76.11
1.00	0.60	0.60	91.78	58.62	78.44	76.43	76.32	76.14
April	Sept.	July	with Classwise reliability in eq. (4.16)					
0.60	0.60	1.00	92.50	78.11	74.49	72.04	79.28	82.77
0.70	0.70	1.00	96.85	77.23	74.07	72.04	80.05	84.27
0.80	0.80	1.00	96.85	77.23	73.99	70.35	79.61	84.09
0.90	0.90	1.00	96.85	77.10	73.42	66.62	78.50	83.60
1.00	1.00	1.00	97.18	76.77	75.23	64.86	78.51	83.60
0.90	1.00	0.90	96.87	76.77	77.28	64.86	78.94	83.66
0.80	1.00	0.80	96.87	76.35	78.11	64.86	79.05	83.58
0.70	1.00	0.70	96.87	76.24	78.11	64.86	79.02	83.54
0.60	1.00	0.60	96.55	76.24	87.49	62.45	80.68	84.03
1.00	0.90	0.90	96.85	77.10	73.42	66.62	78.50	83.60
1.00	0.80	0.80	96.85	77.23	73.99	70.35	79.61	84.09
1.00	0.70	0.70	96.85	77.23	74.07	72.04	80.05	84.27
1.00	0.60	0.60	92.50	78.11	74.49	72.04	79.28	82.77

[1] With the data fusion-based multisource classifier of eq. (4.1) and data set reliability April=July=0.6, Sept. = 1.

[2] With Maximum Likelihood Fusion Rule in eq. (4.5.a).

The classwise reliabilities in eq. (4.16) were observed far better in performance than those in eq. (4.15). This can be easily understood since the classwise reliability in eq.(4.16) indicates more directly the possibility of a local decision being true. Figure 4.6 shows locations of classification occurrences; with the weighted majority rule.

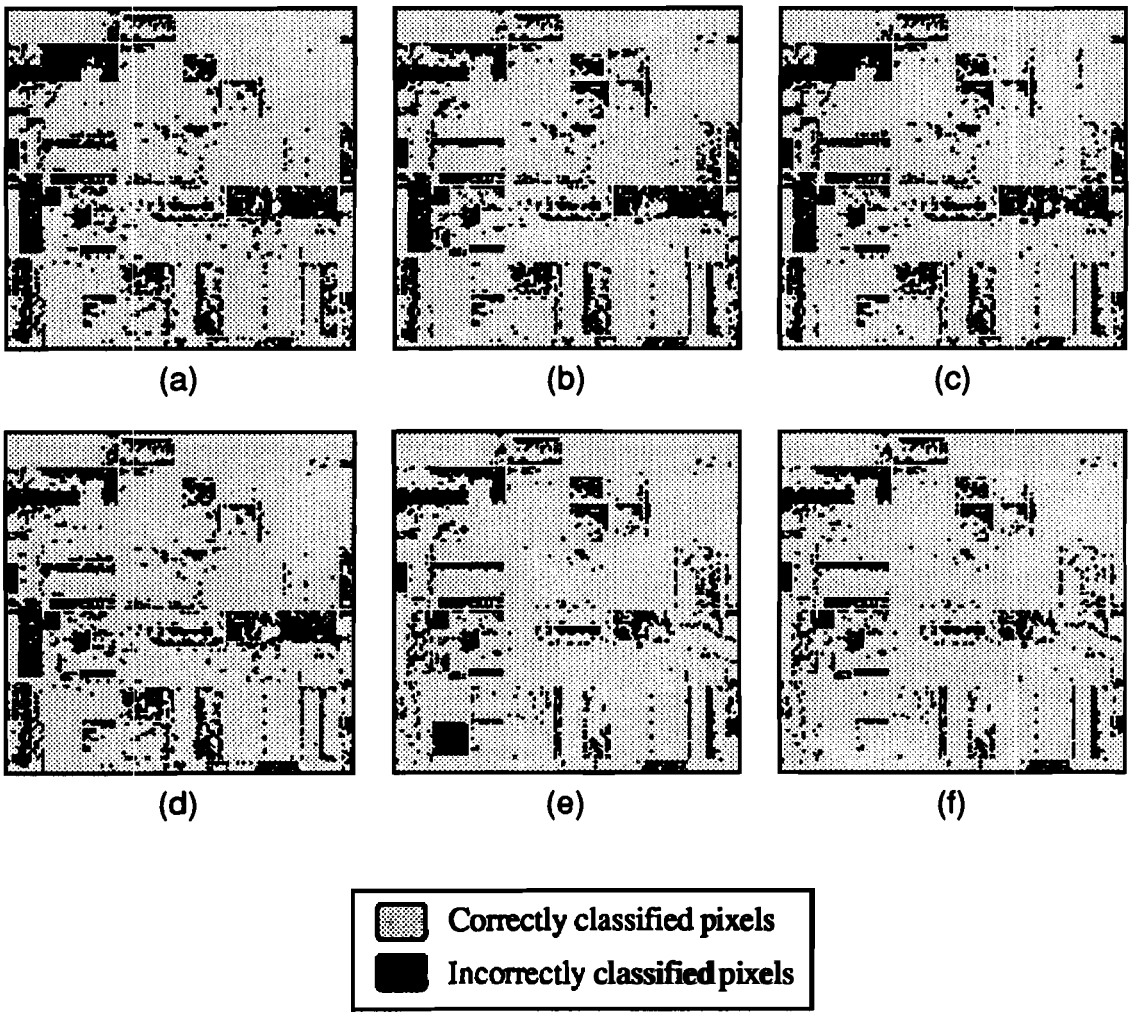


Figure 4.6 Classification Error Maps of July Data with Weighted Majority Decision Fusion with Classwise Reliability in eq. (4.15). (a) With April Data set. (b) Wih September Data set. (c) Wih April and September Data sets.; With Classwise reliability in eq. (4.16). (d) With April Data set. (e) With September Data set. (f) With April and September Data sets.

With eq. (4.16), the weighted majority fusion rule performed much better than the data fusion based rule in eq. (4.1). However, it could not surpass the performance of the maximum likelihood based fusion rule, although it followed very closely the performance. Note that the weighted majority fusion rule requires much less a priori information than the maximum likelihood decision fusion.

Decision fusion rules with cost functions in eq. (4.7.a,b) in a hybrid of maximum likelihood and weighted majority fusion were also tested and showed no significant advantages over the maximum likelihood fusion rule. In the case of the cost function in eq. (4.7.a), the relative magnitude between $A_k(u_k)$'s and the parameter B in (4.12.b) determines a balance of importance between the two decision fusion rules in the global decision. Several different B values were tested. As B became small (near 0), classification performance was dominated by those of the weighted majority fusion rule, and the opposite happened when B became closer to 1. In case of the cost function in eq. (4.7.b), the parameter B_{ML} in eq. (4.14.b) which is essentially a sum of 1 - $A_k(u_k)$'s and the classwise reliabilities $A_k(u_k)$, decides a balance between the two decision fusion rules. Both of the hybrids in eq. (4.7.a,b) performed less successfully compared to the maximum likelihood decision fusion rule.

Although there is further need for research on an optimum selection of data set and classwise reliabilities, multisource classification based on various decision fusion rules discussed in this chapter were observed to perform quite successfully compared to the non-contextual maximum likelihood classifier, or the multisource classifier with feature level fusion. Note that decision fusion approaches are computationally very simple and always applicable to classifying multisource data sets whenever the class decisions of the data sets are available. In contrast, the data fusion-based multisource classifiers combine posterior probabilities of each data set and therefore, all data sets must be describable with statistical probabilities. If data sets are very diverse in terms of their statistical properties, a combination of the posterior probabilities might not be able to produce desirable results since one data set with large ranges of probability values can easily dominate the global decision process. The decision fusion-based approach can be applied, on the contrary, whenever local decisions for each data set can be obtained. With data set and classwise reliability, or the information about conditional probability $P\{u_k \mid u_0\}$'s, it is very straightforward to

control the relative importance of a specific data set, or particular class decisions on the final global decision.

4.8 Conclusion

In this chapter, the problem of multisource classification based on decision fusion was addressed and an optimum decision fusion rule based on Bayesian minimum cost was derived. Three different decision fusion rules were considered with application to multisource data classification.

A maximum likelihood fusion rule was found to be most effective, and it also performed much better than the data fusion based multisource classifier. Although having a limited a priori information requirement compared to the maximum likelihood decision fusion rule, the weighted majority fusion rule performed better than the data fusion-based multisource classifier. Note that both classwise and data set reliabilities can be accommodated in weighted majority decision lusion.

Two different methods were considered in determining classwise reliabilities, and the classwise reliability based on eq. **(4.16)** was found, as expected, to be far more effective than the other.

This decision fusion-based approach in multisource classification or temporal classification is very attractive since it can be always applied to the multisource classification problem irrespective of the diverse nature of data sets whenever local class decisions are provided. This also enables independent processing of each data set separately both in training and actual classification steps.

CHAPTER 5

SPATIAL-TEMPORAL CONTEXTUAL CLASSIFICATION

5.1 Introduction

In this chapter, the spatial contextual classifiers discussed in Chapter 3 and the temporal contextual classifiers in Chapter 4 are combined as suggested in Chapter 2. The cascade classifier (Swain 78a) which was originally developed for bi-temporal data sets can be easily extended for a general multitemporal classification which has more than two temporal data sets, but, as discussed in the previous chapter, the extension requires one to decide on a structure by which temporal information is combined. The spatial-temporal contextual classifier H_{SPTP} in eq. (2.13) is based on an extension of the bi-temporal cascade classifier under the serial structure with which information of each temporal data set is combined in a serial way as shown in Fig. 4.2.(a).

It is also possible to formulate a similar spatial-temporal classifier under the parallel structure of Fig. 4.2.(b) with additional assumptions about class-conditional independence between different temporal sets. A modified spatial-temporal contextual classifier of H_{SPTP} in eq. (2.13) is derived in this chapter, to be suitable for the parallel structure in Fig. 4.2.(b). The decision fusion approach discussed in the previous chapter is also extended for spatial-temporal classification. The local decisions of each temporal data set obtained with a spatial classifier are combined, according to the decision fusion rules discussed previously, for the best global decision.

Experimental results with three temporal Landsat Thematic Mapper (TM) data sets are presented with discussions. Suggestions for future research in the field of the spatial-temporal contextual classification conclude this chapter.

5.2 Spatial-Temporal Contextual Classification Under a Parallel Information Combination Structure

Given p temporal data sets, spatial-temporal contextual classification which is optimal in the sense of maximum a *posteriori* (MAP) probability can be performed with the classifier, $H_{SPTP}(c; r, p)$ presented in eq. (2.13). This can be easily computed using the relations in eq. (2.12) and eq. (2.14). As depicted in Fig. 2.5, this classification scheme is based on a *temporal* classification scenario, that is, as a new temporal data set becomes available, its spatial contextual information extracted by $H_{SP}(\cdot)$ in eq. (2.11) is combined according to the rule in eq. (2.13) with the spatial-temporal contextual information available up to that time in a form of $H_{SPTP}(\cdot)$ in eq. (2.14) so that the updated contextual information can be conveyed to the next classification process of incoming temporal data.

It is also possible to formulate a spatial-temporal contextual classifier in a parallel structure with additional assumptions about class-conditional independence between temporally different data sets. Note that this parallel structure is generally used for multisource classification.

Suppose a classification decision is made among a user-defined set of classes, Ω_0. In the serial structure on which eq. (2.13) is established, Ω_0 is frequently selected as the set of classes in the final temporal data set, Ω_p. However, under the parallel structure, Ω_0 need not be restricted to Ω_p; it can be an arbitrary user-defined set of classes. Denote a random field $c_0(r)$ which indicates a class assigned to a pixel which is spatially located at r on the lattice L. $c_0(r)$ takes a value among the set Ω_0. Two additional assumptions on which the modified spatial-temporal contextual classifier is based, are stated as follow. (For an explanation of notation, refer to Chapter 2).

Assumption 3.

For any k, $1 \leq k \leq p$, any class $c_0 \in \Omega_0$, and for C_{others} which is a subset of $\xi_{C,k-1}$,

$$P\{c_k \mid c_0, C_{others}\} = P\{c_k \mid c_0\} \tag{5.1.a}$$

$$P\{C_{S,k} \mid c_k, c_0, C_{others}\} = P\{C_{S,k} \mid c_k\} \tag{5.1.b}$$

The assumption in eq. (5.1.a) is an extension of eq. (2.3.a) when a class $\mathbf{c_0} = c_0$ among the user defined set Ω_0 is involved. The relation in eq. (5.1.b) is of the same nature as eq. (2.3.b) in that, once the class of pixel $\mathbf{x_k(r)}$ is available, no additional information comes from the class $\mathbf{c_0}$, or the classes of its temporal neighbors. Note that eq. (5.1.a) is very crucial in allowing modification of eq. (2.13) into the parallel structure since it states that the class information of the temporal neighbors, $\mathbf{C_{others}}$, is irrelevant to evaluating the conditional probability of $\mathbf{c_k}$ once the identity of $\mathbf{c_0}$ is available. Based on this assumption, a useful relationship can be derived as follows.

Suppose $\mathbf{C_{others}}$ is a subset of $\xi_{C,k-1}$, and η_k takes either $\mathbf{c_k}$, or $\mathbf{C'_{S,k}}$. Then,

$$P\{\mathbf{C_{others}} \mid \mathbf{c_0}, \eta_k\} = \frac{P\{\eta_k \mid \mathbf{C_{others}}, \mathbf{c_0}\} P\{\mathbf{C_{others}} \mid \mathbf{c_0}\}}{P\{\eta_k \mid \mathbf{c_0}\}}$$

From the assumption in eq. (5.1.a,b), the first term is $P\{\eta_k | \mathbf{c_0}, \mathbf{C_{others}}\} = P\{\eta_k | \mathbf{c_0}\}$, therefore, the following relation holds.

$$P\{\mathbf{C_{others}} \mid \mathbf{c_0}, \eta_k\} = P\{\mathbf{C_{others}} \mid \mathbf{c_0}\} \tag{5.2}$$

A direct application of eq. (5.2) shows that, (data) classes $\mathbf{c_u}$ and $\mathbf{c_t}$ of temporally different data sets, $u \neq t$, are class-conditionally independent as,

$$P\{\mathbf{c_u}, \mathbf{c_t} \mid \mathbf{c_0}\} = P\{\mathbf{c_u} \mid \mathbf{c_0}\}\, P\{\mathbf{c_t} \mid \mathbf{c_0}\}$$

The following is another slightly extended class-conditional independence assumption of eq. (2.14) when $\mathbf{c_0}$ is involved.

Assumption 4.

For any k, $1 \leq k \leq p$, and for any class $\mathbf{c_0} \in \Omega_0$,

$$P\{\mathbf{X_A} \mid \mathbf{C_A}, \mathbf{X_{others}}, \mathbf{C_{others}}, \mathbf{c_0}\} = P\{\mathbf{X_A} \mid \mathbf{C_A}\} \tag{5.3}$$

where,

$\mathbf{X}_A$ is an any non-empty subset of $\mathbf{X}'_{S,k}$.
$\mathbf{C}_A$ is a set of classes corresponding to $\mathbf{X}_A$.
$\mathbf{X}_{others}$ is an any subset of $\xi_{X,p}$ such that $\mathbf{X}_{others} \cap \mathbf{X}'_{S,k} = \phi$.
$\mathbf{C}_{others}$ is an any subset of $\xi_{C,p}$ such that $\mathbf{C}_{others} \cap \mathbf{C}'_{S,k} = \phi$.
($\mathbf{C}_{others}$ is not necessarily a set of classes corresponding to $\mathbf{X}_{others}$).

This assumes class-conditional independence of temporal data sets, irrespective of whether the class $\mathbf{c}_0$ is conditioned or not. The validity of the assumption in eq. (5.3) may be hard to prove, but as suggested in (Lee *et al.* 87), without further a priori correlation information between temporal data sets, this can be a practical assumption to keep classifier complexity and the prior knowledge requirement within a manageable limit.

Suppose $\mathbf{X}_B$ is either $\mathbf{x}_k$, or $\mathbf{X}'_{S,k}$ and $\mathbf{C}_B$ is its corresponding set of classes. Then, from the assumption in eq. (5.3), the following relationship can be established.

$$P\{\mathbf{X}_{others} \mid \mathbf{C}_B, \mathbf{c}_0\} = P\{\mathbf{X}_B \mid \mathbf{c}_0\} \tag{5.4}$$

Since the probability $P\{\mathbf{X}_{others} \mid \mathbf{C}_B, \mathbf{c}_0\}$ is computed as,

$$\sum_{\mathbf{C}_{others}} P\{\mathbf{X}_{others} \mid \mathbf{C}_{others} = C_{others}, \mathbf{C}_B, \mathbf{c}_0\} P\{\mathbf{C}_{others} = C_{others} \mid \mathbf{C}_B, \mathbf{c}_0\} \tag{5.5}$$

Its first term may be written, from the assumption in eq. (5.3), as,

$$P\{\mathbf{X}_{others} \mid \mathbf{C}_{others} = C_{others}, \mathbf{C}_A, \mathbf{c}_0\} = P\{\mathbf{X}_{others} \mid \mathbf{C}_{others} = C_{others}\}$$

From eq. (5.2), the second term is given as $P\{\mathbf{C}_{others} = C_{others} \mid \mathbf{c}_0\}$. Substituting these in eq. (5.5) proves the relationship in eq. (5.4).

Under the two assumptions in eq. (5.1.a,b) and eq. (5.3), a theorem which is useful for deriving a modified spatial-temporal contextual classifier, can be established as,

Theorem 3.

For any t, $1 \leq t \leq p$, and for $\mathbf{X}_A$, $\mathbf{c}_0$, and $\mathbf{X}_{others}$ defined below,

$$P\{\mathbf{X}_A \mid \mathbf{c}_0, \mathbf{X}_{others}\} = P\{\mathbf{X}_A \mid \mathbf{c}_0\} \tag{5.6}$$

where $\mathbf{X}_A$ is either $\mathbf{x}_t$, or $\mathbf{X}'_{S,t}$. $\mathbf{X}_{others}$ is any subset of $\xi_{X,t-1}$.

This can easily be proved using the results in eq. (5.2) and eq. (5.4). Note that the probability $P\{\mathbf{X}_A \mid \mathbf{c}_0, \mathbf{X}_{others}\}$ in eq. (5.6) can be written as,

$$\sum_{C_A} P\{\mathbf{X}_A \mid \mathbf{C}_A = C_A, \mathbf{c}_0, \mathbf{X}_{others}\}\, P\{\mathbf{C}_A = C_A \mid \mathbf{c}_0, \mathbf{X}_{others}\} \tag{5.7}$$

From the class-conditional independence assumption of eq. (5.3),

$$P\{\mathbf{X}_A \mid \mathbf{C}_A = C_A, \mathbf{c}_0, \mathbf{X}_{others}\} = P\{\mathbf{X}_A \mid \mathbf{C}_A = C_A\} \tag{5.8.a}$$

Using Bayes theorem and eq. (5.4), the second term of eq. (5.7) may be expressed as,

$$P\{\mathbf{C}_A = C_A \mid \mathbf{c}_0, \mathbf{X}_{others}\} = P\{\mathbf{C}_A = C_A \mid \mathbf{c}_0\} \tag{5.8.b}$$

Since $P\{\mathbf{X}_A \mid \mathbf{C}_A = C_A\}$ in eq. (5.8.a) is equal to $P\{\mathbf{X}_A \mid \mathbf{C}_A = C_A, \mathbf{c}_0\}$; substituting this and eq. (5.8.b) in eq. (5.7) proves eq. (5.6). Note that a direct application of eq. (5.6) establishes,

$$P\{\mathbf{x}_1, ---, \mathbf{x}_p \mid \mathbf{c}_0\} = \prod_{k=1}^{p} P\{\mathbf{x}_k \mid \mathbf{c}_0\} \tag{5.9}$$

which is frequently assumed in multisource data classification. $\mathbf{x}_1$, ---, $\mathbf{x}_p$ are the feature vectors corresponding to a same spatial location.

With the two additional assumptions 3 and 4 for the case when $\mathbf{c}_0$ is involved, it is straightforward to derive a modified spatial-temporal contextual classifier. Under the parallel information combination structure in Fig. 4.2, decisions with p multitemporal data sets are made among a user-defined set of classes, Ω_0. In a manner similar to eq. (2.2), $H_{SPTP}(c_0; r, k)$, $1 \le k \le p$, is defined as,

$$H_{SPTP}(c_0; r, k) \equiv P\{\mathbf{c}_0(r) = c_0 \mid \mathbf{x}_k = x_k, \mathbf{X}_{S,k} = X_{S,k}, \mathbf{X}_{T,k} = X_{T,k}\} \tag{5.10}$$

When k = 1, $H_{SPTP}(c; r, k)$ is understood as $P\{\mathbf{c}_k = c \mid \mathbf{x}_k = x_k, \mathbf{X}_{S,k} = X_{S,k}\}$ since there are no temporally previous sets, that is, $\mathbf{X}_{T,k}$ is empty. Note that, under the parallel structure, a class decision is made among the set, Ω_0. Spatial-temporal classification is performed by choosing a class $\mathbf{c}_0(r) = c_0$ which maximizes $H_{SPTP}(c_0; r, p)$ among Ω_0. By applying Bayes theorem and the results in eq. (5.6), $H_{SPTP}(c_0; r, k)$ can be expressed, for $2 \le k \le p$, as,

$$H_{SPTP}(c_0; r, k) = A_k \frac{H_{SP}(c_0; r, k)\, H_{TP}(c_0; r, k)}{P\{\mathbf{c}_0(r) = c_0\}} \tag{5.11}$$

The term A_k which is defined in eq. (2.10), is irrelevant to making a class decision c_0 and therefore, it needs not be evaluated. For the case k = 1, $H_{SPTP}(c_0; r, k=1) = H_{SP}(c_0; r, k=1)$. $H_{SP}(c_0; r, k)$ in eq. (5.11) is a spatial contextual part of the k^{th} data set and is defined, for $1 \le k \le p$, as,

$$H_{SP}(c_0; r, k) = P\{\mathbf{c}_0 = c_0 \mid \mathbf{X}_{S,k} = X_{S,k}, \mathbf{x}_k = x_k\} \tag{5.12}$$

This is the same spatial contextual classifier part as eq. (2.11) but formulated in terms of a class in Ω_0. This can be computed, using the class transition matrix $T(\Omega_0 \mid \Omega_k)$, from the spatial contextual part in terms of classes among Ω_k in eq. (2.11) as,

$$H_{SP}(c_0 ; r, k) = \sum_{c \in \Omega_k} H_{SP}(c; r, k)\, P\{c_0(r) = c_0 \mid c_k(r) = c\} \tag{5.13}$$

where c is a class among Ω_k, the set of classes for the k^{th} data set. $H_{SP}(c; r, k)$ is the spatial contextual classifier in terms of data classes in Ω_k and computed using eq. (3.1).

In the same way, the temporal contextual classifier part, for $2 \le k \le p$, which is defined as,

$$H_{TP}(c_0; r, k) = P\{c_0(r) = c_0 \mid X_{T,k}\} \tag{5.14}$$

can expressed, for $2 \le k \le p$, as,

$$H_{TP}(c_0; r, k) = H_{SPTP}(c_0; r, k-1) \tag{5.15}$$

This is different from its serial counterpart in eq. (2.14). The assumption in eq. (5.1.a) is indispensable in establishing this relationship. From eq. (5.11) and eq. (5.15), the spatial-temporal contextual classifier, with p temporal data sets, under a parallel information combination structure, $H_{SPTP}(c_0; r, p)$ can be expressed as,

$$H_{SPTP}(c_0; r, p) = P\{c_0 = c_0\} \prod_{k=1}^{p} \frac{H_{SP}(c_0; r, k)}{P\{c_0 = c_0\}} \tag{5.16}$$

The terms, A_k's are dropped from $H_{SPTP}(c_0; r, p)$ in eq. (5.16) since they are irrelevant in making the decision of c_0. Note that it is in the same form of eq. (4.1)

which is a data fusion-based multisource. A flowchart of the spatial-temporal contextual classifier under the parallel structure is shown in Fig. 5.1.

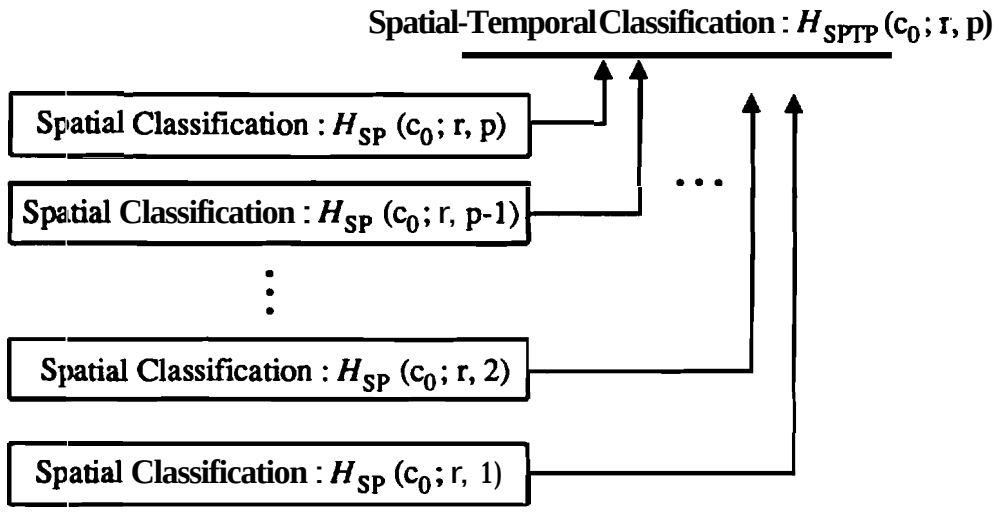


Figure 5.1 Spatial-Temporal Classification Under Parallel Information Combination Structure.

One important difference of the spatial-temporal contextual classifier in eq. (5.16) from that in eq. (2.14) is that there is no interaction between the class $c_0(r)$ and the classes of its spatial neighbors, that is, $c_0(r+v)$, $v \in N_S$. Under the serial structure in eq. (2.14), at each temporal stage, spatial information of that temporal set is combined with its temporal information to execute spatial-temporal contextual classification. But in eq. (5.16), only the spatial contextual information of each temporal data set is combined to make a decision among Ω_0, and there is no interaction between the global decisions of spatially adjacent pixels.

In the same manner as in Chapter 4, the decision fusion approach can be taken for the classifier of eq. (5.16). The class decisions obtained with the spatial contextual classifiers $H_{SP}(c_0; r, k)$, $k = 1, \cdots, p$, are combined together to find a global decision among Ω_0.

The spatial-temporal contextual classifiers in eq. (2.13) and eq. (5.16) are quite general in their scope of application in that, the spatial contextual classification

parts, $H_{SP}(\cdot)$'s can be independently defined according to the particular properties of the data sets. For example, different spatial neighborhoods can be assumed for different temporal data sets. This generalization might be quite useful when sensors with different spatial resolutions are used to acquire temporal data sets. In experiments in this report, for simplicity's sake, only a first order spatial neighborhood system is considered for all given multitemporal data sets.

5.3 Experiments on Spatial-Temporal Contextual Classification

To test the spatio-temporal contextual classifiers in eq. (2.13), or in eq. (5.16) and their modification based on decision fusion, experiments were carried out with the three Landsat Thematic Mapper data sets introduced in previous chapters. Data set descriptions can be found in Chapter 3 and 4. Since the ground truth information shown in Fig. 3.8 was gathered in July and therefore matches best with July data, classification performances were evaluated by comparing classification results of July data with the ground truth map in terms of class-averaged classification accuracy (CAG) and overall classification accuracy (OVA). Classification results with only spatial contexts, or only temporal contexts were presented in previous chapters and, in this chapter, only the results with spatial-temporal contexts are shown.

Table 5.1 Description of Spatial Contextual Part $H_{SP}(\cdot)$ in eq. (2.13).

Classifier	Description of Classifiers
RECU - 1	With the recursive spatial contextual classifier as in eq. (3.28) (With isotropy assumption)
RECU - 2	With the recursive spatial contextual classifier as in eq. (3.28) (Without isotropy assumption)
CM - 1	Spatial correlation context only for homogeneous pixels with eq. (3.29) (With isotropy assumption)
CM - 2	Spatial correlation context only for homogeneous pixels with eq. (3.29) (Without isotropy assumption)

In the first experiment, the July data set was classified with September data as a temporal neighbor set, therefore, the number of temporal data sets, p, was two. The July data set was used as X(2) and the September data set was used as X(1). First, the spatial-temporal classifier in eq. (2.13) was tested. Four different spatial classification schemes, which were introduced in Chapter' 3, were employed for $H_{SP}(\cdot)$ in eq. (2.13) as shown in Table 5.1.

All four spatial classifiers are able to utilize spatial interpixel correlation contexts. In addition, a spatial classifier which utilizes only the spatial interpixel class dependency context in eq. (3.35) was also examined in the experinient. The same class transition matrix as in Chapter 4 was used. For details of spatial and temporal classification, refer to Chapter 3 and 4. Spatial-temporal classification results are shown in Table 5.2. in which the result of a non-contextual maximum likelihood classifier is also included for comparison.

To classify the July data set with the September data set using $H_{SPTP}(\cdot)$ in eq. (2.13), the selected spatial classifier, $H_{SP}(\cdot)$ was applied first to the pixels in the September data set; its classification result was then used in eq. (2.14) to compute $H_{TP}(\cdot)$; and finally eq. (2.13) was used to classify pixels in July data set. Both of the classifications were performed recursively over x-sites and •-sites in Fig. 3.7 until negligible changes of class assignments were attained.

As seen in Table 5.2, in the case of using spatial interpixel correlation contexts, there were 3 ~ 6% overall classification accuracy increases over the non-contextual maximum likelihood classification. Compared to the 8.40% increase over the non-contextual scheme with only spatial interpixel class dependency context, the spatial correlation contexts were not so effective. Considering the additional computational complexity due by including spatial correlation contexts and their relatively low effectiveness compared to the spatial class dependency context case, in the following experiments, only the spatial class dependency context was used in spatial-temporal classification.

The improvement with $H_{SPTP}(\cdot)$ of eq. (2.13) over the non-contextual classifier was very significant for either the temporal, or spatial contexts only classification. The improved classification results of the September data set using spatial context with $H_{SP}(\cdot)$ was very helpful as temporal contextual information in classifying the July data.

Table 5.2 Percent Accuracy Comparison of Classifying July Data with September Data using the Serial Spatial-Temporal Contextual Classifier in eq. (2.13).

Classifier	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
July, ML[1]	90.18	57.72	68.72	77.89	73.63	74.37
July, Spatial Only[2]	94.51	57.28	73.50	80.82	76.53	76.82
With only spatial class label dependency context in eq. (3.35)						
	95.92	75.01	81.32	62.88	78.78	82.77
With interpixel correlation context, without spatial class label dependency context						
RECU - 1	94.33	59.19	73.50	77.38	76.10	77.08
RECU - 2	94.59	58.77	73.00	77.09	75.86	76.96
C M - 1	92.93	68.93	71.19	69.40	75.61	79.00
C M - 2	92.88	68.80	71.03	69.55	75.59	78.98
With interpixel correlation and spatial class label dependency contexts						
RECU - 1	95.65	60.30	73.83	81.41	77.80	78.50
RECU - 2	95.66	59.90	73.17	81.55	77.57	78.32
CM - 1	94.50	69.87	72.18	71.16	76.93	80.30
C M - 2	94.62	69.85	72.02	71.45	76.99	80.36

[1] Non-contextual maximum likelihood classification of July data.
[2] With only spatial interpixel class dependency context.

The same July data set was classified with the parallel spatial-temporal contextual classifier of eq. (5.16), and the results are presented in Table 5.3.

Table 5.3 **Percent Accuracy Comparison of Classifying July Data with September Data using the Parallel Spatial-Temporal Contextual Classifier in eq. (5.16).**

Classifier	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
			Data Fusion Based[1]			
Without Spatial contexts[3]	91.80	63.08	69.88	73.35	74.53	76.67
With Spatial Contexts	95.20	61.81	76.46	76.65	77.53	78.61
			Decision Fusion Based[2]			
without Spatial Contexts[3]	94.19	75.63	68.72	73.79	78.08	82.24
With Spatial Contexts	95.88	79.11	72.67	77.89	81.39	85.04

[1] With eq. (5.16).
[2] Maximum Likelihood Decision Fusion rule.
[3] These are obtained without spatial contexts and are copied from Table 4.5 and 4.7 for comparison.

As discussed in Chapter 4, the data fusion-based and the decision fusion-based approaches were taken. To see the effectiveness of applying spatial classification to each temporal data set, the classification results without partial contexts in Table 4.7 are also included in Table 5.3.

To classify the July data set with the September data set using the data fusion-based $H_{SPTP}(\cdot)$ in eq. (5.16), the spatial classifier with only spatial interpixel class dependency contexts in eq. (3.35) was applied to the September data; the results of September data, $H_{SP}(c;\ r,\ k=1)$'s were translated into $H_{SP}(c_0\ ;\ r,\ k=1)$'s to be used as $H_{TP}(\cdot)$; finally, according to eq. (5.16), $H_{TP}(\cdot)$ and $H_{SP}(c_0\ ;\ r,\ k=2)$ was combined to classify pixels in the July data set. In the decision fusion approach, the class decisions of July and September data with the spatial classifier in eq. (3.35) are combined according to the maximum likelihood decision fusion rule in eq. (4.5.a). Note that in making decisions among Ω_0, there is no spatial interaction between c_0 at r and its spatial neighbors, $c_0(r+v)$, $v \in N_S$.

Compared to the case of temporal classification without spatial contexts in Table 5.3, there were only about 2 ~ 3% overall classification accuracy increase by incorporating spatial contexts in the temporal contextual classification. The data fusion-based classifier in eq. (5.16) was about 4% worse than the serial counterpart in eq. (2.13). This is because there is no consideration of spatial interactions between c_0 at rand its spatial neighbors in eq. (5.16). However, in the case of decision fusion-based the information combination of eq. (5.16), there was a 2.3% classification accuracy increase (OVA) over the result with eq. (2.13). Compared to the non-contextual maximum likelihood classification of the July data, this amounts to a **10.7%,** a significant increase (OVA). The class-averaged accuracy was also increased by 7.8%. Better classification results of each temporal data set by using the spatial contextual classifier brought significant accuracy increases when decision fusion took place.

Figure 5.2 shows the locations where pixels were incorrectly classified. The error map of spatial classification of the July data with only class dependency context in eq. (3.35) is also included for visual comparison.

Compared to the error map of the non-contextual maximum likelihood classifier in Fig. **5.2.(a),** the other error maps in Fig. 5.2 look much cleaner with far fewer isolated errors. This is due to utilizing the spatial-temporal interpixel dependency class context. This cleaner classification result will be much more meaningful in real applications of classification. Also some regions in the July data set which were incorrectly classified with spatial contexts only were correctly classified by utilizing additional spatio-temporal contextual information from the temporal neighbors in the September data set. Therefore we can say that it is very effective to incorporate contextual information from the spatio-temporal neighbors into classification. In the error maps Fig. **5.2.(a)** and (b), there were large incorrectly classified fields in the middle of scene, and they were of the class soybeans, but mostly classified to the class **alfalfa/oats.** When temporal contextual information was incorporated, as shown in Fig. **5.2.(c)** and (d), a considerable number of pixels in those incorrect fields were correctly classified. Note that, there was significant increase in classification accuracy of soybeans in Table 5.3. Figure 5.3 shows corresponding classification maps.

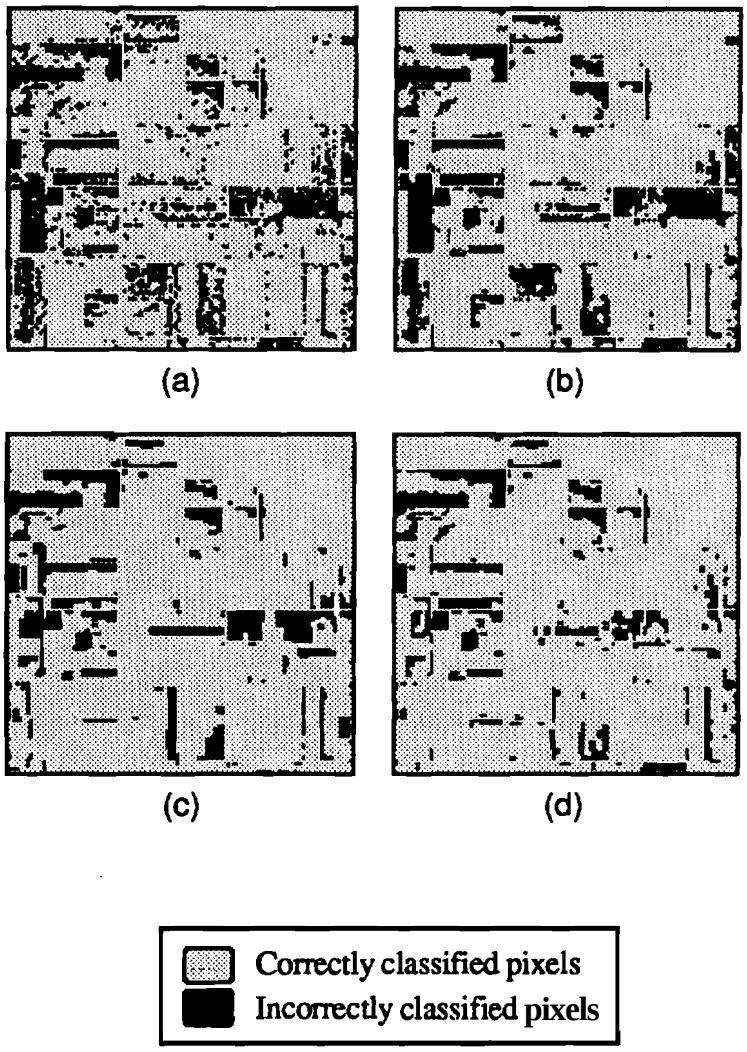


Figure 5.2 Error Maps of Spatial-Temporal Contextual Classification of July Data with September Data. (a) With pixelwise maximum likelihood classifier (no spatial, temporal contexts). (b) With eq. (3.35) (no temporal contexts). (c) With eq. (2.13). (d) With eq. (5.16), maximum likelihood decision fusion based.

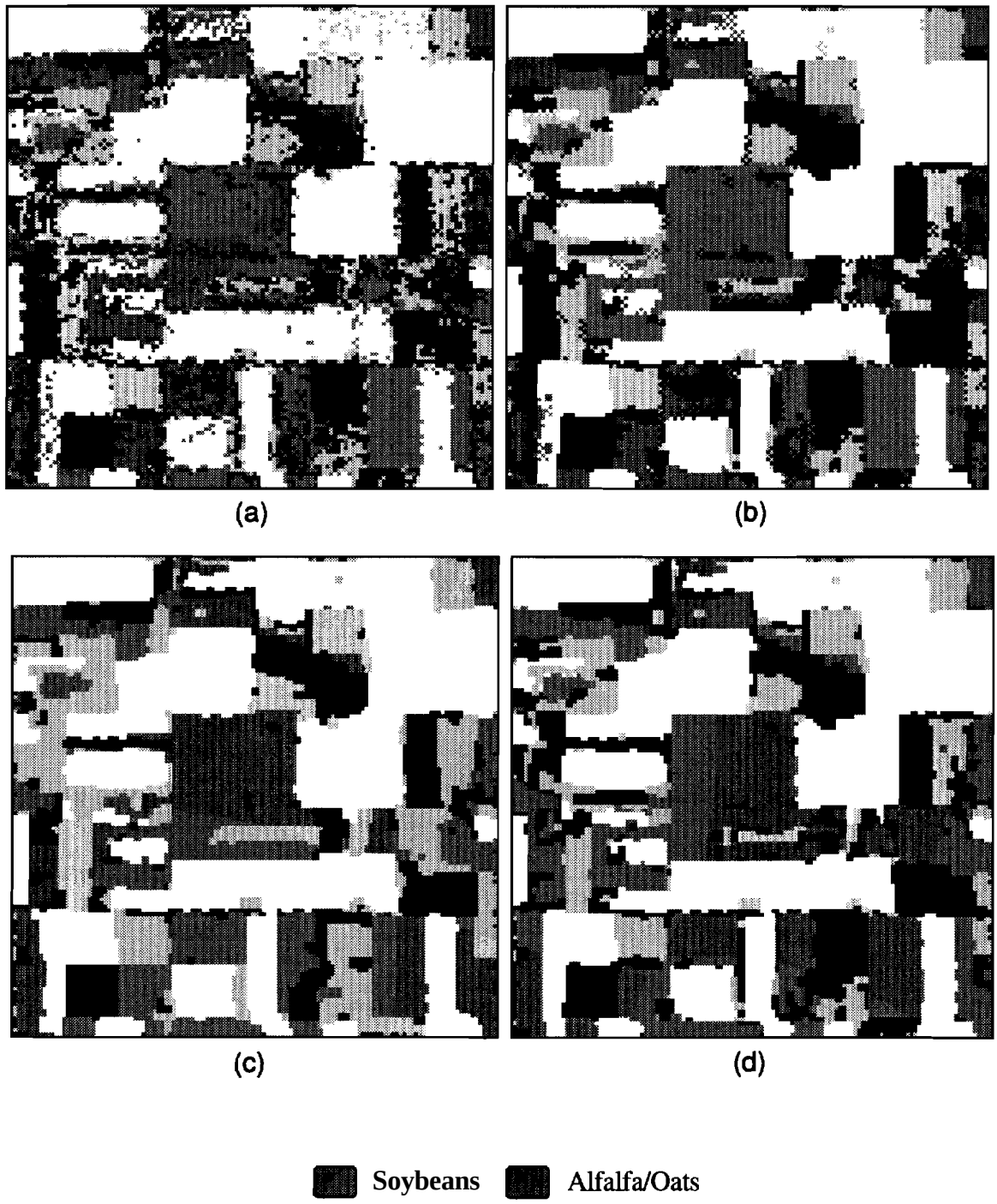


Figure 5.3 Classification Maps of Spatial-Temporal Contextual Classification of July Data with September Data. (a) With a Non-Contextual Maximum Likelihood Classifier. (b) With a Spatial Classifier using eq. (3.35). (c) With Data fusion using eq. (2.13). (d) With maximum likelihood decision fusion using eq. (5.16).

Classification accuracy increases, over non-contextual pixelwise maximum likelihood classifier, by incorporating spatial and/or temporal contextual information are summarized in Fig. 5.4.

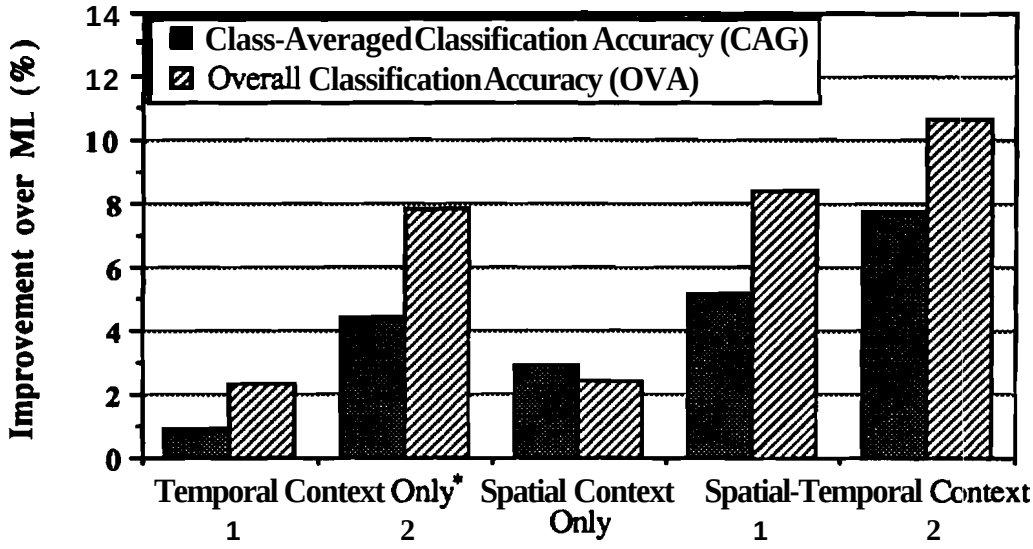


Figure 5.4 Improvement of Classification Accuracy, over a Pixelwise Maximum Likelihood Classifier, by Incorporating Contextual Information in Classifying July Data with September Data as a Temporal Neighbor. 1. Data fusion-based temporal contextual classification (cascade classifier) with eq. (4.1) - serial combination structure.; 2. Maximum likelihood decision fusion-based temporal contextual classification with eq. (4.5.a) - parallel connbination structure.; In spatial classification, only the spatial class dependency context was used with eq. (3.35).

In the case of data fusion-based classification with temporal context only, the improvements shown here are based on the result in Table 4.5, with the data set weights (Sept=July=1). Compared with the best result in terms of overall classification accuracy, which is shown in Table 4.7, the improvements are OVA = 2.43%, CAG= - 0.16%.

In the second experiment, the April data set was also included in classifying July data. with September data. All three temporal data sets were used to examine the effectiveness of the spatial-temporal classifiers discussed in this chapter. Under the serial spatial-temporal contextual classifier in eq. (2.13), the posterior probabilities obtained with spatial classification of April data were combined with those of September data according to eq. (2.14). These results were then combined with $H_{SP}(\cdot)$ of July data to make class decisions. In the parallel data

fusion case of eq. (5.16), the $H_{SP}(\cdot)$ of each data set were combined using eq. (5.16).

Spatial classification results of April data with the classifier of eq. (3.35) are shown in Table 5.4.

Table 5.4 Percent Accuracy Comparison of Spatial Classification of April Data.

b in eq. (3.35)	Wheat	Others	CAG	OVA
0	90.29	89.59	89.94	89.65
2	91.19	90.17	90.68	90.27
4	91.03	90.02	90.52	90.11
6	91.28	90.02	90.65	90.14
8	91.28	89.90	90.59	90.03
10	91.28	89.89	90.58	90.02
12	91.28	89.84	90.56	89.98
14	91.28	89.77	90.52	89.91
16	91.28	89.74	90.51	89.88
18	91.28	89.69	90.48	89.84
20	91.28	89.70	90.49	89.84

Spatial classification of April data with eq. (3.35) didn't make significant differences compared to the pixelwise classification (b = 0 case in Table 5.4). Note that there were defined only 2 classes {wheat, others) for April data. In the data fusion based spatial-temporal classification procedures, the parameter b in eq. (3.35) decides the relative emphasis on spatial class homogeneity compared to the class-conditional likelihood values. If different b values are used for different data sets, then, the data set with the largest b will dominantly affect the classification decisions as discussed in Chapter 4, since that data set is most likely to have the largest range of $H_{SP}(\cdot)$ values. Therefore, the parameter b's must be decided for each data set carefully. With the decision fusion-based classifier, on the contrary, proper values of b's can be independently selected for each data set to best fit each data set.

Table 5.5 **Percent Accuracy Comparison of Classifying July Data with April and September using the Spatial-Temporal Contextual Information.**

Selected Classifier	Percent Classification Accuracy					
	Corn	Soybeans	Wheat	Alfalfa/Oats	CAG	OVA
Classification of July Data with April Data						
Data Fusion : eq. (2.13)	94.32	56.69	84.36	82.43	79.45	77.70
Data Fusion : *eq.* (5.16)	94.64	57.07	77.28	81.70	77.67	77.24
Decision Fusion : *eq.* (4.5.a)	94.51	57.28	90.12	81.63	80.89	78.46
Classification of September Data with April Data						
Data Fusion : eq. (2.13)	83.67	60.00	82.22	54.25	70.03	71.20
Data Fusion : *eq.* (5.16)	83.76	59.71	59.59	53.00	64.01	68.89
Decision Fusion : eq. (4.5.a)	83.79	59.52	87.24	73.57	76.03	73.58
Classification of July Data with April and September Data						
Data Fusion : eq. (2.13)	95.90	72.60	81.65	63.98	78.53	82.03
Data Fusion : eq. (5.16)	95.29	61.66	80.58	78.04	78.89	79.13
Decision Fusion : eq. (4.5.a)	96.56	79.11	89.63	74.23	84.88	86.53

Table 5.5 shows classification results for July data when both September and April data were used as temporally previous data sets. The parameter b was chosen as 0 for April data. The result of the July data classification with temporal context from April data is also included in the table.

As observed in Chapter 4, due to its relative small number of classes, "wheat" compared to others, April data was only marginally effective in improving the overall classification (OVA) as shown in Table 5.5. However, there was a considerable accuracy increase for the class wheat. Compared to the temporal classification case (CAG=79.07, OVA=76.16) in Table 4.4, and the spatial context only case (CAG=76.53, OVA=76.82) in Table 3.5, the spatial-temporal information in classifying July data with April data was useful as shown in Table 5.5.

Decision fusion-based spatial-temporal classification outperformed the others as seen in the previous bi-temporal classification case of July data with September data in Table 5.3. When the September data was classified with the April data, the performances with spatial-temporal information was better than that of the spatial context only (CAG=61.47, OVA=67.86) in Table 3.5. Compared with the temporal context only case (CAG=73.17, OVA=71.06) in Table 4.4, the results with eq. (2.13) were not much different.

Due to the relatively large differences of b values between April (b=0) and September data (b=30), the classes wheat and alfalfa/oats in Table 5.5 were not as accurately classified as with the data fusion-based temporal classifier in Table 4.4. In a separate experiment of classifying September data with April data, it was observed that there were differences in the overall classification accuracy of 3% (a maximum was 73.13%, and a minimum was 69.89%), and in the class-averaged classification accuracy of 5% (a maximum was 75.44%, and a minimum was 70.03%) for various combination of b values for the two temporal data sets. As the b value for the April data set increased, the class-averaged accuracy was seen to increase due to better classification for the classes, wheat and alfalfa/oats.

This effect was also visible in the classification with all three data set together using eq. (2.13). There were differences in the overall classification accuracy of

4.6% (a maximum was 83.82%, and a minimum was 79.18%), and in the class-averaged classification accuracy of 5.6% (a maximum was 81.59%, and a minimum was 76.02%) in the separate experiment of classifying with all three data sets and with various combination of b values for September and April data. The classification result with eq. (2.13) in Table 5.5 turned out to be slightly worse than the result of the spatial context only case (OVA=82.77, CAG=78.78). Due to the relatively low emphasis on April data (b=0), the classes wheat and alfalfa/oats had low classification accuracies, and so did the class-averaged classification accuracy (CAG). However, the decision fusion-based classifier was very successful even in this case. For all four classes, classification accuracies were increased significantly over the non-contextual, and the spatial contextual only classification of July and September data sets. This insensitivity is a direct consequence of dealing with only the class decisions. Notice that it also attained better performances for all classes compared with the same classifier but with only temporal contexts in Table 4.7.

Figure 5.5 shows the locations of classification errors. For visual comparison, the error maps of the non-contextual maximum likelihood classifier and of tho spatial classifier in eq. (3.35) are also presented. When spatial class dependericy context was used, many isolated errors in Fig. 5.5.(a) were removed. But, the large incorrectly classified soybean fields in the middle of scene were considerably corrected by using temporal contexts as seen in Fig. 5.2. Significant portions of the wheat fields in upper right portion of image which were incorrectly classified without April data set, were mostly correctly classified as seen in Fig. 5.5.(d). In Fig. 5.5.(c), wheat fields were still in errors since the ranges of class-conditional likelihood values of April data set were relatively smaller than for the other data set due to its low b parameter value (b=0) in eq. (3.35). Most of the classification errors left in Fig. 5.5.(d) were along field boundaries, in which the problem of mixed pixels might most likely exist. Figure 5.6 shows corresponding classification maps which clearly exhibit the effectiveness of spatial and temporal contextual information

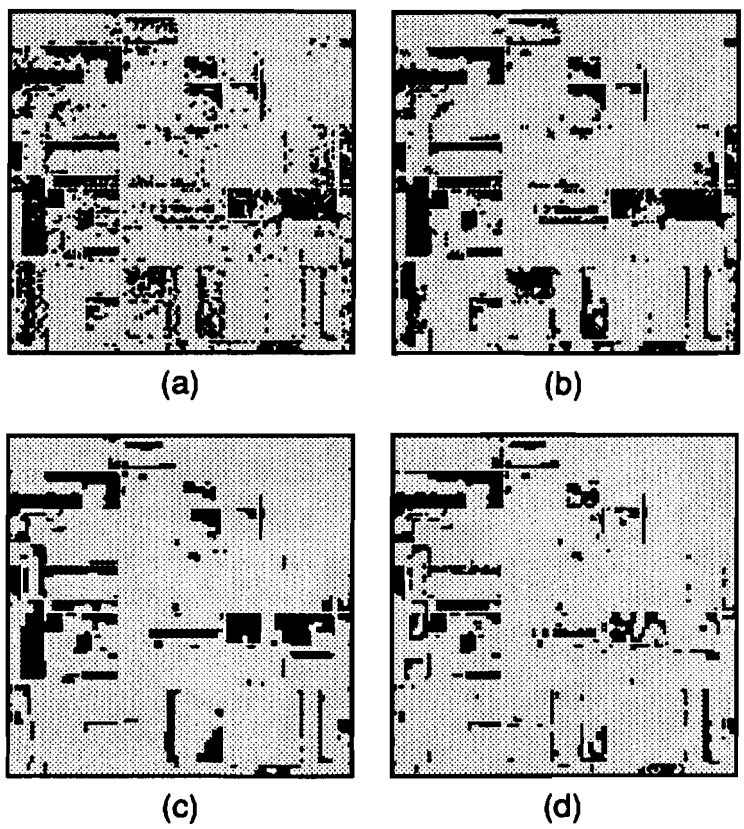


Figure 5.5 Error Maps of Spatial-Temporal Contextual Classification of July Data with April and September Data. (a) With pixelwise maximum likelihood classifier (no spatial, temporal contexts). (b) With eq. (3.35) (no temporal contexts). (c) With eq. (2.13). (d) With eq. (5.16), maximum likelihood decision fusion based.

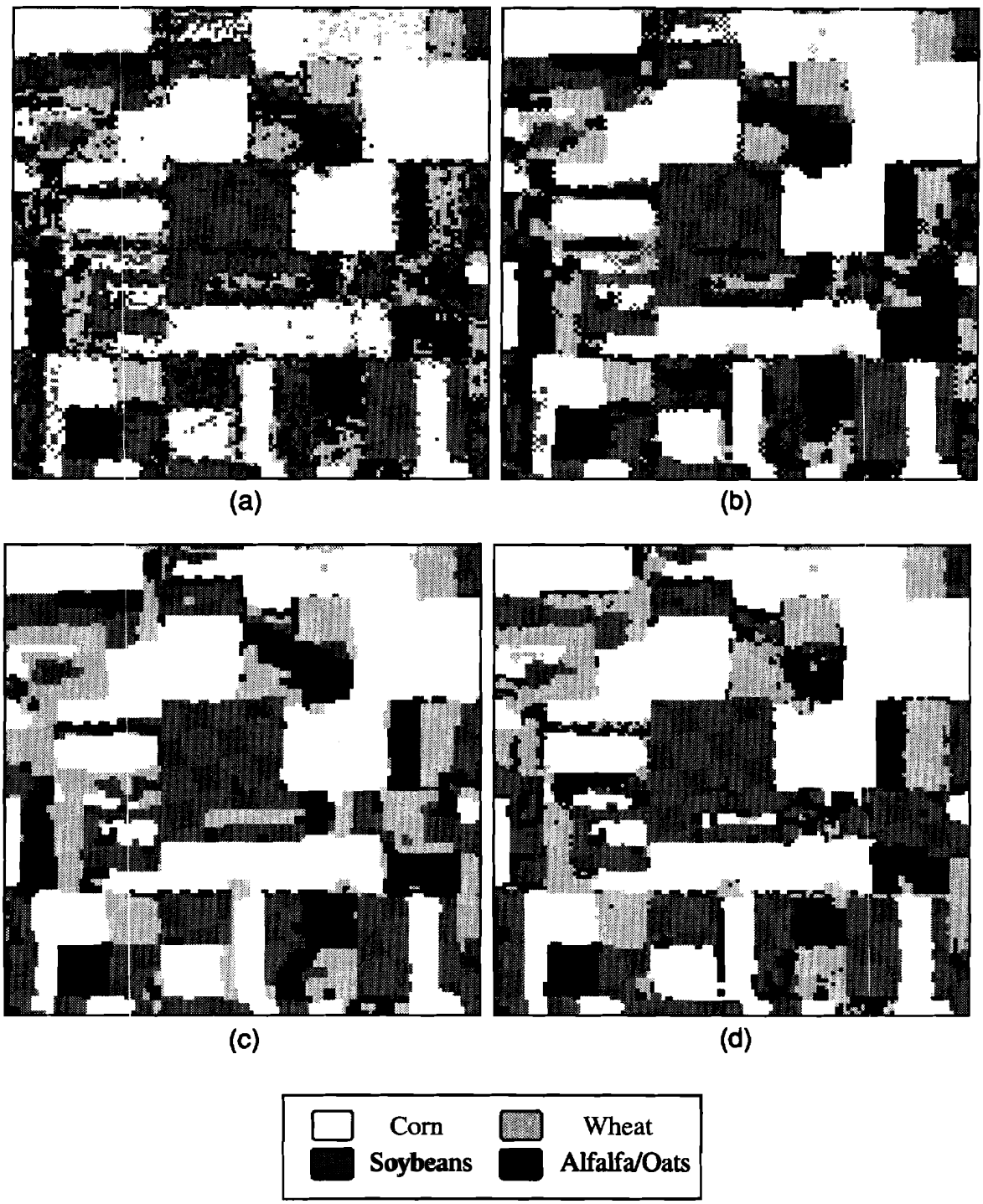


Figure 5.6 Classification Maps of Spatial-Temporal Contextual Classification of July Data with April and September Data. (a) With Non-Contextual Maximum Likelihood Classifier. (b) With Spatial Classifier using eq. (3.35). (c) With Data fusion using eq. (2.13). (d) With maximum likelihood decision fusion using eq. (5.16).

A summary of classification accuracy improvement over the non-contextual maximum likelihood classification of July data by incorporating the spatial-temporal contextual information is given in Fig. 5.7.

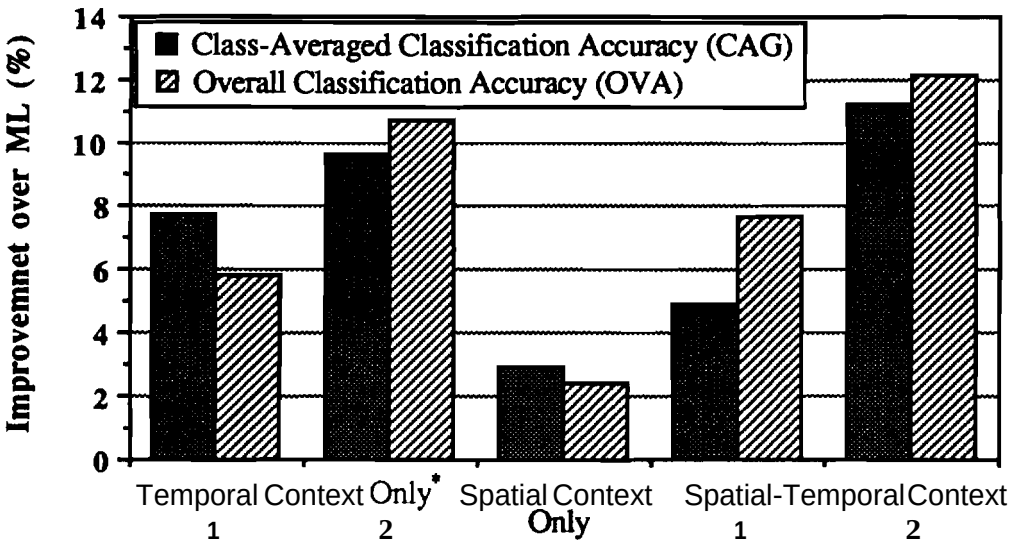


Figure 5.7 Improvement of Classification Accuracy, over a Pixelwise Maximum Likelihood Classifier, by Incorporating Contextual Information for Classifying July Data with April and September data as temporal neighbor sets). 1. Data fusion-based temporal contextual classification (cascade classifier) with eq. (4.1) - serial combination structure.; 2. Maximum likelihood decision fusion-based temporal contextual classification with eq. (4.5.a) - parallel combination structure.; In spatial classification, only spatial class dependency context was used with eq. (3.35).

• In case of data fusion-based classification with the temporal context only, the improvements shown here are based on the best result in terms of overall classification accuracy, which is shown in Table 4.7.

By utilizing both spatial and temporal contextual information, there were classification accuracy improvements as much as 12% over the non-contextual case. The improvements were both for the class-averaged and overall classification accuracies. Spatial context only classification resulted in 2 ~ 3% increase. As seen in previous error maps, the spatial contextual information produced much cleaner classification maps with spatially isolated errors considerably reduced. The classification maps with good delineation of fields will be very useful in may practical application. Temporal contextual information

modeled in term of class transition probabilities was observed to be especially useful in classification. The decision fusion-based combination of multiple data sets was found to perform much better than the data fusion based. It was seen to be insensitive to the inter-relationship of classifiers used for different data sets. This property is expected to be one of the most important requirement for information combination algorithms since it allows independent design and classification of each data set.

5.4 Conclusions

In this chapter, experimental results of the spatial-temporal classifier formulated in Chapter 2 were presented. In addition, with a slight modification, a spatial-temporal contextual classifier under a parallel information combination structure was derived.

The proposed spatial-temporal contextual classifiers exploit the spatial-temporal interpixel class dependency context through spatial prior probabilities and temporal class transition probabilities. The Gibbs random field was used to model the inherent coherence of class labels of spatially adjacent pixels in terms of spatial prior probabilities. Class transition probabilities convey temporal interpixel class dependency context into the classification process.

By allowing the changes of classes over time, it is not necessary to consider the given temporal data sets simultaneously in the training stage and to define additional spectral classes. The number of classes need not be increased even though the number of feature vectors is increased by adding feature vectors of spatio-temporal neighbors. Since this classifier doesn't require processing all the temporal data sets simultaneously, the computational load can be distributed over different times. This classifier is applied to the pixels in a recursive way to yield a computationally efficient contextual classification.

The experiments with three temporal Landsat Thematic Mapper (TM) data sets, taken at April, July, and September, showed significant improvements of classification accuracy over the maximum likelihood non-contextual classification.

In the case of bi-temporal classification of July data with September data, the maximum likelihood decision fusion-based spatial-temporal classifier achieved

classification accuracy increases of about 10.7% in the overall accuracy (OVA) and about **7.8%** in the class-averaged classification accuracy (CAG) over the non-contextual maximum likelihood classification of July data.

In experiments with all three temporal data sets, the spatial-temporal contextual information achieved classification accuracy increases as much as 12% for the overall accuracy (OVA) and 11**%** for the class-averaged classification accuracy (CAG) over the non-contextual maximum likelihood classification of July data. Maximum likelihood decision fusion-based spatial-temporal contextual classifier was found again to be most effective in utilizing spatial-temporal contexts. The resulting classification maps were more meaningful since they had much fewer isolated errors. Classifiers which can utilize potentially important contextual information from spatial, temporal or spatial-temporal neighbors should be quite useful in many real applications, especially where classification accuracy is important.

The degree of usefulness of spatial, temporal or both contextual information in classification may be dependent on data set properties. The spatial class dependency context modeled by a Gibbs random field was found to be very effective in obtaining a more homogeneous class map with much reduced isolated errors. It was not exceptionally computationally demanding under the coding-based recursive approach.

The temporal contextual information based on class transition probabilities was also very useful in improving classification accuracies. When a certain data set is especially effective in extracting a subset of classes, for example, the class wheat in the April data set, its inclusion in the classification process, if properly combined, usually leads to classification accuracy increases. In the experiments, the temporal contextual classification based on decision fusion was observed to have several advantages over the data fusion based counterpart. For example, it would be particularly useful in such cases when the range for posterior probability values are quite different from data set to data set, or when some data sets can not be adequately modeled with statistical probabilities so that posterior probabilities can be computed.

Although all derivations and discussions have been focused on the spatial-temporal contextual classification, as seen in Chapter 4, the contextual classifiers derived in this report can be directly applied to general multisource classification problems. Depending on data set properties, an appropriate classifier, whether it is a spatial classifier or not, can be employed to best utilize the selected data set. The inforimation extracted from each data set can be combined either in a serial or parallel fashion according to the need of application.

5.5 Suggestions for Future Research

The spatial class dependency context was seen to be very effective, but the spatial classifier in eq. (3.35) is based only on local homogeneity of class labels. This spalial class dependency context might not be equally effective in such scenes with relatively small homogeneous fields. In such cases, it needs to be extended to model the general spatial dependency relationship between adjacent class labels. An unsupervised procedure is also necessary for estimating parameters used for such model.

There must be some systematic procedure to decide the data set and classwise reliability or weight factors used for the temporal classifiers in Clhapter 4. Although there are ideas about measuring data set or classwise reliability, assigning specific values for weight factors to be used in classification still remains for further research attention.

When several data sets are combined in classification at the level of data fusion, large differences in the posterior probability ranges often obscure the effect of data set weight factors. Therefore, for the data set weight factors to be fully functional, there must be better way to combine multiple information at the data fusion level.

In temporal classification, the temporal contextual information is conveyed to different data sets using the class transition probabilities, however, there is no estimating procedure available for those. If there are a lot of transition between classes, not only between the same information classes, an unsupervised estimation procedure should be essential.

LIST OF REFERENCES

Benediktsson, J. A. and P. H. Swain, "Consensus Theoretic Classification Methods," IEEE Trans. Systems, Man, and Cybernetics, Vol. 22, No. 4, July/August, pp.688 - 704, 1992

Benediktsson, J. A., P. H. Swain and O. K. Ersoy, "Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data," IEEE Trans. Geosci. Remote Sens., Vol. 28, No. 4, July, pp.540 - 552, 1990

Besag, J., "On the statistical analysis of dirty pictures," J. Royal Statist. Soc. B, Vol. 48, No. 3, pp.259 - 302, 1986

Besag, J., "Spatial Interaction and the Statistical Analysis of Lattice Systems," J. Royal Statist. Soc. Vol. B. 36 pp.192-236, 1974

Chair, Z. and P. K. Varshney, "Optimum Data Fusion in Multiple Sensor Detection Systems," IEEE Trans. Aerospace and Electronic Systems, AES-22, 1986, pp.98 - 101

Chellappa, R., "Stochastic Models in Image Analysis and Processing," Ph.D. Dissertation in School of Electrical Engineering, Purdue University, West Lafayette, Indiana, 1981

Crist, E. P. and W. A. Malia, "A Temporal-Spectral Analysis Technique for Vegetation Application of Landsat," Proceeding of 14th Intern. Symp. on Remote Sensing of Environment, Vol. II, pp.1031 - 1081, 1980

Dattatreya, G. R., "Unsupervised Context Estimation in a Mesh of Pattern Classes for Image Recognition," Pattern Recognition, Vol. 24, No. 7, pp.685 - 694,1991

Derin, H. and P. A. Kelly, "Discrete-Index Markov-Type Random Processes," IEEE Proc., Vol. 77, No. 10, pp.1485-1510, Oct., 1989

Derin, H. and H. Elliot, "Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields," IEEE Trans. Pattern Anal. Machine Intell., Vol. PAMI - 9, No. 1, Jan., pp.39 - 55, 1987

Drake, N. A. *et al.*, "The development of improved algorithms for image processing and classification," Final Report of NERC, Dept. of Geography, University of Reading, U.K., 1987

Eklundh, J. O., H. Yamamoto and A. Rosenfeld, "A Relaxation Method for Multispectral Pixel Classification," IEEE Trans. Pattern Anal. and Mach. Intell., Vol. PAMI-2, No. 1, pp.72-75, Jan., 1980

Fleming, M. D. and R. M. Hoffer, "Computer-Aided Analysis Techniques for an Operational System to Map Forest Lands Utilizing Landsat MSS Data," M.S. Thesis, Purdue University, West Lafayette, Indiana, Dec., 1977

Fukunaga, K., Introduction to Statistical Pattern Recognition, 2nd Edition, Academic Press, New York, 1990

Guo L. J. and J. M. Moore, "Post-Classification Processing for Thematic Mapping Based on Remotely Sensed Image Data," Proceeding of IGARSS 91, pp.2203 - 2206,1991

Haralick, R. M. and H. Joo, "A Context Classifier," IEEE Trans. Geosci. and Remote Sensing, Vol. 24, No. 6, pp.997 - 1007, Nov., 1986

Haralick, R. M., M. C. Zhang and J. B. Campbell, "Multispectral Image Context Classification using the Markov Random Field," Proceeding of IEEE International Conference on Pattern Recognition, pp.190-200, 1984

Haralick, R. M., "Decision making in context," IEEE Trans. Pattern Anal. Machine Intell., Vol. PAMI - 5, No. 4, pp.417 - 428, July, 1983

Haralick, R. M. *et al.*, "Textural features for image classification," IEEE Trans. Systems, Man, and Cybernetics, pp.610 - 621, 1973

Haslett, J., "Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context," Pattern Recognition, Vol. 18, Nos. 3/4, pp.287 - 196, 1985

Hjort, N. L. and E. Mohn, "Topics in the Statistical Analysis of Remotely Sensed Data," Proceedings of the 46th ISI meeting, Tokyo, Sept., IP-21.2, 1987

Hjort, N. L, E. Mohn and G. Strovik, 'Contextual Classification of Remotely Sensed Data Based on Autocorrelation Model," Norwegian Computing Center, Technical Report No. 13, 1985

Hjort, N. L.. and E. Mohn, 'On the Contextual Classification of Data from High Resolution Satellites," Proceedings of 4th Scandinavian Conference on Image Analysis, pp.391-399, June, 1985

Hoballa, I. Y. and P. K. Varshney, "Distributed Bayesian Signal Detection," IEEE Trans. Information Theory, Vol. 35, No. 5, September, 1989, pp.995 - 1000

Hoffer, R. M., M. D. Fleming, L.A. Bartolucci, S. M. Davis, R. F. Nelson, "Digital Processing of Landsat MSS and Topographic Data to Improve Capabilities for Computerized Mapping of Forest Cover Types," LARS Technical Report 011579, Purdue University, West Lafayette, IN 47907, 1979

Hoffer, R. M., "Biological and Physical Considerations in Applying Computer-Aided Analysis Techniques to Remote Sensor Data," Remote Sensing - The Quantitative Approach, edited by P. H. Swain and S. Davis, McGraw-Hill Book Company, New York, 1978

Kalayeh, H. M. and D. A. Landgrebe, "Stochastic Model Utilizing Spectral and Spatial Characteristics" IEEE Trans. Pattern Anal. and Mach. Intell., Vol. PAMI-9, No. 3, pp.457-461, May, 1987

Kalayeh, H. M. and D. A. Landgrebe, "Utilizing Multi-temporal Data by a Stochastic Model," IEEE Trans. Geosci. and Remote Sensing, Vol. 24, No. 5, pp.792 - 795, Sept., 1986

Kalayeh, H. M. and D. A. Landgrebe, "Classification of Remotely Sensed Image Data Using Multitype Information," LARS Technical Report 082782, Purdue University, West Lafayette, IN 47907, 1982

Kashyap, R. L., "Analysis and Synthesis of Image Patterns by Spatial Interaction Models," in Progress in Pattern Recognition, pp.149-186, 1981

Kerkes, J. P. and D. A. Landgrebe, "Simulation of Optical Remote Sensing Systems," IEEE Trans. Geosci. Electron., Vol. 27, No. 6, pp.762-771, 1989

Kettig, R. L. and D. A. Landgrebe, "Classification of multispectral image data by extraction and classification of homogeneous objects," IEEE Trans. Geosci. Eletron., Ge - 14, No. 1, pp.19 - 26, Jan., 1976

Khazenie, N. and M. M. Crawford, "Spatial-Temporal Autocorrelated Model For Contextual Classification," IEEE Trans. Geoscience and Remote Sensing, Vol. 8, No. 4, pp.529 - 539, July, 1990

Kim, Hakil and P. H. Swain, "A method of Classification for Multisource Data in Remote Sensing Based on Interval-Valued Probabilities," Technical Report TR-EE-90-48, July, 1990, School of Electrical Engineering, Purdue University, West Lafayette, IN 47907

Kittler, J., and D. Pairman, "Contextual Pattern Recognition Applied to Cloud Detection and Identification," IEEE Trans. Geoscience and Remote hiensing, Vol. GE-23, No. 6, pp.855 - 863, Nov., 1985

Kittler, J.,, and J. Föglein, "Contextual Classification of Multispectral Pixel Data," Image and Vision Computing, Vol. 2, No. 1, pp.13 - 28, Feb., 1984

Kusaka, T., and Y. Kawata, "Hierarchical Classification of Landsat TM Image using Spectral and Spatial Information," Proceeding of IGARSS 91, pp.2187-2190, 1991

Kusaka, T., H. Egawa and Y. Kawata, "Classification of SPOT Image using Spectral and Spatial Features of Primitive Regions with Nearly Uniform Color," Proceeding of IGARSS 89, pp.649-652, 1989

Landgrebe, D. A., "The Development of a Spectral-Spatial Classifier for Earth Observational Data," Pattern Recognition, Vol. 12, pp.165 - 175, 1980

Landgrebe, D. A., "Useful Information from Multispectral Image Data : Another Look," Remote Sensing - The Quantitative Approach, edited by P. H. Swain and S. Davis, McGraw-Hill Book Company, New York, 1978

Lee, T., J. A. Richards and P. H. Swain, "Probabilistic and evidential approaches for multisource data analysis," IEEE Trans. Geosci. Remote Sens., vol. GE-25, May, pp.283 - 293,1987

Mohn, E, N. L. Hjort and G. O. Storvik, "A Simulation Study of Some Contextual Classification Methods For Remotely Sensed Data," IEEE Trans;. Geosci. and Remote Sensing, Vol. GE-25, No. 6, pp.796 - 804, Nov., 1987

Owen, A.,, "A Neighborhood-Based Classifier for a Landsat Data," Canadian Journal of Statistics, Vol. 12, pp.191-200, 1984

Palubinskas G., "A Comparative Study of Decision Making Algorithms in Images Modeled by Gaussian Markov Random Fields," International Journal of Pattern Recognition and Artificial Intelligence, Vol. 2, No. 4, pp.621-639, 1988

Reibman, A. R. and L. W. Nolte, "Optimal Detection and Performance of Distributed Sensor Systems," IEEE Trans. Aerospace and Electronic Systems, AES-23, 1987, pp.24 - 30

Richards J. A., D. A. Landgrebe and P.H. Swain, "Pixel Labeling by Supervised Probabilistic Relaxation Labeling," IEEE Trans. Pattern Anal. and Mach. Intell., Vol. PAMI-3, pp.188-191, 1981

Rosenfeld A., R. Hummel and S. Zuker, "Scene Labeling by Relaxation Algorithms," IEEE Trans. Systems, Man, and Cybernetics, Vol-SMC-6, pp.420-433, 1976

Sclove S. L., "Pattern Recognition in Image Processing using Interpixel Correlation," IEEE Trans. Pattern Anal. and Mach. Intell., Vol. PAMI-3, pp.206-209, 1981

Swain, P. H., S. B. Vardeman and J. C. Tilton, "Contextual Classification of Multispectral Image Data," Pattern Recognition, Vol-13, No. 6, pp.429-441, 1981

Swain, P. H., "Bayesian Classification in a Time-Varying Environment," IEEE Trans. Systems, Man, and Cybernetics, Vol. 8, No. 12, pp.879 - 883, Dec., 1978a

Swain, P. H., "Fundamentals of Pattern Recognition in Remote Sensing," Remote Sensing - The Quantitative Approach, edited by P. H. Swain and S. Davis, McGraw-Hill Book Company, New York, 1978b

Tang, Z., K. R. Pattipati and D. L. Kleinman, "A Distributed M-ary Hypothesis Testing Problem with Correlated Observations," Proceedings of the 28th IEEE Conference on Decision and Control, Florida, December, 1989, pp.562 - 568

Tenny, R. R. and N. R. Sandell, Jr., "Detection with distributed sensors," IEEE Trans. Aerospace and Electronic Systems, AES-17, July, 1986, pp.501 - 510

Tilton, J. C., S. B. Vardemann, and P. H. Swain, "Estimation of context for statistical classification of multispectral data," IEEE Trans. Geosci. Eletron., Vol. 20, No. 4, pp.445 - 452, 1982

Toussaint, G. T., "The use of context in pattern recognition," Pattern Recognition, Vol. 10, pp.189 - 204, 1978

Townshend, J. R. G., "The Use of Contextual Information in the Classification of Remotely Sensed Data," Photogramm. Eng. Remote Sensing, 49, pp.55 - 64,1983

Tubbs, J. D. and W. O. Alltop, "Measures of Confidence Associated with Combining Classification Results," IEEE Trans. Systems, Man, and Cybernetics, vol. 21, No. 3, May/June, pp.690 - 692, 1991

Warton, S. W., "A contextual classification method for recognizing land use patterns in high resolution remotely sensed data," Pattern Recognition, Vol. 15, No. 4, pp.317 - 324, 1982

Welch, J. R. and K. G. Salter, "A contextual algorithm for pattern recognition and image interpretation," IEEE Trans. Systems, Man, and Cybernetics, pp.610-621, 1973

Yu, T. S. and K. S. Fu, "Recursive contextual classification using a spatial stochastic model," Pattern Recognition, Vol. 16, pp.89-108, 1983

Zenzo S. D., R. Bernstein, S. D. Degloria and H. G. Kolsky, "Gaussian Maximum Likelihood and Contextual Classification Algorithms for Multicrop Classification," IEEE Trans. Geoscience and Remote Sensing, Vol-GE-25, No. 6, Nov., pp.805 - 814, 1987a

Zenzo S. D., S. D. Degloria, R. Bernstein and H. G. Kolsky, "Gaussian Maximum Likelihood and Contextual Classification Algorithms for Multicrop Classification Experiments using Thematic Mapper and Multispectral Scanner Sensor Data," IEEE Trans. Geoscience and Remote Sensing, Vol-GE-25, No. 6, Nov., pp.815 - 824, 1987b

Zhang, M. C. and R. M. Haralick, "Multispectral Image Context Classification Using Stochastic Relaxation," IEEE Trans. Systems, Man, and Cybernetics, Vol. 20, No. 1, pp.128-140, Jan./Feb., 1990

Zhang, Z., H. Shimoda, K. Fukue and T. Sakata, "A New Spatial Classification Algorithm for High Ground Resolution Images," Proceeding of IGARSS 88, pp.509-512, 1988

Appendix A Proofs of Theorems and Lemmas in Chapter 2

The purpose of this appendix is to present formal derivations and proofs of the theorems, lemmas and the **spatio-temporal** contextual classifier addressed in Chapter 2. For an explanation of the notation used, refer to the first part of Chapter 2.

A.l Proofs of Theorems and Lemmas

Since the two assumptions addressed in Chapter 2 are frequently referred in the process of proofs, they are repeated here for easy reference as follows.

Assumption 1.

For any k, $1 \le k \le p$, and for $\mathbf{C}_A$ and $\mathbf{C}_B$ defined below,

$$P\{\mathbf{c}_{k+1} \mid \mathbf{c}_k, \mathbf{C}_A\} = P\{\mathbf{c}_{k+1} \mid \mathbf{c}_k\} \qquad \text{(A.1.a)}$$

$$P\{\mathbf{C}_{S,k} \mid \mathbf{c}_k, \mathbf{C}_B\} = P\{\mathbf{C}_{S,k} \mid \mathbf{c}_k\} \qquad \text{(A.1.b)}$$

where,

$\mathbf{C}_A$ is any non-empty subset of $\xi_{C,k}$ such that $\mathbf{C}_A \cap \{\mathbf{c}_k\} = \phi$. ϕ is an empty set.

$\mathbf{C}_B$ is any non-empty subset of $\xi_{C,k-1}$.

Assumption 2.

For any k, $1 \le k \le p$, and for $\mathbf{X}_A$, $\mathbf{C}_A$, $\mathbf{X}_{others}$ and $\mathbf{C}_{others}$ defined below,

$$P\{\mathbf{X}_A \mid \mathbf{C}_A, \mathbf{X}_{others}, \mathbf{C}_{others}\} = P\{\mathbf{X}_A \mid \mathbf{C}_A\} \qquad \text{(A.2)}$$

where,

$\mathbf{X}_A$ is any non-empty subset of $\mathbf{X}'_{S,k}$.

$\mathbf{C}_A$ is a set of the classes corresponding to $\mathbf{X}_A$.

$\mathbf{X}_{others}$ is any subset of $\xi_{X,p}$ such that $\mathbf{X}_{others} \cap \mathbf{X}'_{S,k} = \phi$.

$\mathbf{C}_{others}$ is any subset of $\xi_{C,p}$ such that $\mathbf{C}_{others} \cap \mathbf{C}'_{S,k} = \phi$.

($\mathbf{C}_{others}$ is not necessarily a set of classes corresponding to $\mathbf{X}_{others}$).

Based on these assumptions, the theorems and lemmas introduced in Chapter 2 can be proved as follows.

***Theorem* 1.**

For any t and u such that $1 \le t \le u \le p$,

$$P\{\eta_u \mid \eta_t, C_{others}\} = P\{\eta_u \mid \eta_t\} = P\{\eta_u \mid c_t\} \qquad (A.3)$$

where,

if $u > t$, η_u is either $\{c_u\}$ or $C'_{s,u}$. η_t is either $\{c_t\}$ or $C'_{s,t}$.

if $u = t$, $\eta_u = C_{s,u}$ and $\eta_t = \{c_t\}$.

C_{others} is any non-empty subset of $\xi_{C,t}$ such that $C_{others} \cap \eta_u = C_{others} \cap \eta_t = \phi$.

Proof of Theorem 1 :

• **When u = t :**

In this case, $\eta_u = C_{s,u}$ and $\eta_t = \{c_t\}$. Since $C_{others} \cap C_{s,u} = C_{others} \cap \{c_t\} = \phi$, note that C_{others} is a non-empty subset of $\xi_{C,t-1}$ and $P\{\eta_u \mid \eta_t, C_{others}\} = P\{C_{s,u} \mid c_t, C_{others}\}$. From the assumption 1 in eq. (A.1.b),

$$P\{C_{s,u} \mid c_t, C_{others}\} = P\{C_{s,u} \mid c_t\} = P\{\eta_u \mid \eta_t\}.$$

• **When u = t + 1 :**

case:

Suppose $\eta_u = \{c_u\} = \{c_{t+1}\}$, *i.e.*, $P\{\eta_u \mid \eta_t, C_{others}\} = P\{c_{t+1} \mid \eta_t, C_{others}\}$.

Since $C_{others} \cap \eta_t = C_{others} \cap \{c_t\} = \phi$, from the assumption 1 in eq. (A.1.a),

$$P\{c_{t+1} \mid \eta_t, C_{others}\} = P\{c_{t+1} \mid c_t\} = P\{c_{t+1} \mid \eta_t\}.$$

case:

Suppose $\eta_u = C'_{s,u} = C'_{s,t+1}$, *i.e.*, $P\{\eta_u \mid \eta_t, C_{others}\} = P\{C'_{s,t+1} \mid \eta_t, C_{others}\}$.

Note $P\{C'_{s,t+1} \mid \eta_t, C_{others}\} = P\{C_{s,t+1} \mid c_{t+1}, \eta_t, C_{others}\} P\{c_{t+1} \mid \eta_t, C_{others}\}$.

Since $C_{others} \cup \eta_t$ is a subset of $\xi_{C,t}$, from the assumption 1 in eq. (A.1.b),

$$P\{C_{s,t+1} \mid c_{t+1}, \eta_t, C_{others}\} = P\{C_{s,t+1} \mid c_{t+1}\} = P\{C_{s,t+1} \mid c_{t+1}, \eta_t\}.$$

In the same way, since $C_{others} \cup [\eta_t - \{c_t\}]$ is a subset of $\xi_{C,t}$ without $\{c_t\}$, from the assumption 1 of eq. (A.1.a),

$$P\{c_{t+1} \mid \eta_t, C_{others}\} = P\{c_{t+1} \mid c_t\} = P\{c_{t+1} \mid \eta_t\}.$$

Therefore,

$$P\{\eta_u \mid \eta_t, C_{others}\} = P\{C_{S,t+1} \mid c_{t+1}, \eta_t\} P\{c_{t+1} \mid \eta_t\} = P\{\eta_u \mid \eta_t\}.$$

When $u > t+1$:

Note that $P\{\eta_{t+k+1} \mid \eta_t, C_{others}\}$ can be written as,

$$P\{\eta_{t+k+1} \mid \eta_t, C_{others}\}$$
$$= \sum_{c_{t+k}} P\{\eta_{t+k+1} \mid c_{t+k}, \eta_t, C_{others}\} P\{c_{t+k} \mid \eta_t, C_{others}\}$$

Suppose eq. (A.3) hold for $u = t + k$, $k \geq 1$, *i.e.*, $P\{\eta_{t+k} \mid \eta_t, C_{others}\} = P\{\eta_{t+k} \mid \eta_t\}$. Then, from this assumption for $u = t + k$,

$$P\{c_{t+k} \mid \eta_t, C_{others}\} = P\{c_{t+k} \mid \eta_t\}.$$

case:

Suppose $\eta_{t+k+1} = \{c_{t+k+1}\}$. From the assumption 1 in eq. (A.1.a),

$$P\{\eta_{t+k+1} \mid c_{t+k}, \eta_t, C_{others}\} = P\{c_{t+k+1} \mid c_{t+k}\} = P\{c_{t+k+1} \mid c_{t+k}, \eta_t\}.$$

Therefore, $P\{\eta_{t+k+1} \mid \eta_t, C_{others}\}$ is computed as,

$$\sum_{c_{t+k}} P\{c_{t+k+1} \mid c_{t+k}, \eta_t\} P\{c_{t+k} \mid \eta_t\} = P\{c_{t+k+1} \mid \eta_t\} = P\{\eta_{t+k+1} \mid \eta_t\}$$

case:

Suppose $\eta_{t+k+1} = C_{S,t+k+1}$, then,

$$P\{\eta_{t+k+1} \mid c_{t+k}, \eta_t, C_{others}\}$$
$$= P\{C_{S,t+k+1} \mid c_{t+k+1}, c_{t+k}, \eta_t, C_{others}\} P\{c_{t+k+1} \mid c_{t+k}, \eta_t, C_{others}\}$$

From the assumption 1 in eq. (A.1.b), $P\{C_{S,t+k+1} \mid c_{t+k+1}, c_{t+k}, \eta_t, C_{others}\}$ is equal to $P\{C_{S,t+k+1} \mid c_{t+k+1}\} = P\{C_{S,t+k+1} \mid c_{t+k+1}, \eta_t\}$ and $P\{c_{t+k+1} \mid c_{t+k}, \eta_t, C_{others}\} = P\{c_{t+k+1} \mid c_{t+k}\} = P\{c_{t+k+1} \mid \eta_t\}$. Therefore,

$$P\{\eta_{t+k+1} \mid c_{t+k}, \eta_t, C_{others}\} = P\{C_{S,t+k+1} \mid c_{t+k+1}, \eta_t\}\, P\{c_{t+k+1} \mid \eta_t\}$$

$$= P\{C'_{S,t+k+1} \mid \eta_t\} = P\{\eta_{t+k+1} \mid \eta_t\}.$$

From case 1, 2, it is proved that if eq. (A.3) holds for $u = t + k$, $k \geq 1$, then it also holds for $u = t + k + 1$. Since eq. (A.3) holds when $k = 1$, by induction, it holds for every t and u such that $1 \leq t \leq u \leq p$.

Proof of second part : $P\{\eta_u \mid \eta_t\} = P\{\eta_u \mid c_t\}$

- When $\eta_t = \{c_t\}$: It is trivial to show $P\{\eta_u \mid \eta_t\} = P\{\eta_u \mid c_t\}$.
- When $\eta_t = C'_{S,t}$: In this case $u \neq t$, $P\{\eta_u \mid \eta_t\} = P\{\eta_u \mid C'_{S,t}\} = P\{\eta_u \mid c_t, C_{S,t}\}$. Since $C_{S,t} \cap \eta_u = C_{S,t} \cap c_t = \phi$, from the result of the first part of Theorem 1,

$$P\{\eta_u \mid c_t, C_{S,t}\} = P\{\eta_u \mid c_t\}.$$

- Q. E. D. -

Lemma 1.

For C_{others}, η_u and η_t defined as in the Theorem 1,

$$P\{C_{others} \mid \eta_u, \eta_t\} = P\{C_{others} \mid \eta_t\} = P\{C_{others} \mid c_t\} \qquad \text{(A.4.a)}$$

$$P\{C_{T,k} \mid c_k, C_{S,k}\} = P\{C_{T,k} \mid c_k\} \qquad \text{(A.4.b)}$$

Proof of eq. (A.4.a) :

Applying Bayes Theorem to the left side of eq. (A.4.a) gives,

$$P\{C_{others} \mid \eta_u, \eta_t\} = P\{\eta_u \mid \eta_t, C_{others}\}P\{\eta_t, C_{others}\} / P\{\eta_u, \eta_t\}.$$

From the Theorem 1, $P\{\eta_u \mid \eta_t, C_{others}\} = P\{\eta_u \mid \eta_t\}$, therefore,

$$P\{\mathbf{C}_{others} \mid \eta_u , \eta_t\} = P\{\eta_t , \mathbf{C}_{others}\} / P\{\eta_t\} = P\{\mathbf{C}_{others} \mid \eta_t\} \qquad (A.5)$$

If η_t is $\{\mathbf{c}_t\}$, it is trivial to show $P\{\mathbf{C}_{others} \mid \eta_t\} = P\{\mathbf{C}_{others} \mid \mathbf{c}_t\}$. If η_t is $\mathbf{C}'_{S,t}$, then, $P\{\mathbf{C}_{others} \mid \mathbf{c}_t\} = P\{\mathbf{C}_{others} \mid \mathbf{c}_t , \mathbf{C}_{S,t}\}$. Applying the result in eq. (A.5) gives, $P\{\mathbf{C}_{others} \mid \mathbf{c}_t , \mathbf{C}_{S,t}\} = P\{\mathbf{C}_{others} \mid \mathbf{c}_t\}$. Therefore, $P\{\mathbf{C}_{others} \mid \eta_u , \eta_t\} = P\{\mathbf{C}_{others} \mid \eta_t\} = P\{\mathbf{C}_{others} \mid \mathbf{c}_t\}$

Proof of eq.(A.4.b) :

Substitute $\mathbf{C}_{others} = \mathbf{C}_{T,k}$, $\eta_t = \{\mathbf{c}_t\}$ and $\eta_u = \mathbf{C}_{S,t}$ in eq. (A.4.a) proves eq. (A.4.b).

- Q. E. D. -

***Theorem* 2.**

For any t and u such that $1 \le t \le u \le p$, and for $\mathbf{X}_A$, η_t and η_u defined as below,

$$P\{\mathbf{X}_A \mid \eta_t, \eta_u\} = P\{\mathbf{X}_A \mid \eta_t\} \qquad (A.6.a)$$

Especially, when $\mathbf{X}_A \cap \mathbf{X}'_{S,t} = \phi$,

$$P\{\mathbf{X}_A \mid \eta_t\} = P\{\mathbf{X}_A \mid \mathbf{c}_t\} \qquad (A.6.b)$$

where,

if $u > t$,

η_t is either $\{\mathbf{c}_t\}$ or $\mathbf{C}'_{S,t}$. η_u is either $\{\mathbf{c}_u\}$ or $\mathbf{C}'_{S,u}$.

$\mathbf{X}_A$ is any non-empty subset of $\xi_{X,t}$ such that $\mathbf{X}_A \cap \mathbf{X}'_{S,t}$ is either ϕ or $\mathbf{X}'_{S,t}$.

if $u = t$,

$\eta_t = \{\mathbf{c}_t\}$ and $\eta_u = \mathbf{C}'_{S,u}$.

$\mathbf{X}_A$ is any non-empty subset of $\xi_{X,t}$.

Proof of Theorem 2 :

Define $\eta_{A,t} = C_A \cap \eta_t$ where C_A is a set of classes corresponding to the pixels in X_A.

case 1 : when $\eta_{A,t} = \phi$.

This implies $X_A \cap X'_{S,t} = \phi$.

$$P\{X_A \mid \eta_t, \eta_u\} = \sum_{C_A} P\{X_A \mid C_A, \eta_t, \eta_u\} P\{C_A \mid \eta_t, \eta_u\}$$

Since $C_A \cap (\eta_t \cup \eta_u) = \phi$, from eq. (A.2), $P\{X_A \mid C_A, \eta_t, \eta_u\} = P\{X_A \mid C_A\}$ and from the lemma 1, $P\{C_A \mid \eta_t, \eta_u\} = P\{C_A \mid \eta_t\}$.

$$P\{X_A \mid \eta_t, \eta_u\} = \sum_{C_A} P\{X_A \mid C_A\} P\{C_A \mid \eta_t\}$$

$$= \sum_{C_A} P\{X_A \mid C_A, \eta_t\} P\{C_A \mid \eta_t\} = P\{X_A \mid \eta_t\} \qquad \text{(A.7.a)}$$

case 2 : when $\eta_{A,t} \neq \phi$.

This implies $u \neq t$ and $\eta_{A,t} = \eta_t$. Let's define $\overline{C}_A = C_A - \eta_{A,t} = C_A - \eta_t$.

when $\overline{C}_A \neq \phi$,

$$P\{X_A \mid \eta_t, \eta_u\} = \sum_{\overline{C}_A} P\{X_A \mid \overline{C}_A, \eta_t, \eta_u\} P\{\overline{C}_A \mid \eta_t, \eta_u\}$$

From the assumption 2, $P\{X_A \mid \overline{C}_A, \eta_t, \eta_u\} = P\{X_A \mid \overline{C}_A, \eta_t\}$ and from the lemma 1, $P\{\overline{C}_A \mid \eta_t, \eta_u\} = P\{\overline{C}_A \mid \eta_t\}$. Therefore, $P\{X_A \mid \eta_t, \eta_u\} = P\{X_A \mid \eta_t\}$.

- when $\overline{C}_A = \phi$, it implies $X_A = X'_{S,t}$ and $\eta_t = C'_{S,t}$. In this case, $P\{X_A \mid \eta_t, \eta_u\} = P\{X'_{S,t} \mid C'_{S,t}, \eta_u\}$ and since $u \neq t$, from the assumption 2, $P\{X'_{S,t} \mid C'_{S,t}, \eta_u\} = P\{X'_{S,t} \mid C'_{S,t}\} = P\{X'_{S,t} \mid \eta_t\}$. Therefore,

$$P\{X_A \mid \eta_t, \eta_u\} = P\{X'_{S,t} \mid \eta_t\} = P\{X_A \mid \eta_t\} \qquad \text{(A.7.b)}$$

With eq. **(A.7.a)** and (A.7. b), eq. (A.6.a) is proved.

proof of eq. (A.6.b) :

- When $\eta_t = \{c_t\}$, it is trivial to show $P\{X_A \mid \eta_t\} = P\{X_A \mid c_t\}$.
- When $\eta_t = C'_{S,t}$, $P\{X_A \mid \eta_t\} = P\{X_A \mid c_t, C_{S,t}\}$ and from eq. (A.7.a), $P\{X_A \mid c_t, C_{S,t}\}$ $= P\{X_A \mid c_t\}$. Therefore, $P\{X_A \mid \eta_t\} = P\{X_A \mid c_t\}$.

- Q.E. D. -

Lemma 2.

$$P\{X_{T,k} \mid c_k, c_{k+1}\} = P\{X_{T,k} \mid c_k\} \quad \text{(A.8.a)}$$

$$P\{X_{T,k} \mid c_k, C_{S,k}\} = P\{X_{T,k} \mid c_k\} \quad \text{(A.8.b)}$$

$$P\{X'_{S,k} \mid c_k, c_{k+1}\} = P\{X'_{S,k} \mid c_k\} \quad \text{(A.8.c)}$$

proof of Lemma 2 :

Substituting $X_A = X_{T,k}$, $\eta_t = \{c_k\}$ and $\eta_u = \{c_{k+1}\}$ in eq. (A.6.a) proves eq. (A.8.a).
substituting $X_A = X_{T,k}$, $\eta_t = \{c_k\}$ and $\eta_u = C_{S,k}$ in eq. (A.6.a) proves eq. (A.8.b).
Substituting $X_A = X'_{S,k}$, $\eta_t = \{c_k\}$ and $\eta_u = \{c_{k+1}\}$ in eq. (A.6.a) proves eq. (A.8.a).

Lemma 3.

For any k, $1 \le k \le p$ and for X_{others} which is any non-empty subset of $\xi_{X\,k-1}$,

$$P\{X'_{S,k} \mid c_k, X_{others}\} = P\{X'_{S,k} \mid c_k\} \quad \text{(A.9)}$$

proof of Lemma 3:

Note that the left -hand side of eq. (A.9) can be written as,

$$P\{X'_{S,k} \mid c_k, X_{others}\} = \sum_{C_{S,k}} P\{X'_{S,k} \mid C'_{S,k}, X_{others}\} P\{C_{S,k} \mid c_k, X_{others}\} \quad \text{(A.10)}$$

From the assumption 2, $P\{X'_{S,k} \mid C'_{S,k}, X_{others}\} = P\{X'_{S,k} \mid C'_{S,k}\}$. By the Bayes rule,

$$P\{C_{S,k} \mid c_k, X_{others}\} = P\{X_{others} \mid C_{S,k}, c_k\} P\{C_{S,k} \mid c_k\} / P\{X_{others} \mid c_k\}$$

Note that,

$$P\{X_{others} \mid C_{S,k}, c_k\} = \sum_{C_{S,k}} P\{X_{others} \mid C_{others}, C_{S,k}, c_k\} P\{C_{others} \mid C_{S,k}, c_k\}$$

and from the assumption 2, $P\{X_{others} \mid C_{others}, C_{S,k}, c_k\} = P\{X_{others} \mid C_{others}\} = P\{X_{others} \mid C_{others}, c_k\}$. According to the lemma 1, $P\{C_{others} \mid C_{S,k}, c_k\} = P\{C_{others} \mid c_k\}$. Therefore, $P\{C_{S,k} \mid X_{others}, c_k\} = P\{C_{S,k} \mid c_k\}$ and,

$$P\{X_{others} \mid C_{S,t}, c_t\} = \sum_{C_{S,t}} P\{X_{others} \mid C_{others}, c_t\} P\{C_{others} \mid c_t\} = P\{X_{others} \mid c_k\}$$

From these results, eq. (A.lO) is

$$P\{X'_{S,k} \mid c_k, X_{others}\} = \sum_{C_{S,k}} P\{X'_{S,k} \mid C'_{S,k}\} P\{C_{S,k} \mid c_k\} = P\{X'_{S,k} \mid c_k\}$$

- Q. E. D. -

A.2 Derivation of Spatio-Temporal Contextual Classifier

In this section, the spatio-temporal contextual classifier given below will be simplified using the properties derived in the previous section. For $k = 2, \cdots, p$ and $c \in \Omega_k$, the spatio-temporal contextual classifier is defined as,

$$H_{SPTP}(c; r, k) = P\{c_k = c \mid x_k = x_k, X_{S,k} = X_{S,k}, X_{T,k} = X_{T,k}\} \qquad (A.11.a)$$

Specially, if $k = 1$, $H_{SPTP}(\cdot)$ is defined as,

$$H_{SPTP}(c; r, k) = P\{c_k = c \mid x_k = x_k, X_{S,k} = X_{S,k}\}, c \in \Omega_k \qquad \text{(A.11.b)}$$

Applying the Bayes rule when k = 2, ---, p, results in,

$$H_{SPTP}(c; r, k) \equiv P\{c_k \mid x_k, X_{S,k}, X_{T,k}\} = P\{c_k \mid X'_{S,k}, X_{T,k}\}$$

The probability $P\{c_k \mid X'_{S,k}, X_{T,k}\}$ can be written as,

$$P\{c_k \mid X'_{S,k}, X_{T,k}\} = \frac{P\{X'_{S,k}, X_{T,k}, c_k\}}{P\{X'_{S,k}, X_{T,k}\}}$$

$$= \frac{P\{X'_{S,k} \mid X_{T,k}, c_k\} P\{X_{T,k} \mid c_k\} P\{c_k\}}{P\{X'_{S,k}, X_{T,k}\}} \qquad \text{(A.12)}$$

Notice, from the lemma 3, (*ie.*, $X_{others} = X_{T,k}$ in eq. (A.9)),

$$P\{X'_{S,k} \mid X_{T,k}, c_k\} = P\{X'_{S,k} \mid c_k\}$$

By using the Bayes theorem,

$$P\{X'_{S,k} \mid c_k\} = P\{c_k \mid X'_{S,k}\} P\{X'_{S,k}\} / P\{c_k\}$$

$$P\{X_{T,k} \mid c_k\} = P\{c_k \mid X_{T,k}\} P\{X_{T,k}\} / P\{c_k\}$$

therefore,

$$P\{c_k \mid X'_{S,k}, X_{T,k}\} = \frac{P\{X'_{S,k}\} P\{X_{T,k}\}}{P\{X'_{S,k}, X_{T,k}\}} \cdot \frac{P\{c_k \mid X'_{S,k}\} P\{c_k \mid X_{T,k}\}}{P\{c_k\}}$$

Let's define $H_{SP}(c; r, k)$, $H_{TP}(c; r, k)$, and A_k, for $c \in \Omega_k$ and $k = 2, ---, p$ as follows.

$$H_{SP}(c; r, k) \equiv P\{c_k = c \mid X'_{S,k} = x'_{S,k}\}$$

$$H_{TP}(c; r, k) \equiv P\{c_k = c \mid X_{T,k} = x_{T,k}\}$$

$$A_k \equiv \frac{P\{X'_{S,k}\}P\{X_{T,k}\}}{P\{X'_{S,k}, X_{T,k}\}}$$

Then, $H_{SPTP}(c; r, k)$ can be written for $c \in \Omega_k$, $k = 2, ---, p$, as,

$$H_{SPTP}(c; r, k) = A_k \frac{H_{SP}(c; r, k)\ H_{TP}(c; r, k)}{P\{c_k = c\}} \qquad \text{(A.13)}$$

The temporal contextual classifier, $H_{TP}(c; r, k)$ can be computed using its previous spatio-temporal contextual part. According to Bayes theorem,

$$\begin{aligned} H_{TP}(c; r, k) &= P\{c_k = c \mid X_{T,k}\} \\ &= P\{X_{T,k} \mid c_k = c\}P\{c_k = c\} / P\{X_{T,k}\} \\ &= \frac{P\{c_k = c\}}{P\{X_{T,k}\}} \sum_{c_{k-1}} P\{X_{T,k} \mid c_k = c, c_{k-1}\}P\{c_{k-1} \mid c_k = c\} \end{aligned} \qquad \text{(A.14)}$$

From the theorem 2 (*i.e.*, substituting, $X_A = X_{T,k} = \xi_{X,k}$, $\eta_t = c_{k-1}$ and $\eta_u = c_k$), the probability $P\{X_{T\,k} \mid c_k, c_{k-1}\}$ is equal to $P\{X_{T\,k} \mid c_{k-1}\}$ which can be computed as,

$$P\{X_{T,k} \mid c_{k-1}\} = P\{c_{k-1} \mid X'_{S,k-1}, X_{T,k-1}\}P\{X'_{S,k-1}, X_{T,k-1}\} / P\{c_{k-1}\}$$

Notice also that $P\{c_{k-1} \mid X'_{S,k-1}, X_{T,k-1}\}$ can be written as $H_{SPTP}(c_{k-1}; r, k-1)$ and $P\{X'_{S,k-1}, X_{T,k-1}\} = P\{X_{T,k}\}$; substituting these yields,

$$P\{X_{T,k} \mid c_{k-1}\} = \frac{P\{X_{T,k}\}}{P\{c_{k-1}\}} H_{SPTP}(c_{k-1}; r, k-1)$$

The temporal contextual classifier part, $H_{TP}(c; r, k)$ in eq. (A.14) is now written as,

$$H_{TP}(c; r, k) = P\{c_k = c\} \sum_{c_{k-1} = d} H_{SPTP}(c_{k-1} = d; r, k-1) \frac{P\{c_{k-1} = d \mid c_k = c\}}{P\{c_{k-1} = d\}}$$

Applying the Bayes theorem yields,

$$P\{c_k = c\} \frac{P\{c_{k-1} = d \mid c_k = c\}}{P\{c_{k-1} = d\}} = P\{c_k = c \mid c_{k-1} = d\}$$

therefore, $H_{TP}(c; r, k)$ can be computed as,

$$H_{TP}(c; r, k) = \sum_{d \in \Omega_{k-1}} H_{SPTP}(d; r, k-1)\, P\{c_k = c \mid c_{k-1} = d\}$$

In a summary of previous results, for $k = 2, \cdots, p$, and $c \in \Omega_k$,

$$H_{SPTP}(c; r, k) = A_k \frac{H_{SP}(c; r, k)\, H_{TP}(c; r, k)}{P\{c_k = c\}}$$

where,

$$H_{SP}(c; r, k) \equiv P\{c_k = c \mid X'_{S,k} = X'_{S,k}\}$$

$$H_{TP}(c; r, k) \equiv P\{c_k = c \mid X_{T,k} = X_{T,k}\}$$

and,

$$H_{TP}(c; r, k) = \sum_{d \in \Omega_{k-1}} H_{SPTP}(d; r, k-1)\, P\{\mathbf{c}_k = c \mid \mathbf{c}_{k-1} = d\}$$

In case $k = 1$,

$$H_{SPTP}(c; r, k) = P\{\mathbf{c}_k = c \mid \mathbf{x}_k = x_k, \mathbf{X}_{S,k} = X_{S,k}\} = H_{SP}(c; r, k)$$

This concludes the derivation of the spatio-temporal contextual classifier.

Appendix B Program List for Spatial-Temporal Classification

Program list for the spatial and temporal classifiers discussed in this report is available upon request.