Investigation of the Random Forest Framework for Classification of Hyperspectral Data

JiSoo Ham, Yangchi Chen, Melba M. Crawford, *Senior Member, IEEE*, and Joydeep Ghosh, *Senior Member, IEEE*

***Abstract*—Statistical classification of byperspectral data is challenging because the inputs are high in dimension and represent multiple classes that are sometimes quite mixed, while the amount and quality of ground truth in the form of labeled data is typically limited. The resulting classifiers are often unstable and have poor generalization. This paper investigates two approaches based on the concept of random forests of classifiers implemented within a binary hierarchical multiclassifier system, with the goal of achieving improved generalization of the classifier in analysis of hyperspectral data, particularly when the quantity of training data is limited. A new classifier is proposed that incorporates bagging of training samples and adaptive random subspace feature selection within a binary hierarchical classifier (BHC), such that the number of features that is selected at each node of the tree is dependent on the quantity of associated training data. Results are compared to a random forest implementation based on the framework of classification and regression trees. For both methods, classification results obtained from experiments on data acquired by the National Aeronautics and Space Administration (NASA) Airborne Visible/Infrared Imaging Spectrometer instrument over the Kennedy Space Center, Florida, and by Hyperion on the NASA Earth Observing 1 satellite over the Okavango Delta of Botswana are superior to those from the original best basis BHC algorithm and a random subspace extension of the BHC.**

***Index Terms*—Binary hierarchical classifier (BHC), classification, classification and regression trees (CART), Hyperion, hyperspectral, Okavango Delta, random forests, random subspace feature selection.**

I. Introduction

THE INCREASING availability of data from hyperspectral sensors, particularly with the launch of the Hyperion instrument on the National Aeronautics and Space Administration (NASA) Earth Observation 1 (EO-1) satellite, has generated tremendous interest in the remote sensing community. These instruments characterize spectral signatures with much greater detail than traditional multispectral sensors and thereby can potentially provide improved discrimination of targets [1]. However, hyperspectral data also present difficult challenges for supervised statistical classification, where labeled training data are used to estimate the parameters of the label-conditional probability density functions [2]. The dimensionality of the data is high ($\sim$200); there are often tens of classes C; and the quantity of training data is often small. Sample statistics of training data may also not be representative of the true probability distributions of the individual class signatures, particularly for remote, inaccessible areas where training data are logistically difficult and expensive to acquire. Generalization of the resulting classifiers is often poor, thereby resulting in poor quality mapping over extended areas.

Various approaches have been investigated to mitigate the impact of *small sample sizes and high dimensionality*, which are inherently coupled issues, since the adequacy of a data sample depends on the data dimensionality, among other factors [3]. For example, regularization methods try to stabilize the covariance matrix by weighting the sample covariance matrix and a pooled covariance matrix or by shrinking the sample covariance matrix toward the identity matrix [4]. While this may reduce the variance of the parameter estimates, the bias of the estimates can increase dramatically. Alternatively, the input space can be transformed into a reduced feature space via feature selection [5] or feature extraction. Although these two approaches reduce the effect of the high-dimensionality problem, feature selection methods are often trapped in a local optimal feature subset, while feature extraction methods lose the interpretability of the original features. Another way of dealing with a small training set is to augment it with unlabeled data and then use semisupervised learning techniques. These methods have been shown to enhance supervised classification [6], [7]. However, convergence of the updating scheme can be problematic, and it is affected by selection of the initial training samples and by outliers.

In analysis of hyperspectral data, Lee and Landgrebe proposed methods for feature extraction based on decision boundaries that maximize separation of data in multiple two-class problems [8]. These decision boundary feature extraction (DBFE) methods are often effective for two-class problems, but do not exploit correlation between sequential bands. Jia and Richards developed the segmented principal components transformation (SPCT) whereby the original bands are grouped into subsets of highly correlated adjacent bands to which the Karhunen–Loeve transform is applied. The most significant principal components are then selected from each subset to yield a feature vector with reduced dimension [9]. The approach treats interband correlation globally and does not guarantee good discrimination capability because the principal components transformation preserves variance in the data rather than maximizing discrimination between classes. Kumar *et al.* investigated band-combining techniques, motivated by

Manuscript received March 24, 2004; revised December 7, 2004. This work was supported in part by the National Aeronautics and Space Administration Earth Observing 1 Program under Grant NCC5-463, in part by the Terrestrial Sciences Program of the Army Research Office under Grant DAAG55-98-1-0287, and in part by the National Science Foundation under Grant IIS-0312471.

J. Ham, Y. Chen, and M. M. Crawford are with the Center for Space Research, The University of Texas at Austin, Austin, TX 78759 USA (e-mail: jham@csr.utexas.edu; yanji@csr.utexas.edu; crawford@csr.utexas.edu).

J. Ghosh is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: ghosh@ece.utexas.edu).

Digital Object Identifier 10.1109/TGRS.2004.842481

best basis functions, as a means of feature extraction in a pairwise classifier framework [10]. Adjacent bands are selected for merging (alternate splitting) in a bottom-up (alternate top down) fashion using the product of a correlation measure and a Fisher discriminant. Morgan *et al.* [11] suggested a similar correlation-based band-combining approach, in conjunction with a covariance shrinkage method, for both a top-down and bottom-up hierarchical classifier to ameliorate the small training data problem.

The theory and practice of classifier ensembles also provide ways of alleviating sample size and high-dimensionality concerns [12]. Bagging involves bootstrapped sampling of the original data, generating a classifier specific to each sample, and then averaging the classifier outputs [13]. This method takes advantage of data reuse, but when the training dataset in the (sub-)sample is very small, the potential for improved diversity and reduced impact of outliers is offset by degradation in individual classifier performance [14]. Boosting also combines weak individual classifiers to develop an improved classifier, but by reweighting training data to increase sensitivity to incorrectly classified training observations. While boosting can improve performance for large training samples, it is not useful for small sample problems, particularly in the presence of outliers. When the input space is large, random subspace (RS) feature selection can potentially provide improved classifier diversity, while stabilizing parameter estimates, by randomly reducing the number of inputs to each classifier in the ensemble and constructing multiple classifiers in the resulting random input space [15], [16]. The method is potentially attractive for problems with redundant input features (e.g., hyperspectral data) and when outliers exist in the training data. Recently, approaches referred to as "random forests of classifiers" have been proposed. These involve developing multiple trees from randomly sampled subspaces of input features, then combining the resulting outputs via voting or a maximum *a posteriori* rule [17]. These methods typically achieve superior generalization for small training samples, but are computationally intensive.

Land cover classification problems usually involve a large number of classes, i.e., the *output space* is large. Output decomposition using binary classifiers in a multiclassifier framework has been shown to be more successful than traditional 1-of-C classifiers for many problems involving large output spaces [18]. Decomposition methods using pairwise classifiers, error-correcting output codes (ECOC) [19], and binary decision trees have all been investigated in this context (see [11] for an overview). Pairwise classifiers develop a separate classifier for each pair of classes, thereby resulting in $O(C^2)$ classifiers that must be combined to determine the final class label. These methods often yield simple classifiers with excellent discrimination for specific pairs, but are generally inefficient for problems with a large number of output classes. In the ECOC, a C-class problem is decomposed into $\bar{C}$ binary problems, whereby the original class is then encoded into a $\bar{C}$ binary vector of a coding matrix. It has been shown that the ECOC method yields robust, stable classifiers. However, since the code matrix design is not based on the characteristics of the classes it represents, interpretability of the classifier is limited.

Binary trees, which often provide an attractive approach for decomposing large output space problems, can be constructed using a variety of splitting functions involving single or multiple features and output classes. To address the high-dimensional output problem while exploiting the affinity for spectrally similar classes, Kumar *et al.* proposed a binary hierarchical classifier (BHC) [20] to decompose a $(C > 2)$-class problem into a binary hierarchy of $(C - 1)$ simpler two-class problems that can be solved using a corresponding hierarchy of classifiers, each based on a simple linear discriminant. The method was extended by Morgan *et al.* [11] for small training samples using an adaptive best basis BHC, which exploits the class-specific correlation structure between sequential bands of hyperspectral data and utilizes an adaptive regularization approach to stabilize covariance estimates. An adaptive random subspace feature selection approach was also investigated within the BHC framework (RS-BHC) as a means of improving classifier performance when the number of training samples is extremely small [21].

In this paper, we investigate a random forest of binary classifiers as a means of increasing diversity of hierarchical classifiers. We evaluate the results obtained for trees produced by our BHC classifier and the original classification and regression trees (CART)-based random forest method [17]. For the BHC, the goal is to exploit the advantages of natural class affinity, while improving generalization in classification of hyperspectral data when the number of training samples is small. The CART-based approach is not directly affected by small sample size statistics and potentially provides greater diversity within the forest, but typically produces trees of enormous size if the output space is large. The paper is organized as follows: the best basis (BB-BHC), random subspace (RS-BHC), and random forest (RF-BHC) implementations of the BHC method and the CART-based framework (RF-CART) are all described in Section II; classification results using the random forest approaches obtained for data acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over the Kennedy Space Center, Florida, and EO-1 Hyperion over the Okavango Delta, Botswana, are presented in Section III and compared to those obtained from the BB-BHC and RS-BHC. Results from all the methods are evaluated, and new directions for future work are suggested in Section IV.

II. Random Forest Binary Hierarchical Classification Method

The top-down BHC framework recursively decomposes a C-class problem into $C - 1$ two-(meta)class problems via a deterministic simulated annealing method [20]. The root classifier tries to optimally partition the original set of classes into two disjoint metaclasses while simultaneously determining the Fisher discriminant that separates these two subsets. This procedure is recursed, i.e., the metaclass Ω_n at node n is partitioned into two metaclasses $(\Omega_{2n}, \Omega_{2n+1})$, until the original C classes are obtained at the leaves. The tree structure, as shown in Fig. 1, allows the more natural, easier discriminations to be accomplished earlier. Fewer classes are involved in the partitioning at lower levels of the BHC hierarchy. Thus, while the classification task typically becomes simpler, the number

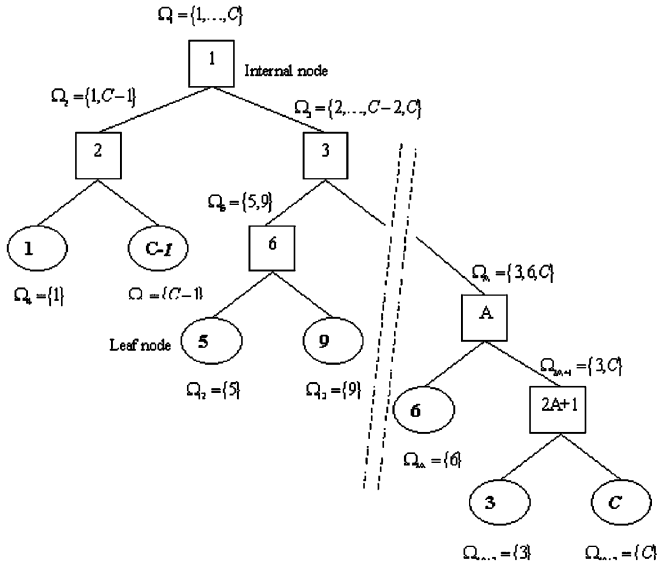


Fig. 1. Binary hierarchical (multi)-classifier framework for solving a C-class problem.

of relevant training samples also decreases. The BB-BHC ameliorates this effect by utilizing an ancestor covariance matrix while exploiting the interband serial correlation through an adaptive class-dependent band aggregation process [11]. A band-combining step is performed on highly correlated spectrally adjacent bands prior to the partitioning of metaclasses, thereby reducing the number of inputs relative to the number of training data points. Bands are aggregated until a user-defined ratio R between the number of training samples for the respective (meta)classes and input dimension is achieved. Typically, R is selected to be at least 5.

The BHC method is extended in the RS-BHC approach by utilizing the random subspace method as a postprocessing stage to tree construction, with the goal of reducing the number of inputs while refining decision boundaries [22]. The BB-BHC method is used to first construct the hierarchy, and then random subspace sampling is performed at each node of the tree where the criterion for R is not satisfied. For each (meta)class m with n_m vector-valued observations, $X_m = (X_1, \ldots, X_{n_m})$, a subset of elements of $X_i^{=}(x_{i1}, \ldots, x_{ik})$ with dimension $p_m = n_m/R < k$ is then randomly selected from the k-dimensional set of features. The resulting modified training set $X_m^r = (X_1^r, \ldots, X_{n_m}^r)$ consists of observation vectors $X_i^r = (x_{i1}^r, \ldots, x_{ip}^r)$ where the same subset of features is selected for each element $X_i^r \in X^r, (i = 1, \ldots, n_m)$. The number of random subspaces selected at each such node is $N_s = (k/p_m) \times F$, where the value of F is a user-supplied input. A discriminant vector is constructed for each random subspace, and the N_s vectors are combined at each node of the hierarchy via majority voting. Our empirical evidence indicates that good results are typically achieved for $2 < F < 4$, which provides adequate coverage of the feature space. Improvement in classification accuracy is not significant for $F > 4$.

The random forest implementation of the BHC (RF-BHC) extends the RS-BHC by incorporating random subspace feature selection in the actual development of the tree. This is particularly advantageous, as random subspace sampling is performed by the RS-BHC only at nodes where the ratio R is not exceeded. Thus, subsampling of the input features typically occurs only at lower levels of the tree, thereby limiting diversity. For moderate sized training samples, bagging can increase diversity of the multiclassifier system, so a bootstrap sample of observations is selected for each tree in the RF-BHC. At each metaclass node m, a random subspace of features of dimension $p_m = \min(p_m, N_f)$ is selected to determine the decision boundary for the classifier at that node, where N_f is a user-selected input. To guarantee greater diversity, we choose $N_f \ll k$. The tree is then developed using the resulting set of features selected at each node. The process is repeated to grow a forest of identically, independently distributed random vectors associated with the individual trees.

The fundamental difference between BHC and other decision trees is that the former focuses on decomposing the output space; partitioning of the input space occurs as a consequence. Both RF-BHC and RF-CART use the random forest ensemble method to increase the diversity of each base learning module, then combine results of the individual modules (trees). While Breiman's CART-based random forest follows a typical binary divide-and-conquer hierarchical scheme, it differs from the BHC in the base learning module. The BHC uses the generalized framework for associative modular learning systems (GAMLS) [23] algorithm to split each node into metaclasses that are separated by the maximum Fisher distance. Using a sequence of binary tests, CART seeks the split that maximizes the reduction of the impurity of the parent nodes and its two child nodes as measured by the Gini index [24]. The most discriminating feature is selected to perform the split. Used in the random forest context, a random subspace of the original k features is selected at each node of the tree, and the most discriminating feature is then selected. Further, unlike the actual CART method, the RF-CART approach does not perform pruning of nodes, as pruning reduces diversity of trees in the forest. Analogous to the RF-BHC, each tree is grown using a bootstrap sample of the training set.

III. Results

Hyperspectral data from the following two sources were analyzed in this paper.

1) *Kennedy Space Center, Florida*: The NASA AVIRIS instrument acquired data over the Kennedy Space Center (KSC), Florida, on March 23, 1996. AVIRIS acquires data in 224 bands of 10-nm width with center wavelengths from 400–2500 nm. The KSC data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. After removing water absorption and low signal-to-noise (SNR) bands, 176 bands were used for the analysis. Training data were selected using land cover maps derived from color infrared photography provided by KSC and Landsat Thematic Mapper (TM) imagery. The vegetation classification scheme was developed by KSC personnel in an effort to define functional types that are discernible at the spatial resolution of Landsat and these AVIRIS data. Discrimination of land cover for this environment is difficult due to the similarity of spectral signatures for certain

TABLE I
CLASS CODES, NAMES, AND NUMBER OF TRAINING SAMPLES FOR KENNEDY SPACE CENTER AVIRIS

	Class	No. samples
1	Scrub	761 (14.6%)
2	Willow swamp	243 (4.66%)
3	Cabbage palm hammock	256 (4.92%)
4	Cabbage palm/oak hammock	252 (4.84%)
5	Slash pine	161 (3.07%)
6	Oak/broadleaf hammock	229 (4.38%)
7	Hardwood swamp	105 (2.0%)
8	Graminoid marsh	431 (8.27%)
9	Spartina marsh	520 (9.99%)
10	Cattail marsh	404 (7.76%)
11	Salt marsh	419 (8.04%)
12	Mud flats	503 (9.66%)
13	Water	927 (17.8%)

TABLE II
CLASS CODES, NAMES, AND NUMBER OF TRAINING AND SPATIALLY DISJOINT TEST SAMPLES FOR BOTSWANA HYPERION DATA

	Class	No. samples, Test set	No. samples, S. D. Test set
1	Water	270 (8.31%)	126 (5.05%)
2	Hippo grass	101 (3.09%)	162 (6.5%)
3	Floodplain grasses1	251 (7.74%)	158 (6.34%)
4	Floodplain grasses2	215 (6.63%)	165 (6.62%)
5	Reeds1	269 (8.27%)	168 (6.74%)
6	Riparian	269 (8.27%)	211 (8.46%)
7	Firescar2	259 (7.98%)	176 (7.06%)
8	Island interior	203 (6.26%)	154 (6.17%)
9	Acacia woodlands	314 (9.67%)	151 (6.05%)
10	Acacia shrublands	248 (7.65%)	190 (7.62%)
11	Acacia grasslands	305 (9.38%)	358 (14.35%)
12	Short mopane	181 (5.56%)	153 (6.13%)
13	Mixed mopane	268 (8.27%)	233 (9.34%)
14	Exposed soils	95 (2.92%)	89 (3.57%)

vegetation types. For classification purposes, 13 classes representing the various land cover types that occur in this environment were defined for the site (Table I). Classes 4 and 6 represent mixed classes.

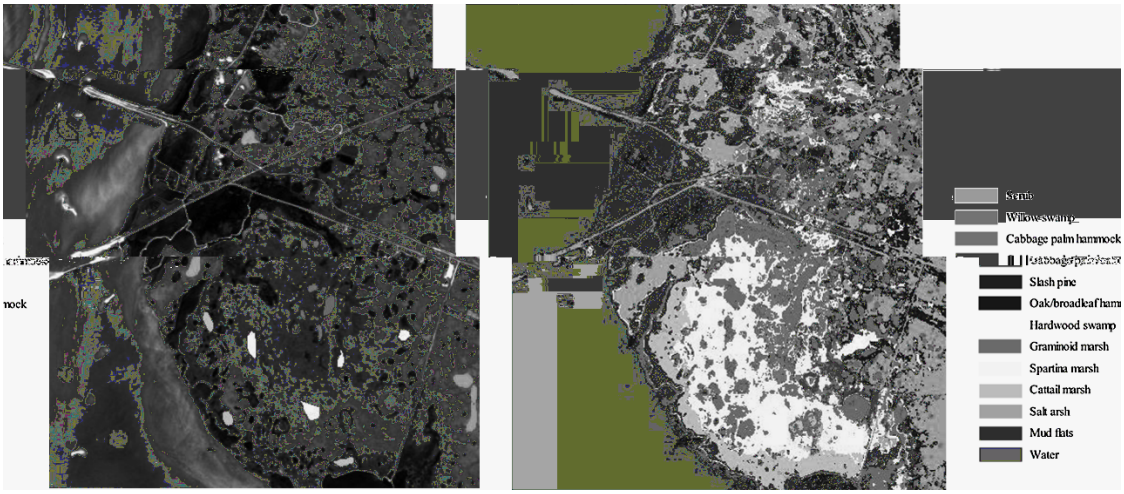
2) *Okavango Delta, Botswana*: The NASA EO-1 satellite acquired a sequence of data over the Okavango Delta, Botswana in 2001–2004. The Hyperion sensor on EO-1 acquires data at 30-m pixel resolution over a 7.7-km strip in 242 bands covering the 400–2500-nm portion of the spectrum in 10-nm windows. Preprocessing of the data was performed by the University of Texas Center for Space Research to mitigate the effects of bad detectors, interdetector miscalibration, and intermittent anomalies. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features: [10–55, 82–97, 102–119, 134–164, 187–220]. The data analyzed in this study, acquired May 31, 2001, consist of observations from 14 identified classes representing the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the delta [22]. These classes were chosen to reflect the impact of flooding on vegetation in the study area. The class names and corresponding numbers of ground truth observations used in the experiments are listed in Table II. Classes 3 and 4 are both floodplain grasses that are seasonally inundated, but differ in their hydroperiod (the amount of time inundated). Classes 9–11 represent different mixtures of acacia woodlands, shrublands, and grasslands and are named according to the dominant class. Training data were selected manually using a combination of global positioning system-located vegetation surveys, aerial photography from the Aquarap (2000) project, and 2.6-m resolution IKONOS multispectral imagery. The class priors for both datasets, as indicated by the labeled data, are only moderately skewed. For simplicity, we assume the class priors to be equal while developing the BHC classifier. This assumption shall be reconsidered later.

For both datasets, ten randomly sampled partitions of the labeled data were subsampled such that 75% of the data samples were used for training and 25% for testing. In order to investigate the impact of the quantity of training data on classifier performance, each of the training datasets was then randomly subsampled to create samples whose sizes corresponded to 50%, 30%, and 15% of the original labeled data. All classifiers were evaluated against the same ten testing samples comprised of 25% of the original labeled data in order to isolate the impact of sample size.

Experiments were performed using the BB-BHC, RS-BHC, RF-BHC, and RF-CART. Although authors recommend various values for the dimension of the random subspace and the number of trees in a random forest, there do not appear to have been any systematic studies of the issue to date. In the results reported here, the ratio R was set at 5, and the value of F was 4 for the RS-BHC method. In our experiments, the dimension of the random subspace was determined adaptively in the BHC, but was always selected such that the value of R was at least 5. For the RF-BHC, the value of N_f was selected to be 20. In order to have somewhat comparable inputs, 20 input features were randomly selected in the RF-CART method. One hundred trees were grown for each experiment, as our sensitivity studies showed that larger forests did not provide improved results for these datasets.

A. Results: Original Training and Test Areas

Kennedy Space Center: The true-color image shown in Fig. 2(a), along with the classification results obtained from the RF-BHC in Fig. 2(b), shows the spatial distribution of classes and training sites over the 614×512 pixel study area. Average classification accuracies for test data and associated



(a) (b)

Fig. 2. (a) AVIRIS data, (Bands 31, 21, 11) acquired over KSC, training sites overlaid. (b) Classified image of KSC AVIRIS data using RF-BHC classifier.

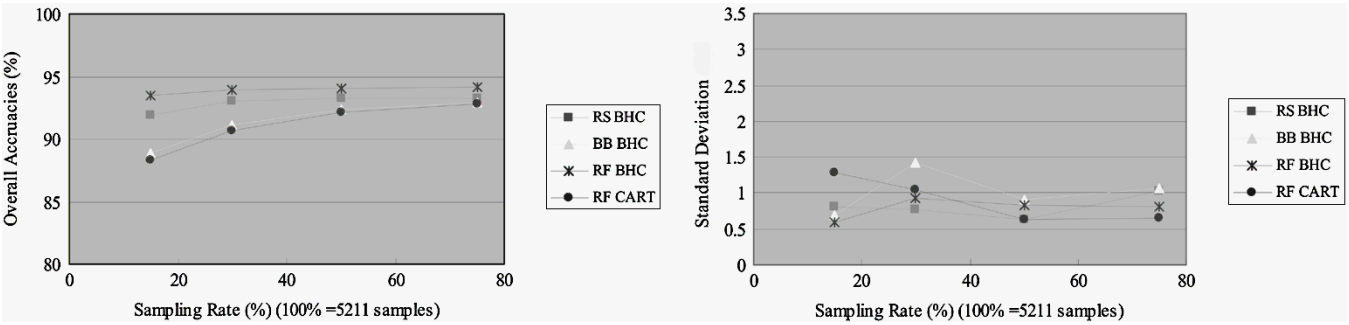


Fig. 3. Average and standard deviation of classification accuracies for AVIRIS test data.

standard deviations for the ten experiments conducted with each classifier are plotted in Fig. 3. The overall trends in accuracies relative to the quantity of training data are similar for all methods when applied to the test dataset. At the 75% sampling rate, the accuracies of all methods are nearly the same, although the RF-BHC yields somewhat higher accuracies than the other methods. The results obtained by the BB-BHC and RF-CART methods are very similar over all sampling rates, with RF-CART yielding slightly lower accuracies. These methods consistently produced the lowest overall average accuracies. The RS-BHC yielded approximately the same accuracies as the BB-BHC at the 75% sampling rate, but improved relative to the BB-BHC and the RF-CART approach at lower sampling rates. For the BHC-based methods, this appears to demonstrate the value of reduced redundancy in the input space and improvements achieved by better tuning of the decision boundaries, even though the tree structure is identical to the BB-BHC and random sampling of the feature space is not required until lower levels of the tree (particularly for the higher training data fractions). Results were also obtained using the original RS-BHC and best basis aggregated data. The BB-RS-BHC consistently yielded slightly lower accuracies than the RS-BHC because of the reduced diversity of trees, but results were not statistically different and are not reported here. The overall average accuracy of the RF-BHC is consistently the highest and improves relative to other BHC methods and RF-CART, as the fraction of training data is reduced.

The RF-BHC is also the most stable method over all training fractions, as measured by the standard deviation of the accuracies. The standard deviations of the accuracies of the random-subspace-based methods appear to benefit from the diversity of the input space. The standard deviation of the accuracies obtained by the BB-BHC increased dramatically at the 30% sampling rate because it was necessary to aggregate a large number of bands to satisfy the ratio R. The problem, which is manifested both in the tree building and in the decision boundary of the BB-BHC, is offset in the determination of the RS-BHC decision boundary. It should be noted that although the standard deviation of BB-BHC decreases at the 15% sampling rate, the associated average classification accuracy is also poor, further demonstrating it is uniformly inferior at low sampling rates. The reduced accuracy of RF-CART at low sampling rates, relative to the BHC-based random forest methods, is attributed to the value of the inherent exploitation of class affinities by the BHC approaches. Further, although the standard deviation of the accuracies for the RF-CART approach is low for high sampling rates, it increases consistently as the sampling rate of the training data is reduced, likely because the discrimination capability of the single best feature within a small random sample of inputs may be quite variable. Further, the benefits of bagging

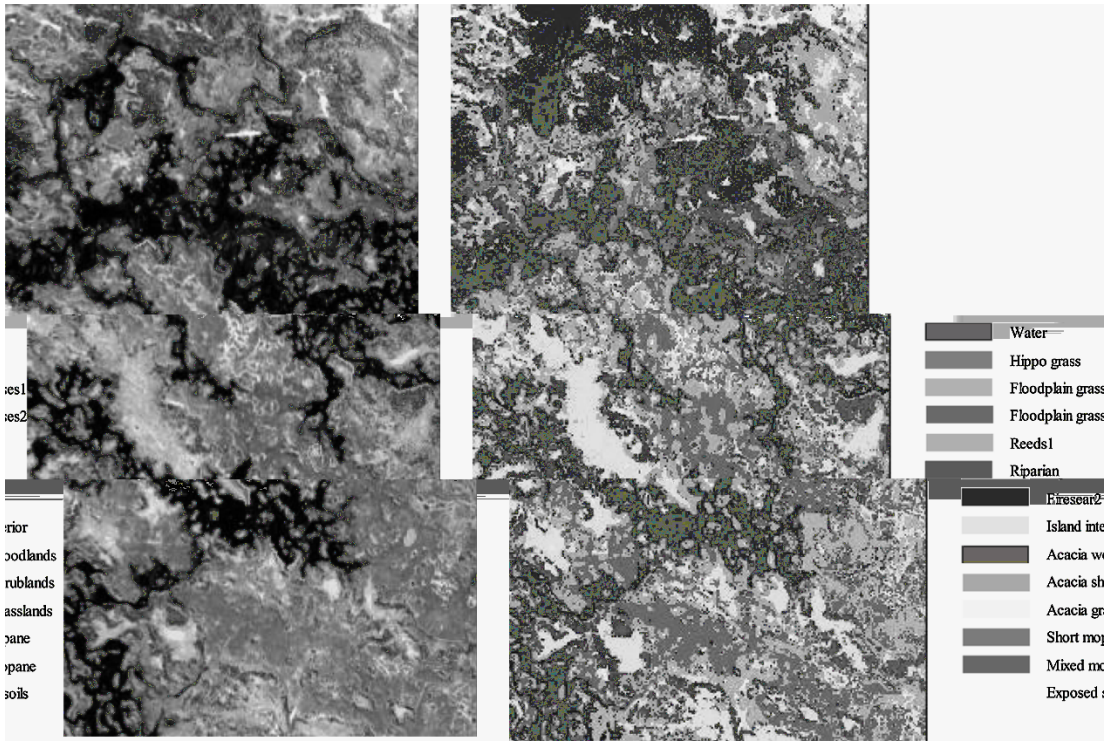


Fig. 4. (a) Hyperion data, (Bands 51, 149, 31) acquired over Okavango Delta, training sites overlaid. (b) Classified image of Hyperion data over Okavango Delta using RF-BHC classifier.

the training sample occur at the higher sampling rates for both the RF-BHC and RF-CART methods.

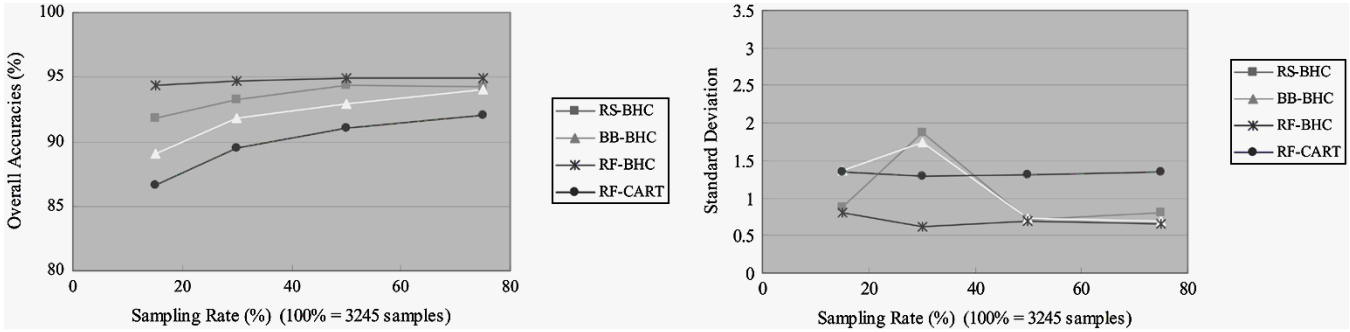
Okavango Delta: The RGB image in Fig. 4(a) and the classification results obtained by the RF-BHC in Fig. 4(b) show that the spatial distribution of classes is extremely complex over this 256×1476 pixel area. Using the same random sampling strategy as for the KSC data, results were obtained at each percentage for all four classifiers. Plots of classification accuracies at the various sampling rates are shown in Fig. 5. The overall trends in accuracies relative to the fraction of training data are similar those of the KSC AVIRIS test data. Among the classifiers, RF-BHC yielded the highest classification accuracy on all training sample fractions, and performance degraded only slightly at lower sampling rates. The standard deviations of the accuracies obtained using the RF-BHC are low and remain nearly constant over the various sampling rates. At the 30% sampling rate, the standard deviations of the accuracies yielded by the BB-BHC and RS-BHC are substantially higher. For the BB-BHC, this again appears to be due to the amount of band aggregation required to achieve the ratio R. Unlike the KSC case, the RS-BHC is apparently unable to mitigate problems associated with band aggregation during the tree construction phase for the Okavango data.

The difference in results produced by the RF-BHC and RF-CART methods was unexpected, since both utilize an ensemble of 100 trees to build a stronger classifier. Previous research by Tumer and Ghosh [14] indicated that the accuracy of an ensemble method relies on the diversity of the base classifier. To investigate the performance of the individual trees, we further analyzed the performance of both random forest methods at the 75% sampling rate. The average accuracy over the set of individual trees developed by the RF-BHC is 89.2%, and the standard deviation is 1.3. The overall accuracy for the RF-BHC, which is determined by simple voting, increased by 5.7% to 94.9%. For the RF-CART method, the average accuracy obtained using 100 trees is 84.2%, with standard deviation 1.3. The ensemble of these 100 trees, using simple voting utilized in the original Brieman random forest, yielded a 7.8% increase to 92%. For this type of classification problem, it appears that the BHC is a better base classifier than CART, although CART realizes substantial improvement when trees are combined.

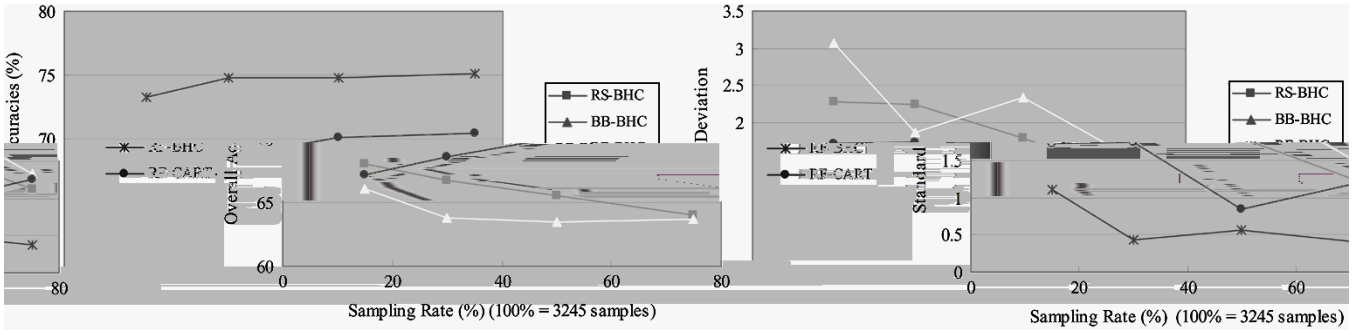
B. Generalization to Spatially Disjoint Areas

Traditionally, the training and test data are spatially colocated and can thus be assumed to be samples from the same distribution. In practice, however, it is also useful to estimate how a classifier will perform in areas that are somewhat different, in order to indicate how much additional data labeling and retraining is needed to make the model applicable to much larger areas. With this goal in mind, a "spatially disjoint" test set was also acquired from a geographically separate location at the Botswana site and used to evaluate the classifiers developed previously.

These spatially disjoint data have somewhat different characteristics from the training/test data, so the performance of all classifiers is reduced, as expected. Still, as with the test data, the BB-BHC yielded the lowest overall average accuracy at all sampling rates. The incremental improvement in average accuracy achieved by the random subspace method increases with



(a)



(b)

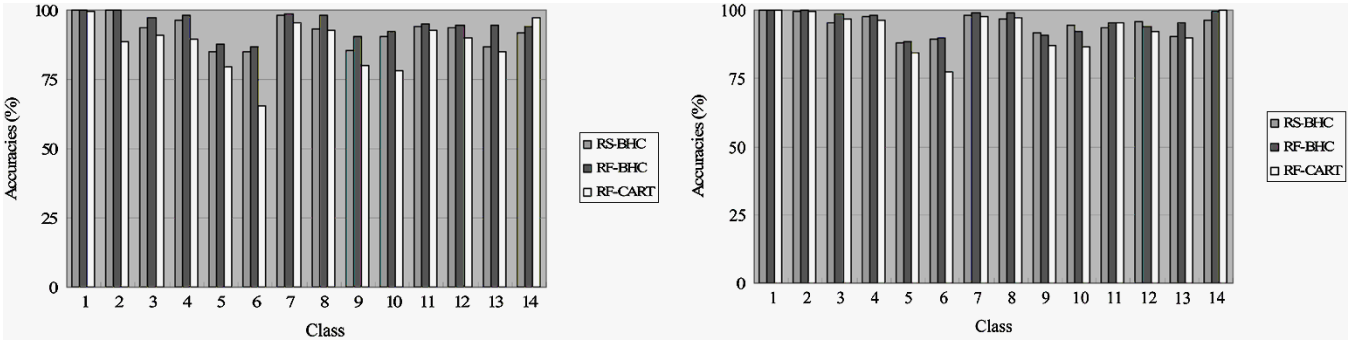
Fig. 5. (a) Average and standard deviation of classification accuracies for Hyperion test sets. (b) Average and standard deviation of classification accuracies for Hyperion spatially disjoint test sets.

reduced sampling rates, but is not statistically significant, as the standard deviations of the accuracies also increase substantially with lower sampling rates. The RF-BHC implementation and the RF-CART method yielded higher accuracies for the spatially disjoint test data at all sampling rates than both the BB-BHC and the RS-BHC, thereby demonstrating the greater generalization of these approaches. Similar to results from the test data, the RF-BHC consistently produced the highest overall average accuracies for the spatially disjoint test set, indicating the value of exploiting class affinity, coupled with the increased diversity of trees achieved by forcing random sampling of the input space at all nodes. The RF-CART method also achieved good generalization, as indicated by its performance on these spatially disjoint test data, although results for the original test set were inferior to the other methods. This is attributed both to the diversity that it achieves and its reduced dependence on the training sample statistics. Similar to the test data, the performance of both the RF-BHC and RF-CART methods was further investigated for the 100 trees obtained from the Hyperion spatially disjoint test set at the 75% sampling rate. The average classification accuracy over the set of individual RF-BHC trees is 68.2%, and the standard deviation is 2.9. The ensemble random forest result using simple voting is 75.2%, an increase of 7%. For RF-CART the values are 60.8% and 2.4, respectively. The classification accuracy increased by 9.6% to 70.4% when the 100 trees were combined using simple voting.

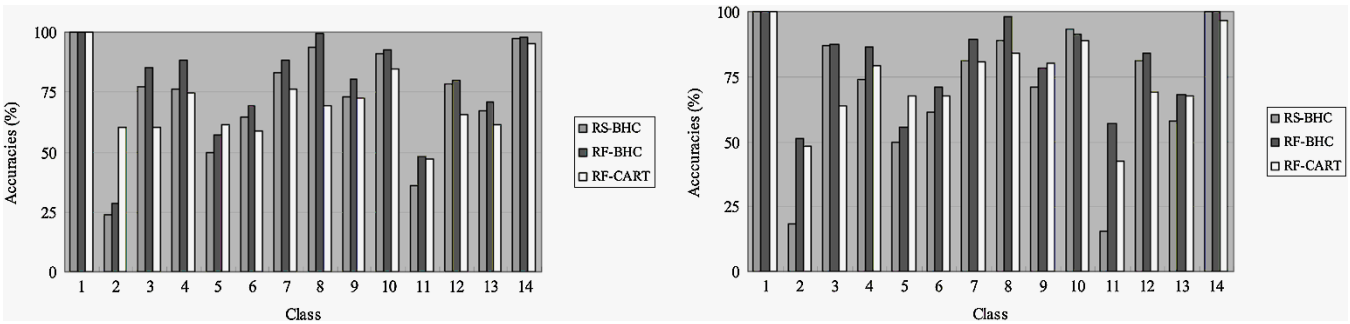
Since RF-BHC and RF-CART use the same random forest framework, their differences lie both in the tree construction and the underlying classifier. For the remotely sensed data in this study, the BHC exploits class affinity, while the good performance of the CART-like method on the spatially disjoint test set suggests that it provides more diversity. To further investigate this issue, we calculated the entropy, a nonpairwise diversity measure [12], of trees obtained from both RF-BHC and RF-CART (Table III). The results indicate that RF-CART method produces more diverse trees than RF-BHC at all four sampling rates. RF-CART achieves an 9.6% increase in accuracies via the ensemble, while RF-BHC results improve only 7%, thereby reinforcing the idea that ensemble methods benefit more from combining diverse classifiers. Further, as the sampling rate increases, the diversity of RF-BHC trees decreases. Under the same situation, however, the diversity of the RF-CART forest remains comparatively consistent. This means that the RF-BHC inputs become more homogeneous as the number of samples increases, while RF-CART does not follow the same trend. Overall, the advantages of an ensemble approach are clear as the RS-BHC used only one tree structure rather than an ensemble of potentially different trees, which significantly reduced the generalization of its classification accuracies on the spatially disjoint test set.

TABLE III
ENTROPY-BASED DIVERSITY OF ENSEMBLE MEMBERS OBSERVED FOR THE SPATIALLY DISJOINT BOTSWANA HYPERION DATA AT DIFFERENT SAMPLING RATES

Methods \ Sampling Rate		15%	30%	50%	75%
RF-BHC	Average	0.440	0.383	0.345	0.326
	Std	0.013	0.008	0.006	0.007
RF-CART	Average	0.516	0.476	0.460	0.460
	Std	0.024	0.020	0.011	0.010



(a)



(b)

Fig. 6. (a) Class-dependent accuracies for Hyperion test set. (Left) 15% and (right) 75% sampling rate. (b) Class-dependent accuracies for Hyperion spatially disjoint test set. (Left) 15% and (right) 75% sampling rate.

The differences between the overall accuracies for the spatially disjoint test set and those for the test set are quite remarkable. As noted earlier, the test data are spatially colocated with the training data, whereas the spatially disjoint test set is not. Clearly, either the class priors or the class-conditional feature distributions (or both) are substantially different, at least for some classes in the more remote area. This motivated us to further investigate class-specific results. Class-dependent accuracies for the Hyperion test and spatially disjoint test sets are provided in Fig. 6, and the detailed confusion matrix for the RF-BHC is contained in Table IV. Results in Table II indicate that the priors were indeed somewhat different. In particular, there were relatively more samples of Classes 2 and 11, and less of Classes 1 and 9. However, while false negative errors increase for Classes 2 and 11, there is no overall clear trend. For example, classification accuracies for Class 1 (water) which is spectrally quite distinct, are unaffected by the change of priors. Moreover, several class accuracies are now much lower than 80%, while others are almost unaffected. This leads us to believe that change in class-conditional distributions in certain classes that are spectrally quite similar is the main cause of the marked degradation in their classification accuracies. In particular, the overall classification accuracies of RS-BHC, RF-BHC and RF-CART methods are strongly influenced by the performance of Classes 2 and 11. Class 2, hippo grass, which grows within the river channels, has a small training sample and is spectrally similar to water, as many pixels are mixed with water. Class 11, acacia grasslands, is a mixed class that is most often confused with other grasses or acacia shrubs, which is also a mixed class.

Using Fig. 6 (b), we can also compare the class-specific accuracies for the spatially disjoint test set for the RS-BHC, RF-BHC and RF-CART approaches. Consistent with the overall accuracies, the performance of the RF-BHC is generally better than RF-CART method at both the 75% and 15% sampling rates. Similarly, the RS-BHC yields consistently lower accuracies, particularly for Classes 2 and 11. Although higher classification accuracies were achieved for Class 2 by RF-CART than the two BHC methods, it is not statistically significant, as the standard deviation of the average sample accuracy is more than 12.

In comparing the overall computational requirements of the BHC-based and RF-CART methods, there are several trade-offs. BHC-based trees always solve $C - 1$ binary problems.

TABLE IV
CONFUSION MATRIX FOR HYPERION SPATIALLY DISJOINT TEST SET AT 75% SAMPLING RATE, RF-BHC CLASSIFIER

Classified\Actual (col)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
1	126	44	0	0	0	0	1	0	0	0	0	0	0	0
2	0	84	0	0	4	0	0	0	0	0	0	0	0	0
3	0	0	139	18	3	0	5	0	1	0	0	0	0	0
4	0	0	0	144	4	2	0	0	0	0	0	0	0	0
5	0	0	0	2	96	23	0	0	0	0	0	0	0	0
6	0	15	0	0	16	149	0	0	30	0	0	0	2	0
7	0	2	0	0	0	0	160	0	0	0	0	0	0	0
8	0	0	0	0	12	0	0	150	0	0	92	1	0	0
9	0	17	4	1	4	36	0	0	115	0	0	0	60	0
10	0	0	15	0	11	0	7	0	1	172	38	0	6	0
11	0	0	0	0	0	0	0	0	0	17	208	0	0	0
12	0	0	0	0	7	0	0	0	0	0	1	129	2	0
13	0	0	0	0	9	1	3	0	4	1	6	23	163	0
14	0	0	0	0	2	0	0	4	0	0	13	0	0	89
Total	126	162	158	165	168	211	176	154	151	190	358	153	233	89
Class accuracies	100	51.9	88	87.3	57.1	70.6	90.9	97.4	76.2	90.5	58.1	84.3	70	100

At the 75% sampling rate, the average CART decision tree for Botswana Hyperion data contained 326 nodes (standard deviation =16). For this 14-class problem, the BHC tree had only $27\ (1 + 13 * 2)$ nodes. For the same experiment, the CPU time for the RF-CART method was 8 min 42 s, while it was 1 h 4 min 4 s for the RF-BHC. Both experiments were performed on a 3-GHz Pentium IV CPU machine. The RF-BHC required more CPU time than the RF-CART method because GAMLS is a deterministic simulated annealing algorithm. It should be noted that while neither algorithm was coded as an operational method, average timing results reflect their relative computational requirements.

IV. Conclusion and Future Work

The primary purpose of the study was to investigate the performance of random feature subset selection methods in terms of generalization. The secondary goal was to investigate the performance of the methods when applied to data acquired by Hyperion data, which have low SNR. The performance of an implementation that focused on tuning decision boundaries of the BHC, and that of two random forest approaches was investigated. Classification accuracies achieved by ensemble methods rely heavily on achieving diversity within the ensemble. The conflicting effects of improved SNR and reduced spectral resolution from band aggregation appear to be positively complemented by the improved diversity achieved by the RS-BHC through random sampling of the original features. We also noted that the change in classification accuracies achieved by using a forest rather than a single tree indicates that the RF-CART method actually achieves greater incremental benefit from the ensemble than the RF-BHC. Thus, the ensemble both exploits the greater diversity provided by the single feature splits and mitigates the potential impact of selecting features that are redundant or have poor discrimination capability.

A critical characteristic of the BHC is that it exploits the natural groupings of similar classes, which often occur in remotely sensed data acquired over natural landscapes. This provides a natural hierarchy that is often well handled by the simple Fisher discriminant. The random forest methods all yielded superior results for both test and spatially disjoint test data at our two study sites, with the improvement being greater for the spatially disjoint test set, thereby indicating improved generalization to extended areas. For these data, RF-BHC produced stable results over all sampling rates. Additional study is required to better characterize this issue. In this context, elimination of irrelevant and possibly redundant input features should also be considered in the RF-BHC. Other classifiers, such as the ECOC and support vector machines, should also be investigated within the RF-BHC framework. Overall, the RF-BHC methods appear to be quite promising in terms of generalization, but should be applied to many more datasets with different characteristics in order to better assess their overall performance. Also, much work remains to be done on determining how to improve performance on extended areas represented by the spatially disjoint dataset, especially, since both the class mixtures and class-conditional spectral properties can change in such situations. If this problem can be solved, then one can more confidently label much larger regions than those directly described by the available labeled data. For mixed classes, the issue may be mitigated in some cases by determining relative abundances of component classes via unmixing of hyperspectral data, if representative signatures of pure classes can be obtained [25], [26]. Approaches for representing spatially nonstationary spectral signatures may also be appropriate.

Acknowledgment

The authors thank A. Neuenschwander (UT Center for Space Research) for help in preprocessing the Hyperion data and interpreting the overall classification results.

References

[1] J. S. Pearlman, P. S. Berry, C. C. Segal, J. Shapanski, D. Beiso, and S. L. Carman, "Hyperion: A space-based imaging spectrometer," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1160–1173, Jun. 2003.

[2] D. Landgrebe, "Hyperspectral image data analysis as a high dimensional signal processing problem," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.

[3] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, Mar. 1991.

[4] S. Tadjudin and D. A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 4, pp. 2113–2118, Jul. 1999.

[5] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, Jul. 2001.

[6] Q. Jackson and D. Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set," *IEEE Trans. Geosci.Remote Sens.*, vol. 39, no. 12, pp. 2664–2679, Dec. 2001.

[7] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Computational Learning Theory*, 1998, pp. 92–100.

[8] C. Lee and D. Landgrebe, "Decision boundary feature extraction for neural networks," *IEEE Trans. Neural Networks*, vol. 8, no. 1, pp. 75–83, Jan. 1997.

[9] X. Jia and J. A. Richards, "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 1, pp. 538–542, Jan. 1999.

[10] S. Kumar, J. Ghosh, and M. M. Crawford, "Best basis feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 29, no. 7, pp. 1368–1379, Jul. 2001.

[11] J. T. Morgan, A. Henneguelle, M. M. Crawford, J. Ghosh, and A. Neuenschwander, "Adaptive feature spaces for land cover classification with limited ground truth," *Intl. J. Pattern Recognit. Artif. Intell.*, vol. 18, no. 5, pp. 777–800, 2004.

[12] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: Wiley, 2004.

[13] L. Breiman, "Bagging predictors," *Mach. Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[14] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," *Connection Sci.*, vol. 8, no. 3/4, pp. 385–404, 1996.

[15] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[16] M. Skurichina and R. P. W. Duin, "Bagging, boosting, and the random subspace method for linear classifiers," *Int. J. Pattern Anal. Appl.*, vol. 5, no. 2, pp. 121–135, 2002.

[17] L. Breiman, "Random forests," *Mach. Learning*, vol. 45, pp. 5–32, 2001.

[18] J. Furnkranz, "Round robin classification," *J. Mach. Learning Res.*, vol. 2, pp. 721–747, 2002.

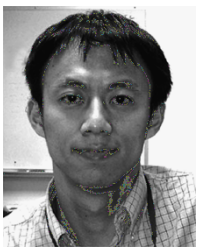
[19] T. G. Dietterich and R. Bakiri, "Solving multiclass learning problems using error correcting output codes," *J. Artif. Intell. Res.*, vol. 2, no. 1, pp. 263–286, 1995.

[20] S. Kumar, J. Ghosh, and M. M. Crawford, "Hierarchical fusion of multiple classifiers for hyperspectral data analysis," *Int. J. Pattern Anal. Appl.*, vol. 5, no. 2, pp. 210–220, 2002.

[21] M. Crawford, J. Ham, and J. Ghosh, "Robust classifiers for hyperspectral data analysis using limited training data," presented at the *2003 Tyrrhenian International Workshop on Remote Sensing*, Elba Island, Italy, Sep. 15–18, 2003.
[22] A. L. Neuenschwander, M. M. Crawford, and S. Ringrose, "Results of the EO-1 experiment—Use of Earth Observing-1 Advanced Land Imager (ALI) data to assess the vegetational response to flooding in the Okavango Delta, Botswana," *Int. J. Remote Sens.*, 2005, to be published.
[23] S. Kumar and J. Ghosh, "GAMLS: A generalized framework for associative modular learning systems," in *Proc. Applications and Science of Comp. Intelligence II*, Orlando, FL, 1999, pp. 24–34.
[24] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
[25] J. W. Boardman, "Geometric mixture analysis of imaging spectrometry data," in *Proc. IGARSS*, vol. 4, 1994, pp. 2369–2371.
[26] G. P. Asner and K. B. Heidebrecht, "Spectral unmixing of vegetation, soil, and dry carbon in arid regions: Comparing multi-spectral and hyperspectral observations," *Int. J. Remote Sens.*, vol. 23, pp. 3939–3958, 2002.

JiSoo Ham received the B.S. degree from Ewha Womans University, Seoul, Korea, in 1995, and M.A. from the University of Texas (UT), Austin, in 1997, both in mathematics. She is currently pursuing the Ph.D. degree at UT.

She is currently a Research Assistant with the Center for Space Research, UT.



Yangchi Chen received the B.S. degree in naval architecture and ocean engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1996, and the M.S. degree in operations research and industrial engineering from the University of Texas (UT), Austin, in 2001. He is currently pursuing the Ph.D. degree at UT.

He is currently a Graduate Research Assistant with the Center for Space Research, UT.

Melba M. Crawford (M'89–SM'05) received the B.S. degree in civil engineering and the M.S. degree in civil and environmental engineering from the University of Illinois, Urbana, in 1970 and 1973, respectively, and the Ph.D. degree in industrial and systems engineering from The Ohio State University, Columbus, in 1981.

She holds an Engineering Foundation endowed professorship in mechanical engineering at the University of Texas (UT), Austin, and has been a faculty member of UT since 1980. She has been affiliated with the UT Center for Space Research since 1988 and is currently Head of the Remote Sensing and Image Processing group. She teaches in the graduate program in Operations Research within the Mechanical Engineering Department and is an Associate Director of the University of Texas Environmental Science Institute. She is a member of the NASA Earth System Science and Applications Advisory Committee (ESSAAC) and was a member of the NASA EO-1 Science Validation team for the Advanced Land Imager and Hyperion. Her research interests are in statistical pattern recognition, data fusion, and signal and image processing as applied to the analysis of remotely sensed data, with a current emphasis on hyperspectral and lidar data. She has more than 100 publications in scientific journals, conference proceedings, and technical reports.

Dr. Crawford was named a Member of the First Class of Jefferson Science Fellows at the U.S. Department of State in 2004. She and her colleagues recently received the outstanding paper award at IICAI-2003. She has also received the UT Outstanding Faculty Excellence Award, the General Dynamics Teaching Excellence Award, the Halliburton Education Foundation Award for Excellence in Teaching, and the Outstanding Graduate Faculty Award for the University of Texas at Austin. She is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and Vice President of the IEEE Geosciences and Remote Sensing Society Administrative Committee.

Joydeep Ghosh (SM'04) received the B.Tech degree from the Indian Institute of Technology, Kanpur, in 1983, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1988.

He joined the Department of Electrical and Computer Engineering, University of Texas, Austin, in 1989, and is currently a Full Professor since 1998 and a holder of the Archie Straiton Endowed Fellowship. He directs the Laboratory for Artificial Neural Systems, where his research group is studying the theory and applications of adaptive pattern recognition, data mining including web mining, and multilearner systems. He has published more than 200 refereed papers and edited 12 books. He is currently an Associate Editor of *Pattern Recognition*, *Neural Computing Surveys*, and the *International Journal of Smart Engineering Design*.

Dr. Ghosh has received nine Best Paper Awards, including the 1992 Darlington Prize for the Best Journal Paper from IEEE Circuits and Systems Society, Best Theory Paper at SDM 04, and the Best Applications Paper at ANNIE'97. He has served as Conference Co-Chair of Artificial Neural Networks in Engineering (ANNIE)'93 -ANNIE'96, ANNIE'98–2003, and in the program or organizing committees of several conferences on data mining and neural networks each year. More recently, he coorganized workshops on web mining (with the SIAM International Conference on Data Mining, 2001 and 2002) and High Dimensional Clustering (SDM 2003). He was a Plenary Speaker for ANNIE'97, MCS2002, and was Letters Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS (1998–2000).