

**Adaptive Bayesian Contextual Classification
Based on Markov Random Fields**

Qiong Jackson, *Student IEEE*, David Landgrebe, *Life Fellow, IEEE*
School of Electrical & Computer Engineering
Purdue University

Copyright © 2002 IEEE. Reprinted from the *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 40, No. 11, pp 2454-2463, November, 2002.

This material is posted here with permission of the IEEE. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by sending a blank email message to pubs-permissions@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Abstract-In this paper an Adaptive Bayesian Contextual classification procedure that utilizes both spectral and spatial interpixel dependency contexts in estimation of statistics and classification is proposed. Essentially, this classifier is the constructive coupling of an adaptive classification procedure and a Bayesian contextual classification procedure. In this classifier, the joint prior probabilities of the classes of each pixel and its spatial neighbors are modeled by the Markov Random Field. The estimation of statistics and classification are performed in a recursive manner to allow the establishment of the positive feedback process in a computationally efficient manner. Experiments with real hyperspectral data show that, starting with a small training sample set, this classifier can reach classification accuracies similar to that obtained by a pixelwise MLC with a very large training sample set. Additionally, classification maps are produced which have significantly less speckle error.

Index Terms-Adaptive iterative classification procedure, Bayesian contextual classification procedure, iterative conditional mode (ICM), hyperspectral data, semi-labeled samples

I. INTRODUCTION

Hyperspectral image data acquired by new generation sensors contain extremely rich spectral attributes, which offer the potential to discriminate more detailed classes with high classification accuracy using a conventional Maximum Likelihood Pixel Classifier (MLC). However, two difficulties inhibit this potential. First of all, the large number of classes of interest combined with the large number of spectral bands available requires a large number of training samples. Unfortunately training samples are generally expensive and tedious to obtain. As a result, generally the class statistics have to be estimated with a limited training sample set. Hence the estimated class statistics are less accurate and the subsequent classifier performance is less accurate than need be. Additionally, in a conventional MLC, it is explicitly assumed that the spectral properties of a pixel are independent of the spectral properties of its adjacent pixels. The MLC has difficulty distinguishing the pixels that come from different land-cover classes but have very similar spectral properties. The result is often a snow-like classification map.

Since, in general, certain ground cover classes may be more likely to be placed adjacently than others, there is more than trivial information available from the relative assignments of the classes of neighboring pixels. Also, in many remotely sensed images, objects on the ground are much greater than the pixel element size so neighboring pixels are more likely to come from the same class and form a homogeneous region. Therefore, a supervised contextual classifier that utilizes both spectral and spatial contextual information may be able to better discriminate between the pixels with similar spectral attributes but located in different regions. This should allow reduction of the speckle error

and improve the classification performance significantly. However, this type of classifier also faces the problem of the small training sample size where the class conditional probability has to be estimated in the analysis of hyperspectral data.

In [1], it has been demonstrated that an adaptive pixel maximum likelihood classifier (MLC) may alleviate the small training sample problem by including semi-labeled samples along with the training samples during the process of estimation of statistics. As illustrated in Figure 1, essentially it is formed by adding a feedback loop (highlighted by dark arrows) to a conventional ML classifier. The classifier starts with the initial classification where only training samples are used to estimate the statistics. The initial classification assigns a tentative class label to each unlabeled sample according to the ML decision rule, and thus unlabeled samples become semi-labeled samples because class label information is partially obtained. At the following iteration, semi-labeled samples together with the training samples are used to re-estimate the statistics. To control the influence of each semi-labeled sample for subsequent estimation of statistics, full weight is assigned to a training sample and reduced weight is assigned to semi-labeled samples. The key to successful performance of this classifier is to establish a positive feedback process wherein, during each iteration, the estimation of statistics can be improved based on the higher classification accuracy of the previous iteration. In return, much higher classification accuracy can be achieved as the iteration process proceeds. We have shown in [1] that the higher accuracy at each iteration and a large number of semi-labeled samples can allow the establishment of this positive feedback and lead to rapid convergence of classification accuracy. However, as with a

conventional MLC, performance of this adaptive pixel MLC is limited by using just spectral information.

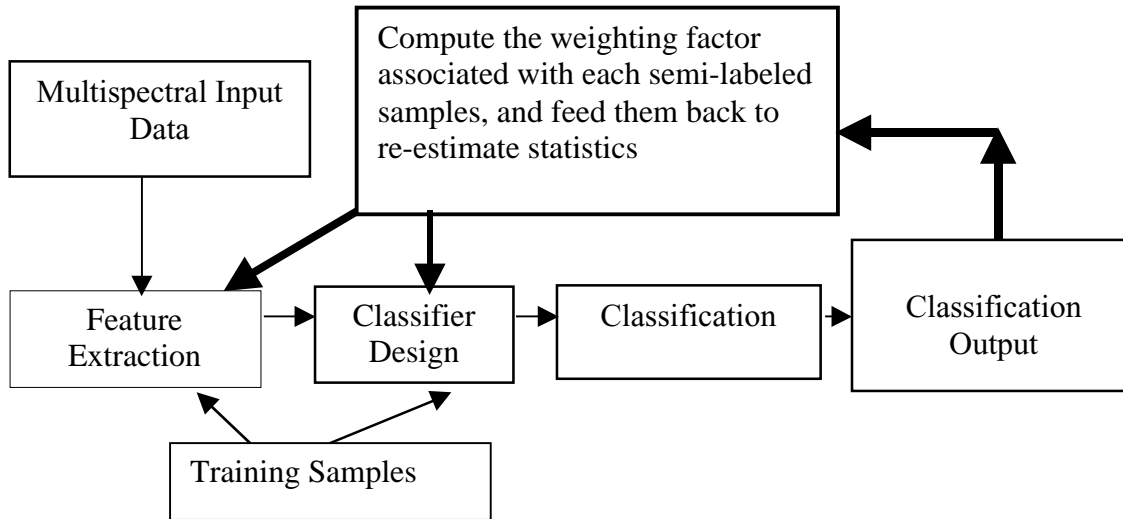


Fig. 1 An adaptive classification procedure

In this paper, an adaptive Bayesian contextual classifier that utilizes both spectral and spatial interpixel dependency contexts in estimation of statistics and classification is proposed. Essentially, the proposed classifier is the combination of a Bayesian contextual classifier and an adaptive classification procedure. In this classifier, only interpixel class dependency context is considered, and the joint prior probabilities of the classes of each pixel and its spatial neighbors are modeled by the Markov Random Field. As an adaptive classification procedure, the estimation of statistics and classification are performed in a recursive manner. Because usually a contextual classifier achieves higher accuracy than a pixelwise MLC, the proposed classifier has several advantages over the adaptive MLC. First of all, the positive feedback might be easier to establish. Secondly, it might converge faster. Third, the final accuracy might be higher with much less speckle errors.

Compared with a conventional one-pass contextual classification, this approach should mitigate the small training sample problem in the analysis of hyperspectral data.

This paper is organized as follows. In the following section, the Bayesian contextual formulation of an image based on a Markov Random Field (MRF) model is reviewed. One of the methods to solve it, called Iterative Conditional Mode (ICM)[2], is then presented. In section III, the proposed combination of adaptive classification with the ICM [2] is defined. The experimental results of this proposed method with real hyperspectral data are presented in section IV. In the final section, conclusions are summarized.

II. BAYESIAN FORMULATION AND THE ICM

Multivariate image \mathbf{X} is composed of p -dimensional pixels where $\mathbf{X}_k(s)$, and $\{k=1, 2, \dots, p\}$, and $s=(i,j)$ denotes a two-dimensional index, an image lattice point at the i^{th} row and j^{th} column. Let u denote the field that contains the classification of each pixel in \mathbf{X} . Points in u can take values in the set $\{1, 2, \dots, L\}$, where L is the number of classes. The multivariate image \mathbf{X} is then classified by finding a field of class labels \hat{u}_{MAP} such that

$$u_{MAP} = \arg \max_u \{p(u | X)\} = \arg \max_u \{p(X | u)p(u)\} \quad (1)$$

where \hat{u}_{MAP} is referred to as a MAP estimate of the field of class labels which maximizes the posterior probability in Eq. (1). Therefore, the modeling of both the prior probability distribution $p(u)$ and class-conditional distribution $p(\mathbf{X}/u)$ becomes an essential task. Note

that the estimate Eq.(1) becomes the pixel-wise noncontextual classifier if the prior probability does not have any contextual consequence in formulating Eq.(1).

In most image lattice problems, available information stems from two different sources: observation on image sites for a given occurrence of the problem, and a priori knowledge about the restrictions imposed on the simultaneous labeling of connected neighboring units. This second source of information reflects statistical dependencies between the labels of neighboring sites. Markov random field (MRF) theory [2, 3, 4, 5, 6] provides a convenient and consistent way to model such context-dependent information. The MRF s-Gibbs equivalence, established by Hammersley and Clifford , and further developed by Besag [3], gives an explicit formula for the joint distribution of MRF s.

For a Markov random field u , the conditional distribution of a point in the field, given all other points, is only dependent on its neighbors: $p\{u(s) | u(S - s)\} = p\{u(s) | u(\partial s)\}$. Here \mathbf{S} is an image lattice and $\mathbf{S} - s$ denotes a set of points in \mathbf{S} excluding s , ∂s denotes the neighboring pixels of s . The first order neighborhood system is usually defined as the four pixels surrounding a given pixel, and higher orders are defined by adding corner pixels to a lower order neighborhood system. A clique is defined as a subset of points in \mathbf{S} such that if s and r are two points contained in a clique c , then s and r are neighbors, and the order of a clique is the number of points (sites) in the clique. The neighborhood system and the corresponding cliques are illustrated in Figure 2.

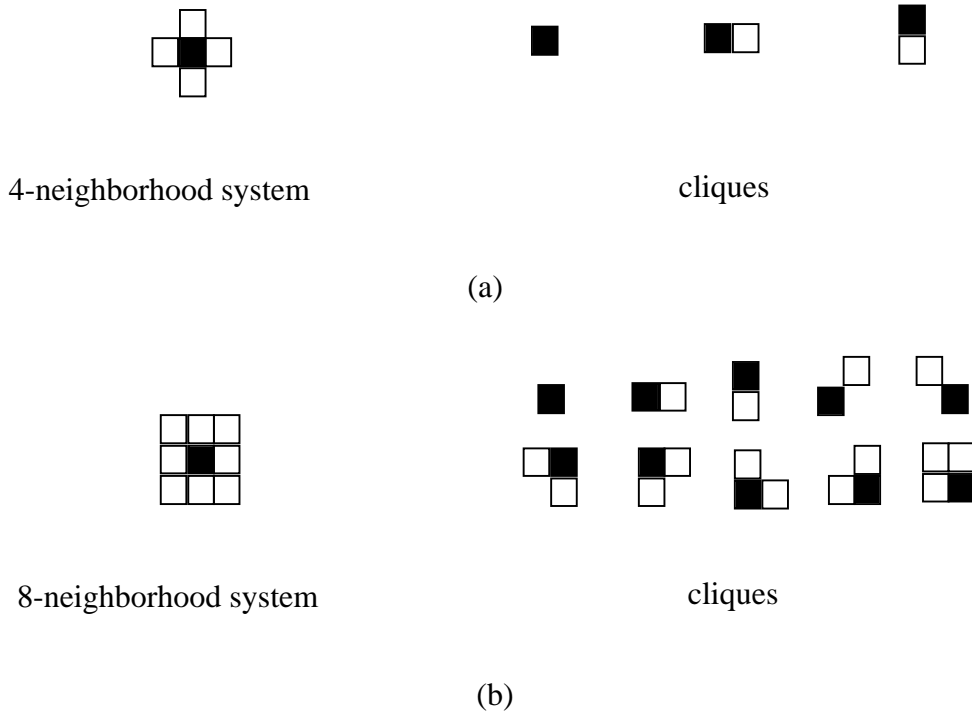


Fig. 2 Neighborhoods system and corresponding cliques

The a priori probability of the labeling $p(u)$ defines an MRF. According to the Hammersley-Clifford theorem, for a given neighbor system, $p(u)$ can be expressed as a Gibbs distribution:

$$p(u) = \frac{1}{Z} \exp[-\sum_c V^c(u)] \quad (2)$$

where Z is a normalizing constant called a partition coefficient, and V^c is an arbitrary function of u on the clique c . C is defined as the set of all cliques.

Together with the joint class-conditional distribution $p\{\mathbf{X}|u\}$ and prior distribution of Eq.(2), the MAP estimates of true class labels as given by Eq. (1) becomes:

$$\hat{u}_{MAP} = \arg \min_u \{-\ln p(X | u) + \sum_C V^c(u)\} \quad (3)$$

The minimization of (3) is essential in order to derive a MAP estimate of u , \hat{u}_{MAP} . In [7], it is pointed out that the one-dimensional dynamic programming in [8] or simulated annealing method in [4] are computational expensive, and the global minimization still suffers from falling into a local minimum. In [2], a method called ICM is developed to approximate \hat{u}_{MAP} using assumptions to reduce the computational complexity. Instead of attempting to optimize in one step by the above-suggested methods, the ICM is computationally feasible since it updates the class assignments iteratively so that inverting a large matrix is avoided. To apply the ICM method, Eq. (1) is modified to conform to the task based on two main assumptions, which are:

(1) Each pixel value is class-conditionally independent, such that:

$$p(X | u) = \prod_{s=1}^N p\{X(s) | u(s)\}$$

where N is the total number of pixels in the image.

(2) The class labels are the realization of a Markov random field, and their probability mass functions are identical, i.e.,

$$p(u(s) | u(S - s)) = p\{u(s) | u(\partial s)\}$$

Suppose that the objective is to estimate the class label of a pixel given the estimates of class labels for all other pixels inside the rectangular lattice. Then the optimization of Eq.(1) becomes:

$$\hat{u}(s) = \arg \max_u \{p(u(s) | X, \hat{u}(S - s))\} \quad (4)$$

Note that $u(s)$ denotes a class label at $s \in \mathbf{S}$. Applying the Bayes rule and considering the Markov property of (2), the argument of Eq.(4) becomes

$$p\{u(s) | X, u(S-s)\} \propto p\{X | u(s), u(S-s)\} p\{u(s) | u(\partial s)\} \quad (5)$$

The first term of the right hand side of Eq. (5) becomes

$$p\{X | u(s), u(S-s)\} = p\{X(s) | u(s)\} p\{X(S-s) | u(S-s)\} \quad (6)$$

by virtue of the assumption (1). Since the class assignment of all other pixels except $u(s)$ inside the lattice are already made, the term $p\{X(S-s) | \hat{u}(S-s)\}$ is not a factor affecting the optimization. Therefore, Eq. (4) in connection with Eqs.(5) and (6) becomes

$$\begin{aligned} u(s) &= \arg \max_{u(s)} \{p(u(s) | X, u(S-s))\} \\ &= \arg \max_{u(s)} [p\{X | u(s), u(S-s)\} p\{u(s) | u(S-s)\}] \\ &= \arg \max_{u(s)} [p\{X(s) | u(s)\} p\{u(s) | u(\partial s)\}] \end{aligned} \quad (7)$$

Assume the class conditional distribution can be represented by a Gaussian distribution,

$$p\{X(s) | u(s)\} = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_u|^{\frac{1}{2}}} \exp[-\frac{1}{2} \{(X(s) - \mu_u)^T \Sigma_u^{-1} (X(s) - \mu_u)\}] \quad (8)$$

where the μ_u , Σ_u are the mean vector and covariance matrix for the class u , respectively.

Concerning energies of cliques of order 2 (2-point clique) and restricting to 4-neighborhood system, for the sake of mathematical and computational convenience, most MRF image models are assumed to be homogeneous and isotropic. Then V^c is independent of the location of clique c in \mathbf{S} and independent of the orientation of c . Under these assumptions, the M-Level MRF model is frequently used for an image segmentation problem [11]:

$$V(u(s), u(s')) = \begin{cases} 0 & \text{if } u(s) = u(s') \\ \beta & \text{otherwise} \end{cases}$$

where β is a constant coefficient, which can be estimated from the image or empirically determined. It is a weight emphasizing the significance of interaction among adjacent pixels inside a clique. Therefore, the class conditional probability mass function of $p\{u(s) | \hat{u}(\partial s)\}$ becomes

$$p\{u(s) | \hat{u}(\partial s)\} = \frac{1}{Z} \exp[-\beta \sum_{s' \in c(s)} \{1 - \delta(u(s) - u(s'))\}]$$

Then Eq.(7) is equivalent to:

$$\begin{aligned} u(s) &= \arg \max_{1 \leq u \leq L} [p\{X(s) | u(s)\} p\{u(s) | u(\partial s)\}] \\ &= \arg \min_{1 \leq u \leq L} [-\ln p\{X(s) | u(s)\} - \ln p\{u(s) | u(\partial s)\}] \\ &= \arg \min_{1 \leq u \leq L} [\ln |\Sigma_u| + (X(s) - \mu_u)^T \Sigma_u^{-1} (X(s) - \mu_u) + 2m\beta + const.] \end{aligned} \quad (9)$$

Here, m is the number of occurrences of the class different from $u(s)$ in the clique containing s . The term const. doesn't depend on the particular class assignment to the pixels. Essentially, by starting with the initial ML classification outputs, ICM [2] solves Eq. (9) repeatedly until convergence is reached where the class label doesn't change much.

III ADAPTIVE BAYESIAN CONTEXTUAL CLASSIFIER: THE COMBINATION OF AN ADAPTIVE CLASSIFIER WITH BAYESIAN CONTEXTUAL ITERATION CONDITIONAL MODES (ICM)

In this section, the new adaptive Bayesian contextual classifier is developed that combines the adaptive procedure proposed in [1] with the Bayesian Contextual Iteration Conditional Modes (ICM) [2]. In this new classifier, contextual information is incorporated into the process of a weighting factor computation and classification. There

are two reasons for this operation. One is to further emphasize the positive effect from the correctly classified semi-labeled samples and discourage the negative influence from the misclassified semi-labeled ones, and the second is to enhance the classification using contextual information in addition to the likelihood. In a manner similar to the adaptive procedure and ICM, this new method is also an iterative process. It starts with initial parameter estimates ϕ^0 (including the mean vectors and covariance matrices for all classes) using training samples only and then uses them to perform the initial classification. Based on the initial classification, it repeats the estimation of statistics and classification at each iteration using training samples and semi-labeled samples until convergence is reached.

Assume the initial class conditional statistics and classification has been obtained by using the training samples, and all L classes can be represented by Gaussian distributions. Denote $y = (y_{i1}, \dots, y_{im_i})$ as the training samples for the i^{th} class, whose pdf is $f_i(y/\phi_i)$, and $x = (x_{i1}, \dots, x_{in_i})$ are the semi-labeled samples that have been classified to the i^{th} class. Here m_i is the number of training samples for i^{th} class, and n_i is the number of semi-labeled samples classified to the i^{th} class, and ϕ_i represents the set of parameters for the i^{th} class.

The procedure for this method is defined as follows:

Cycle 1 (Initial Cycle)

- 1) Use only **training samples** to estimate statistics, and then perform classification using a ML classifier

2a) Perform classification using a MAP classifier based on the classification map from the ML:

$$X(s) \in u \Leftrightarrow u(s) = \arg \min_{1 \leq u \leq L} [\ln |\Sigma_u| + (X(s) - \mu_u)^T \Sigma_u^{-1} (X(s) - \mu_u) + 2m\beta] \quad (10)$$

where β is empirically determined.

2b) Perform classification using a postprocessing classifier based on the classification map from the ML

$$X(s) \in u \Leftrightarrow u(s) = \arg \max_{u(s)} [p\{u(s) | u(\partial s)\}] = \arg \min_{1 \leq u(s) \leq L} [2m] \quad (11)$$

The purpose of using the postprocessing classifier is to compare the results from the MAP classifier

Cycle 2:

1) Compute weighting factors using **contextual information** together with the **likelihood** based on the classification results from the MAP classifier in step (2a) from the previous cycle

$$w_{ij}^c = \frac{p(x_{ij} | \phi_u^c) p(u(s) | u(\partial s))}{\sum_{k=1}^L p(x_{ij} | \phi_k^c) p(k(s) | k(\partial s))} \quad (12)$$

Note that unit weight is assigned to each training sample.

2) Obtain the class conditional statistics by maximizing the mixed log likelihood of **training samples** and of **semi-labeled samples**, which are obtained from the MAP classifier in step (2a) from the previous cycle.

$$\mu_i^+ = \frac{\sum_{j=1}^{m_i} y_{ij} + \sum_{j=1}^{n_i} w_{ij}^c x_{ij}}{m_i + \sum_{j=1}^{n_i} w_{ij}^c} \quad (13a)$$

$$\Sigma_i^+ = \frac{\sum_{j=1}^{m_i} (y_{ij} - \mu_i^+)(y_{ij} - \mu_i^+)^T + \sum_{j=1}^{n_i} w_{ij}^c (x_{ij} - \mu_i^+)(x_{ij} - \mu_i^+)^T}{m_i + \sum_{j=1}^{n_i} w_{ij}^c} \quad (13b)$$

Note that as indicated in Eq. (13a) and Eq. (13b), the estimated statistics are affected by training samples and semi-labeled samples.

3) Perform classification based on the maximum likelihood (ML) classification rule:

$$X(s) \in u \Leftrightarrow u(s) = \arg \min_{1 \leq u \leq L} [\ln |\Sigma_u^+| + (X(s) - \mu_u^+)^T (\Sigma_u^+)^{-1} (X(s) - \mu_u^+)] \quad (14)$$

4a) Perform classification using the MAP classifier based on the classification map from the MLC (step 3):

$$X(s) \in u \Leftrightarrow u(s) = \arg \min_{1 \leq u \leq L} [\ln |\Sigma_u^+| + (X(s) - \mu_u^+)^T (\Sigma_u^+)^{-1} (X(s) - \mu_u^+) + 2m\beta] \quad (15)$$

4b) Perform classification using the postprocessing classifier based on the classification map from the MLC (step 3)

$$X(s) \in u \Leftrightarrow u(s) = \arg \max_{u(s)} [p\{u(s) | u(\partial s)\}] = \arg \min_{1 \leq u(s) \leq L} [2m]$$

The steps of the cycle 2 are repeated until convergence is reached where the classification results have small change. The flow chart in Fig. 3 illustrates one complete cycle of the adaptive contextual classifier.

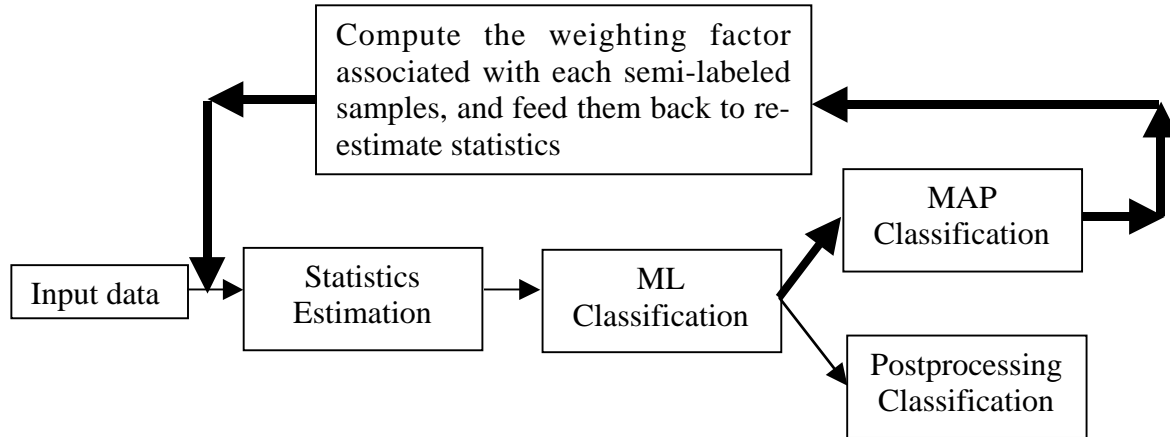


Fig. 3 One complete cycle of the Adaptive Bayesian Contextual Classifier

Note that as an adaptive pixel-wise ML classifier, in this Adaptive Bayesian Contextual classification procedure, the label of each semi-labeled sample is updated after each classification, including ML, MAP, and postprocessing classification at each cycle, and the weight of each semi-labeled sample is updated after each cycle.

Correspondingly, the class conditional statistics are updated at each cycle as well. For notational purposes, in the following, the ML, MAP and Postprocessing classifiers at each cycle are denoted as ABC-ML, ABC-MAP and ABC-Postprocessing classifiers, respectively.

Several modifications have been made in this new classifier. First, the contextual information in addition to likelihood is utilized to enhance the performance of semi-labels in terms of their influence of class conditional estimation of statistics and to improve the classifier performance. Second, the semi-labeled samples generated from the ABC-MAP classifier, instead of the ABC-ML classifier at the previous cycle in addition to training samples, are used to update the current class conditional statistics. Third, each cycle is

started with the ABC-ML classifier instead of the ABC-MAP classifier as a conventional ICM does.

The reason for the third modification is as follows. First, it has been shown that in the ICM starting with the classification results from a ML classifier, in general the MAP classifier outperforms the ML classifier [9, 10]. Even though a postprocessing classifier may be able to improve classification accuracy also by reducing the speckle error, it is more likely to be overdone and lead to loss of more details than using the ABC-MAP classifier. In other words, semi-labeled samples generated from the ABC-MAP classifier should contain more correctly classified samples. Because the accuracy of statistics estimation is strongly related the accuracy of classified samples, better estimation of statistics may result by using the semi-labeled samples generated by ABC-MAP than by the ABC-ML classifier or the ABC-Postprocessing classifier. Second, with good estimation of statistics, the ABC-ML classifier may be able to recover more details, and it is less likely to bias the minority class with small numbers of pixels than the ABC-MAP classifier or the ABC-Postprocessing classifier. Since the ultimate objective here is to generate a classification map with high quality, i.e., high classification accuracy with less speckle but with adequate details, the ABC-ML classifier is chosen to start each cycle to produce the classification results with as much detail as possible. After that the ABC-MAP, or the ABC-Postprocessing classifier is used to further improve classification accuracy by removing the speckle error that usually can be corrected by using contextual information, for instance, spatial proximity. In the following section, an experiment with the proposed algorithm is conducted with hyperspectral data and the results are presented.

IV EXPERIMENTAL RESULTS AND DISCUSSION

For this experiment, the data is part of an airborne hyperspectral data flightline over the Washington DC mall, which were collected by the HYDICE scanner. In this data set there are 210 bands in the 0.4 to 2.4 μm region of the visible and infrared spectrum. In the analysis, the water absorption bands are removed and the remaining 191 bands are used. Since the data has high spatial resolution (about 5 meters), the testing samples and training samples are manually selected by visual inspection with the aid of a SAR image and Digital Elevation Data for the same scene. The detailed information about training and testing samples are shown in Table 1. Even though the training and testing samples can be identified in this case, selecting these many testing samples was a daunting task that took about 3 hours. By comparison, it only took about 15 minutes to select training samples. There is no overlap between training fields and testing fields. The initial statistics are estimated in the original space and then they are used to perform feature extraction using Discriminant Analysis Feature Extraction (DAFE) [12]. To reduce the computation load, 10 reliable features extracted from the 191 spectral bands are used to form a new subspace. The class conditional statistics corresponding to this new subspace are estimated, and then classification is performed in this new subspace. At the following iterations, similar steps are followed except that semi-labeled samples in addition to labeled samples are used to estimate statistics in the original space and the subspace.

Table 1 Training and Testing Samples

(Sub)Class	Training Samples	Testing Samples
<i>roof1</i>	24	151
<i>roof2</i>	20	1459
<i>roof3</i>	23	377
<i>roof4</i>	18	321
<i>roof5</i>	18	280
<i>road1</i>	24	4770
<i>road2</i>	24	876
<i>path</i>	24	586
<i>shadow</i>	23	236
<i>tree</i>	21	1202
<i>grass</i>	21	3229
Total	240	13487

The desired scene classes are *Roof*, *Road*, *Path*, *Trees*, and *Grass*. However, two of these classes are spectrally multimodal and must be modeled by using several subclasses. Thus, five subclasses were used to form the class *Roof*, and two were used to model *Road*. In addition, the class *Shadow* was added so that the list of classes is suitably exhaustive.

This data set is a challenge to analyze for several reasons. First, classes are complex. There is a large diversity in the materials used in constructing rooftops, and consequently no single spectral response is representative of the class *Roof*. Even though some of the subclasses are spectrally quite different, some are quite similar. Subclasses of *Roof*, and *Road* are spectrally similar in that they may be made of similar materials, for instance, asphalt. Third, this data was collected during the dry season; most of lawns are not well grown and as a result, the class *Grass* and *Path* are difficult to differentiate, since some areas of grass are nearly bare soil, which is spectrally similar to the gravel of *Path*.

In table 2, the overall classification accuracy and Kappa statistics obtained by three types of classifiers of the proposed Adaptive Bayesian Classification (ABC) procedure during each cycle with various values of β is illustrated. Figure 4 illustrates the classification accuracy at each cycle obtained by the Bayesian Contextual Classification procedure with $\beta=32$ and the adaptive MLC [1]. Figure 5 shows the variation in classification accuracy of the adaptive contextual classification procedure with β . Here the class accuracy is the ratio of the number of the correctly classified test samples to the number of test samples from a class under consideration. The group accuracy is the ratio of the number of the correctly classified test samples from a functional (information) class to the number of test samples from the same information class. For example, the number of testing samples for the functional class *roof* is the summation of testing samples from the class *roof1 to roof5*. Pixels from one subclass that are classified into a different subclass of the same information class are thus not counted as errors.

The following results may be observed: 1) Adaptive Bayesian Contextual (ABC) classification procedure outperforms the adaptive ML classifier significantly. At each cycle after the initial cycle, three classifiers from ABC, ABC-ML, ABC-MAP and ABC-Postprocessing, achieve higher accuracies than the adaptive MLC. Additionally, the ABC classification procedure converges faster than the adaptive ML classifier. 2) For the ABC classification procedure, both the overall class and group classification accuracies increase as the iteration process progresses. After just three cycles the classification accuracy obtained by ABC-MAP converges with a net increment of about 13% for the class, and about 7% for the group. 3) At each cycle, the ABC-MAP and the ABC-

Postprocessing classifier achieve the higher overall class and group classification accuracies than the ABC-ML classifier does. This indicates that contextual information does help to reduce the speckle error and accordingly improve classification performance.

4) During the first cycle the classification accuracy increment from the ABC-ML to ABC-MAP is about 3% for the class and 2% for the group. However, the classification accuracy increase for the ABC-ML at the second cycle is more than twice that amount, i.e., about 8% for the class and 4% for the group. This indicates that using additional contextual information does improve the classification performance, but the improvement is limited. Essentially, the significant improvement of the classification performance may stem from better statistics estimates produced by the adaptive method.

5) ABC classification performance increases as β becomes large. However, the group classification accuracy doesn't change much. This indicates that the classification result is not very sensitive to the value of β if it is large enough, for example, greater or equal to 16 in this case.

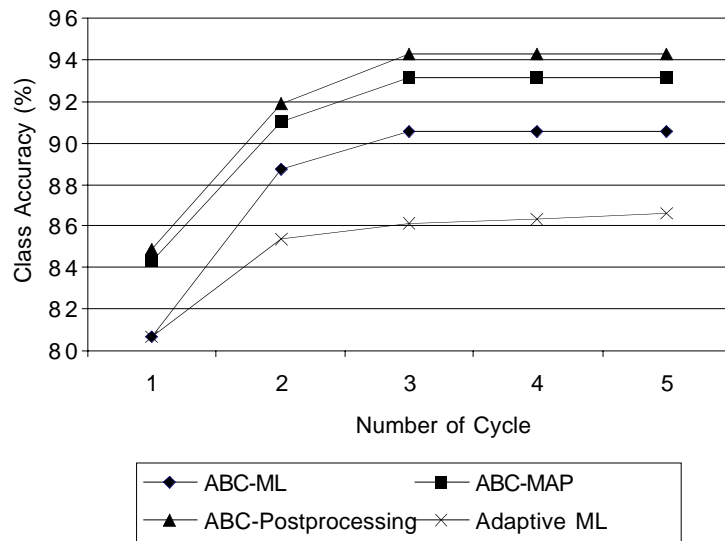
It may seem unexpected that the classification accuracies for both ABC-ML and ABC-Processing classifier vary with β because the decision rules for these two classifiers do not explicit contain β at all. However, in the content of the adaptive Bayesian contextual classification procedure, they should be related β . This may be explained as follows. The ABC-ML is coupled into the loop of the adaptive Bayesian contextual classification procedure, and during each cycle, it uses the statistics estimated by training samples together with semi-labeled samples. Since the class labels of semi-samples are assigned by the ABC-MAP classification rule during the previous cycle, the estimated

statistics at the present cycle are related β . For this reason, the classification outputs for the ABC—ML classifier are affected by β . Similarly, the performance of the ABC-Postprocessing classifier is also a function of β , because this classifier performs classification by reclassifying the classification outputs from the ABC-ML at the same cycle.

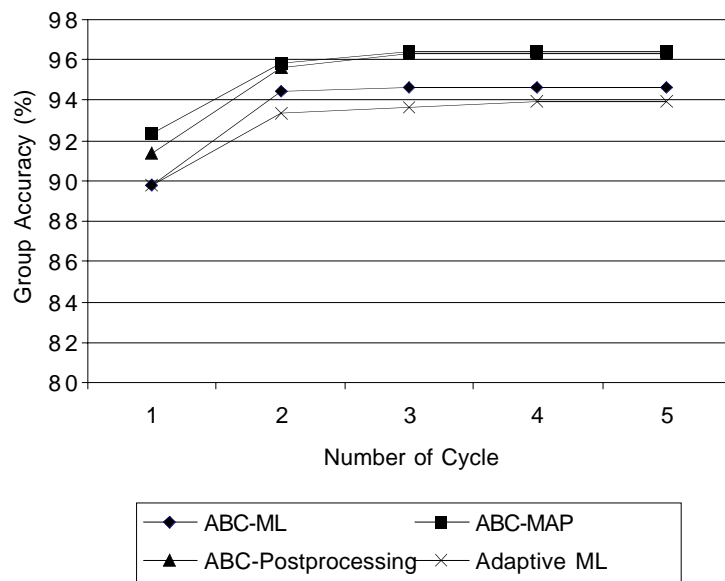
Table 2. The Performance of the Adaptive Bayesian Contextual Classification Procedure.

The accuracy measured by the testing samples in each cycle is given followed by the Kappa statistic in parentheses. The results are given for the average of the subclasses (Class) and after the subclasses have been grouped into the final classes (Group). The Resubstitution results are those for the training samples.

		Accuracy (Kappa Statistics) (%)	$\beta=1$	$\beta=2$	$\beta=4$	$\beta=8$	$\beta=16$	$\beta=32$
Cycle 1	ML	Class	80.7(76.9)	80.7(76.9)	80.7(76.9)	80.7(76.9)	80.7(76.9)	80.7(76.9)
		Group	89.8(86.2)	89.8(86.2)	89.8(86.2)	89.8(86.2)	89.8(86.2)	89.8(86.2)
	MAP	Class	81.8(78.2)	82.6(79.1)	83.3(79.9)	83.7(80.4)	84.1(80.8)	84.3(81.0)
		Group	84.9(81.8)	90.6(87.3)	91.0(87.8)	91.5(88.5)	92.0(89.2)	92.4(89.7)
	Post-Processing	Class	84.9(81.8)	84.9(81.8)	84.9(81.8)	84.9(81.8)	84.9(81.8)	84.9(81.8)
		Group	91.4(88.3)	91.4(88.3)	91.4(88.3)	91.4(88.3)	91.4(88.3)	91.4(88.3)
Cycle 2	ML	Class	86.5(83.9)	86.9(84.1)	87.4(84.6)	87.6(85.1)	88.4(85.8)	88.7(86.2)
		Group	94.0(91.8)	94.0(91.8)	93.9(91.8)	94.0(91.8)	94.2(92.1)	94.4(92.3)
	MAP	Class	88.0(85.4)	88.1(86.7)	89.2(86.8)	89.9(87.6)	90.5(88.3)	91.0(89.0)
		Group	94.9(93.0)	95.2(93.5)	95.3(93.6)	95.5(93.8)	95.8(94.3)	95.8(94.3)
	Post-Processing	Class	90.0(87.7)	90.3(88.1)	90.5(88.4)	91.0(88.9)	91.0(88.9)	91.9(90.0)
		Group	95.5(93.9)	95.5(93.9)	95.4(93.7)	95.5(93.9)	95.5(93.9)	95.6(93.9)
Cycle 3	ML	Class	86.4(84.0)	87.1(84.2)	87.8(85.1)	88.7(86.1)	89.7(87.3)	90.5(88.3)
		Group	92.9(90.3)	93.3(91.7)	94.0(91.8)	94.2(92.1)	94.4(92.4)	94.6(92.8)
	MAP	Class	87.4(85.6)	88.4(85.8)	89.3(86.9)	90.4(88.2)	92.1(90.2)	93.1(91.4)
		Group	93.8(91.4)	94.7(92.8)	95.1(93.3)	95.5(93.8)	96.1(94.6)	96.4(95.0)
	Post-Processing	Class	89.6(87.8)	90.5(88.2)	91.1(89.0)	91.9(89.9)	93.2(91.6)	94.3(92.9)
		Group	95.5(93.8)	95.6(94.0)	95.8(94.2)	96.0(94.5)	96.3(94.9)	96.6(95.3)
Resubstitution		Class	95.6(94.5)					
		Group	96.7(95.5)					



(a) Class accuracy



(b) Group accuracy

Fig. 4 Progression of the classification accuracy with $\beta=32$: (a) Class accuracy
(b) Group accuracy

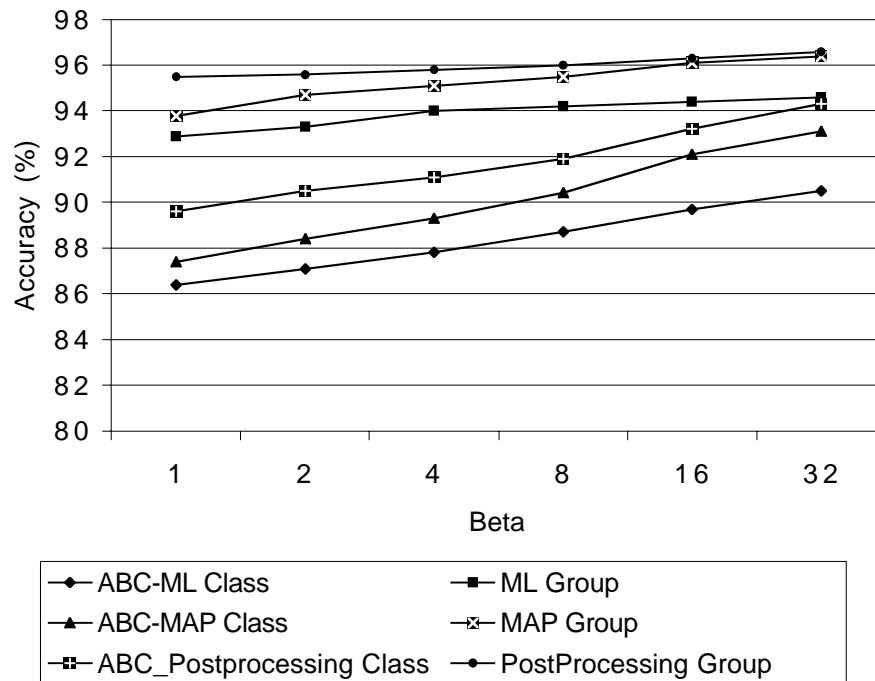


Fig. 5 Classification accuracy versus β in the Adaptive Bayesian Contextual classifier at the last cycle

Even though there are a large number of samples (13,487) in the test fields, this is only about one seventh of the total number of samples (95,456) in the data set. Therefore, the classification accuracy for test fields can only provide an incomplete characterization of this proposed ABC classification procedure's performance. It is worthwhile to examine the thematic maps (a color representation of the classification output where each color represents one class) of the segmented images, which is another way to access the quality of the classification. Fig. 6a through 6c depict the classification results during the initial cycle with $\beta=32$ where the conventional ML, MAP and Postprocessing classifiers are used.

During the initial cycle, the classification performance of the conventional ML and MAP and Postprocessing classifiers at the subspace is not good due to two reasons. First, the extracted features that form the subspace are not optimal because the initial statistics estimates which are used to extract features with DAFE [12] at the original space are quite poor due to the limited training samples. Second, even at the subspace the estimated statistics are still not precise because the number of training samples is still smaller than the number of parameters.

In Fig. 6a it is seen that classification errors occur in many places and some of them are highlighted by rectangles and ovals in Figs. 6b-d. These errors are mostly due to incorrectly estimated statistics and, to a lesser extent, the spectral similarity between classes. For instance, there is a great deal of similarity in the spectral response between information classes *Roof* and *Road*, between *Path* and *Grass*, and between *Tree* and *Grass*. In Figures 6b and 6c speckle errors may be observed either due to the spectral similarity or due to corrupted pixels. For instance, some pixels of the class *Road* are contaminated by the presence of gravel, puddles, and cars etc.. This type of error is greatly reduced by the MAP or the Postprocessing classifiers. However, errors of the first type (highlighted by ovals) still remain. In some areas the MAP or the Postprocessing classifiers create additional errors (highlighted by hexagons) beyond those generated by the MLC. This type of error leads to loss of details, which show the side-effects of these two classifiers. At some areas, the Postprocessing classifier cause more side-effect than the MAPC. This indicates that with additional contextual information the classification performance may be improved by using a conventional MAPC or a Postprocessing

classifier. However, this improvement is certainly limited if the initial classification accuracy obtained by a conventional ML is poor.

During the third cycle, which is shown in Fig. 7a, the classification errors have been greatly reduced and most of detail lost in the first cycle has been recovered by the ABC-MLC. In particular, the lawn areas that have been misclassified as trees have been substantially correctly classified at this cycle. This may be attributed to improved estimation of statistics. However, speckle errors still remain in certain regions, for example, the regions that are highlighted by rectangles. As a result, even with good statistics, the ABC-MLC could not differentiate between the classes with similar spectral responses as well or those classes with contaminated pixels. On the other hand, with additional contextual information, this type of error can be mostly removed by the ABC-MAPC. As shown in Fig. 7b, the thematic map generated by the ABC-MAPC are visually clean and pleasant.

Upon comparing Fig. 7a with Fig. 7c, it is observed that the ABC-MLC outperforms the adaptive MLC. In particular, two classes, *Grass and Trees*, are better separated by the ABC-MLC. Also, the objects on the thematic map shown in Fig. 7b are better defined than those in Fig. 6c, in particular, *Roofs, Grass, and Trees*. This indicates that the ABC-MAPC performs better than the conventional MAPC.

To benchmark the performance of the adaptive Bayesian contextual classification method, all testing samples are used as training samples, and then classification is

performed by a conventional MLC. Subsequently, the MLC performance is tested by the same set of testing samples. The thematic map of the segmented image is shown in Fig. 7d. With the large training sets, two classes, *Grass* and *Trees*, are nicely identified. However, there are some undesired effects. There are many pixels from *Path*, and five subclasses of *Roofs* that are incorrectly identified as *Road*. The possible explanation is as follows: there are a number of testing samples for *Road* that are not good as training samples for the reason that these pixels have been contaminated by the presence of cars, puddles, etc. As a result, the estimated statistics for *Road* using these samples are not quite as precise, leading to the confusion among *Road*, *Path* and *Roofs*. Consequently, the pixels from the classes, *Roofs* or *Path*, might be more likely incorrectly classified as *Road*.

In addition, there are many speckle errors that are mostly scattered on the regions where roads and roofs are located. Some of these speckles are not truly errors for the reason that due to the presence of the cars on the road the pixels on road areas are not truly road. However, some speckles errors may be errors due the similar spectral response of different classes. This further indicates the essential drawback of a pixel classifier, that is, even with good statistics estimates, speckle errors may be unavoidable.

V CONCLUSION

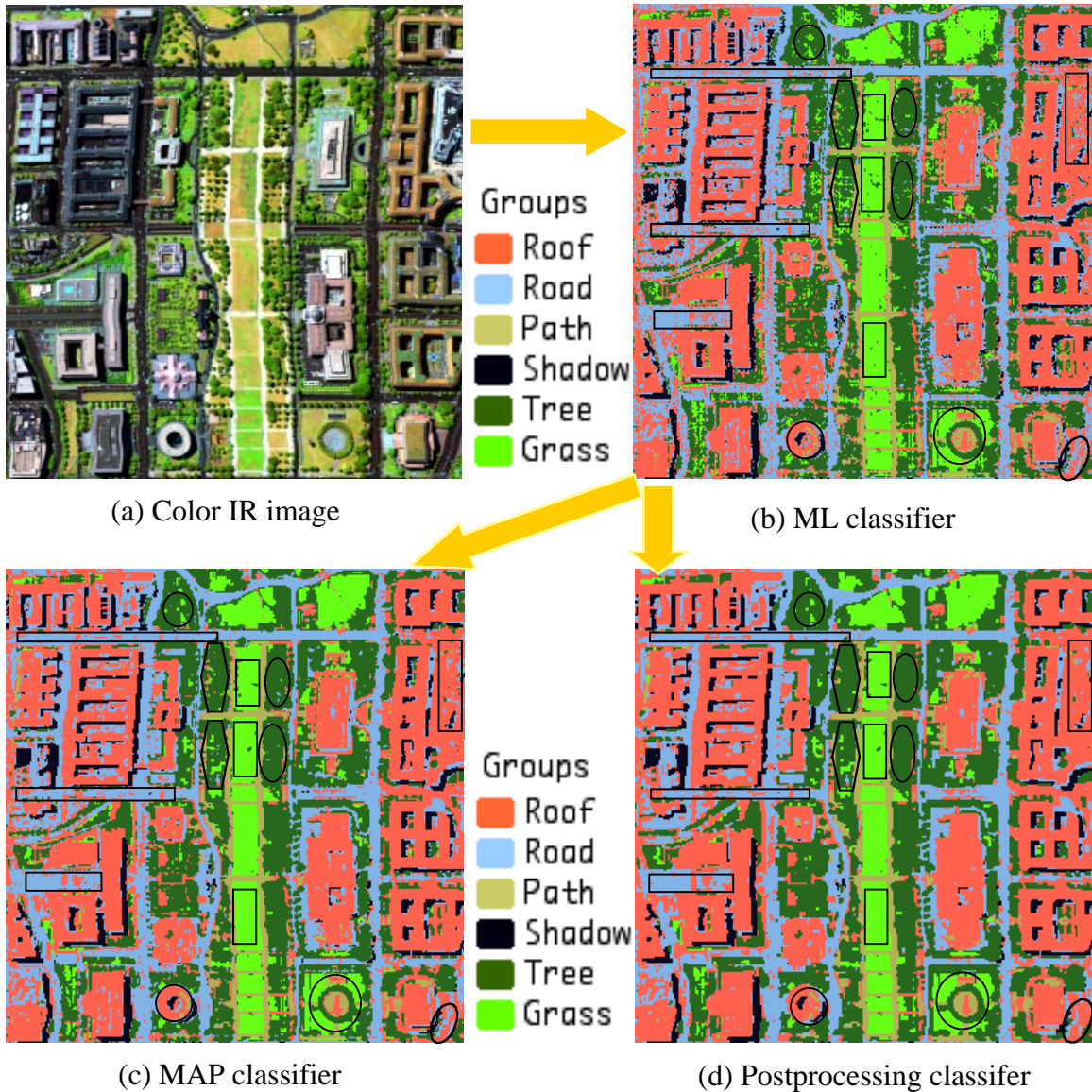
In this paper, an Adaptive Bayesian Contextual classification procedure based on Markov Random Fields is developed. In this procedure, the adaptive classifier and the Bayesian contextual classifier that is approximated by ICM [2] are integrated. As a result,

the advantages of both classifiers are incorporated. As an adaptive ML classifier, the proposed procedure can iteratively enhance statistics estimates and improve classification performance with a limited training sample set. As with a contextual classifier, it can therefore improve the classification accuracy by reducing the speckle errors due to spectral similarity between classes that are very difficult to differentiate by a pixel-wise ML classifier.

The experimental results with hyperspectral data further reveal the benefits of this classification procedure. Starting with a limited training sample set, this method is able to steadily raise classification accuracy and eventually drive it close to the optimal value. The total improvement in the classification accuracy is significant and the convergence rate is fast even though a simple sub-optimal contextual classifier is used. This is significant because the classifier ICM has a reputation of slow convergence when it is used alone [11].

Overall, the proposed procedure is conceptually simple, easy to implement, fast to run, and has high performance. Here, the very simple and efficient sub-optimal contextual classifier, ICM, is integrated with the simple ML classifier. The high performance is achieved because these techniques are combined in a constructive way so that their individual shortcomings can be reduced and their advantages can be amplified. It is specifically advantageous when the pixels have strong local (short distance) statistics independence.

As with the adaptive ML classifier developed in [1], the adaptive Bayesian contextual classification procedure provides a means to mitigate the limitations imposed by the Hughes effect [13] (small training sample problem). In addition, it offers a robust classification procedure that can significantly reduce the analyst's effort in terms of the quantity and quality of training samples selected. This is important because training samples are generally expensive or tedious to obtain. Also, this means the dependence on the skill level of the analyst may be greatly reduced.



- Regions highlighted by the rectangles: speckle errors here may be due to confusions between classes generated by the ML classifier, but most of them are corrected by the MAP and the Postprocessing classifiers.
- Regions highlighted by the hexagons: partial details achieved by the ML classifier, but then lost by the MAP and the Postprocessing classifiers, and the Postprocessing classifier causes even more loss than the MAP classifier
- Regions highlighted by the ovals: classification errors here may be due to bad estimation of statistics with limited training samples that occurs in the ML classifier, and could not be corrected by the MAP or the Postprocessing classifiers and some areas even made it worse.

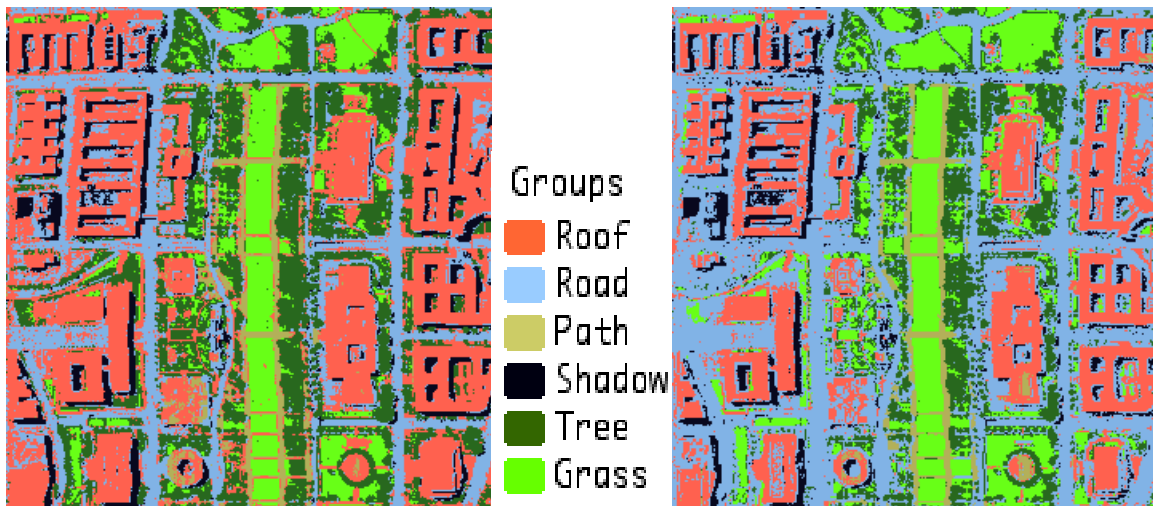
Fig. 6 Original image and the thematic maps of the segmented images during the first cycle with $\beta=32$



(a) ABC-ML classifier

(b) ABC-MAP classifier

- Regions highlighted by the ovals: classification errors occurring during the first cycles may be due to bad statistics estimates but have been corrected by the ABC-ML classifier at this cycle with improved statistics estimates
- Regions highlighted by the hexagons: details lost in the first cycle and then are recovered by the ABC-ML classifier during this cycle, then most of them have been maintained in the subsequent ABC-MAP classifier
- Regions highlighted by the rectangles: speckle errors remain in the results of the ABC-ML classifier at this cycle, but corrected by the ABC-MAP classifier



(c) Adaptive ML classifier

(d) ML classifier with all testing samples as training samples

Fig. 7 Thematic maps (grouped classes) of the segmented images: (a) ABC-ML classifier, (b) ABC-MAP classifier at the third cycle, (c) Adaptive ML classifier at the fifth cycle and (d) ML classifier with all testing samples as training samples

REFERENCES

- [1] Q. Jackson and D. Landgrebe, "An Adaptive Classifier Design for High-Dimensional Data Analysis with a Limited Training Data Set," *IEEE Transactions on Geoscience and Remote Sensing*, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 39, No. 12, pp. 2664-2679, Dec. 2001.
- [2] J. Besag, On the statistical analysis of dirty pictures, *J. Royal Statist. Soc.*, vol. 68, pp.259-302, 1986.
- [3] J. Besag, Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statist. Soc.*, vol. 36, no. 2, pp. 192-236, 1974.
- [4] S. German and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration, *IEEE Trans. Pat. An. Mach, Intell*, vol. PAMI-6, no. 6, pp. 721-741, Nov. 1984.
- [5] R. Kinderman and J. L. Snell, Markov random fields and their applications, *Amer. Math. Soc.*, vol. 1, pp. 1-142, 1980.
- [6] S. Z. Li, *Markov Random Field Modeling in Computer Vision*, Berlin, Germany: Springer-Verlag, 1995.
- [7] C. Bouman and B. Liu, Multiple resolution segmentation of textured images, *IEEE Trans. Pattern. Anal. Machine Intell.*, vol. 13, no. 2, pp. 99-113, 1991.
- [8] B. Gidas, A renormalization group approach to image processing problems, *IEEE Trans. Pat. An. Mach, Intell.*, vol. 11, no. 2, pp. 164-180, Feb. 1989.
- [9] B. Jeon and D. A. Landgrebe, Spatio-temporal contextual classification of remotely sensed multispectral data, Proc. of 1990 IEEE Intern. Conf. on Syst., Man, and Cybern., Los Angeles, CA, pp. 342-344
- [10] Yonhong Jung and Philip H. Swain, Bayesian contextual classification based on modified M-estimates and markov random fields , *IEEE Trans. Geosci. Remote Sensing*, vol.34, no. 1, pp. 68-75, Jan. 1996.
- [11] C. Bouman, class materials for the class *Digital Image processing II*, Purdue University, Fall 2000.
- [12] K. Fukunaga, *Intro. Statistical Pattern Recognition*, San Diego: Academic Press Inc., 1990.
- [13] G. F. Hughes, On the mean accuracy of statistical pattern recognition , *IEEE Trans. Information Theory*, Vol. IT-14, No. 1, pp 55-63, 1968