Ensemble Multiple Kernel Active Learning For Classification of Multisource Remote Sensing Data

Yuhang Zhang, *Student Member, IEEE*, Hsiuhan Lexie Yang, *Student Member, IEEE*, Saurabh Prasad, *Senior Member, IEEE*, Edoardo Pasolli, *Member, IEEE*, Jinha Jung, and Melba Crawford, *Fellow, IEEE*

***Abstract*—Incorporating disparate features from multiple sources can provide valuable diverse information for remote sensing data analysis. However, multisource remote sensing data require large quantities of labeled data to train robust supervised classifiers, which are often difficult and expensive to acquire. A mixture-of-kernel approach can facilitate the construction of an effective formulation for acquiring useful samples via active learning (AL). In this paper, we propose an ensemble multiple kernel active learning (EnsembleMKL-AL) framework that incorporates different types of features extracted from multisensor remote sensing data (hyperspectral imagery and LiDAR data) for robust classification. An ensemble of probabilistic multiple kernel classifiers is embedded into a maximum disagreement-based AL system, which adaptively optimizes the kernel for each source during the AL process. At the end of each learning step, a decision fusion strategy is implemented to make a final decision based on the probabilistic outputs. The proposed framework is tested in a multisource environment, including different types of features extracted from hyperspectral and LiDAR data. The experimental results validate the efficacy of the proposed approach. In addition, we demonstrate that using ensemble classifiers and a large number of disparate but relevant features can further improve the performance of an AL-based classification approach.**

***Index Terms*—Active learning (AL), ensemble classification, multiple kernel learning, multisource data.**

I. Introduction

MULTISOURCE remote sensing data analysis has been an emerging research topic in recent years, due to the development of relevant remote sensing technologies, e.g., very high resolution (VHR) optical, multispectral, and hyperspectral sensors, synthetic aperture radar (SAR), and light detection and ranging (LiDAR) systems. Given sufficient labeled ground reference data, supervised learning methods are effective for analysis of these data, for problems including classification, spectral unmixing, and anomaly detection. Unfortunately, the performance of supervised learning models is heavily dependent on the availability of representative labeled data for training, which in real-world applications are usually expensive and time-consuming to obtain. Moreover, the very high dimensionality of feature spaces resulting from hyperspectral imagery makes it difficult to design reliable classifiers with a limited quantity of labeled data. However, manual selection of training data from imagery, a common practice, is subjective (particularly if accomplished via visual interpretation) and tends to introduce redundancy into the supervised classifier because data are selected in spatially contiguous patches, and thus slow the training process. Therefore, it is important to collect training data that are most informative and useful for the underlying classification task.

Active learning (AL) was introduced for such tasks in the machine learning community [1], and has been demonstrated to be useful for classification of remote sensing data [2], [3]. Unlike traditional passive learning, where labeled data are used to train the classifier and unlabeled samples are subsequently classified, in AL, users can interact with the classifier, both providing capability to select the most informative samples and allowing adaptation in dynamic environments. In the AL framework, classifiers are initially trained on a very limited set of training samples, but additional informative and representative samples are identified from the abundant unlabeled data, labeled, and then inducted into this set, thereby growing the training dataset in a systematic way. The goal is to minimize the cost related to the sample labeling process while maximizing the discrimination capabilities.

In recent years, many AL strategies have been proposed for remote sensing data classification. One group is specific to margin-based classification approaches, such as support vector machine (SVM) classifiers. In this context, margin sampling (MS) represents a simple but powerful strategy [4], [5], where the importance of samples is based on the distance to the hyperplane, which indicates the level of uncertainty and its importance toward learning the decision boundary. Samples whose distance to the hyperplane is small are likely to be

Manuscript received April 10, 2014; revised July 28, 2014; accepted September 02, 2014. This work was supported in part by NASA AIST Grant 11-0077, in part by the Hyperspectral Image Analysis Group at the University of Houston, and in part by the Laboratory for Applications of Remote Sensing at Purdue University.

Y. Zhang and S. Prasad are with the Hyperspectral Image Analysis Group, Electrical and Computer Engineering Department, University of Houston, Houston, TX 77004 USA (e-mail: yzhang87@uh.edu; saurabh.prasad@ieee.org).

H. L. Yang, E. Pasolli, and M. Crawford are with the Laboratory for Applications of Remote Sensing, within Civil Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: hhyang@purdue.edu; epasolli@purdue.edu; mcrawford@purdue.edu).

J. Jung is with the School of Science and Engineering, Texas A&M University Corpus Christi, Corpus Christi, TX 78412 USA (e-mail: jinha.jung@tamucc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2014.2359136

support vectors, and thus more important for learning the classifier. In [6], instead of using the distance to the hyperplane as selection measure, the original classification problem is reformulated into a new binary problem where the goal is to discriminate between significant and nonsignificant samples. While exploiting the kernel space induced by spectral features has been demonstrated as a successful framework for MS, an enhanced kernel space can be constructed by including multiple features or sources within a multiple kernel learning (MKL) framework, as demonstrated in our preliminary work reported in [7]. Although MS is effective in many AL problems, limitations include the following. 1) It can only be applied to margin-based methods, in which decisions are made on the distance to a separating hyperplane. 2) It is not suitable when multiple classifiers are involved in the learning process, since no inter-classifier-information among the samples is considered.

Another popular family of AL strategies quantifies the uncertainty of samples by considering a committee of learners [8]. Each member of the committee builds its own learning model, and consequently labels the samples in the candidate pool. The algorithm then selects the samples which have the maximum disagreement for different classification models in the committee. Among the strategies that can be utilized to construct committees, a recently proposed approach for hyperspectral imagery utilizes feature subsets as a proxy for multiple views (each view being a member in this ensemble/committee) [9].

Other methods are based on posterior probabilities. In [10], using a maximum-likelihood classifier, the samples whose inclusion in the training set maximizes the changes in the posterior distribution are selected. In [11], samples are selected as a function of entropy of the corresponding class label. Another strategy is represented by the breaking-ties criterion [12], in which the difference between the two largest posterior probabilities is considered. More recently, researchers have incorporated spatial information [13], [14], and have incorporated the AL framework in practical operational scenarios [15], [16]. Moreover, while the number of samples is an appealing way to formulate the labeling cost, the time spent for labeling or the distance traveled in the field can be more appropriate for certain applications [17].

Although several AL strategies have been proposed in the literature, they have been applied mostly for single-sensor remote sensing data. Little research has been conducted in multisensor scenarios. Multiple sensors can provide diverse information, and a combination of different features has the potential to provide a more "complete" representation of objects on the ground. However, the high dimensionality of the resulting feature space necessitates appropriate classification approaches. In recent work, composite kernel learning (CKL) and MKL have been shown to outperform traditional single-kernel machines for some scenarios. Both are based on a combination (typically linear) of different base kernels, where each kernel is dedicated to a particular type of feature (e.g. a unique source). CKL was first proposed to improve the classification of hyperspectral images by combining spectral and spatial information [18], and further improved in [19] by removing the constraints on weight parameters. MKL has outperformed traditional single kernel approaches in different applications [20], [21] and has also been applied successfully for remote sensing [22]–[24]. However, previous research has been investigated in a fixed training set scenario, and little attention has been given to the integration of MKL and AL frameworks.

A preliminary attempt to integrate MKL and AL was recently discussed in our preliminary work [7], in which a MKL-AL framework based on the MS selection strategy was proposed to integrate multiple features or various sensors in an adaptive manner for effective learning. We demonstrated that combining SimpleMKL [25] and MS can be an effective and robust AL approach for multiple disparate features (derived from one or more sensors), considering that tuning the kernel parameters and determining the weights in the linear mixture of kernels is not trivial. Although adaptive kernels can be developed for each source, the MS-based selection strategy is not source-specific, which can result in a potentially sub-optimal use of each source.

The goal of this paper is to develop a robust source-specific AL framework that can be applied in a multisource environment to select important samples for labeling. In particular, we propose a novel ensemble multiple kernel active learning (EnsembleMKL-AL) system based on the maximum disagreement query strategy that incorporates different types of features extracted from multisensor remote sensing data.

The remainder of the paper is organized as follows. The proposed EnsembleMKL-AL framework is presented in Section II. In Section III, we briefly introduce the different types of features considered in this paper, including two LiDAR derived features, extended multiattribute profiles (EMAPs) and object-based texture features. Experimental results and analysis are presented in Section IV. Finally, concluding remarks are included in Section V.

II. Proposed Method

A. Framework

We propose an EnsembleMKL-AL framework for robust classification of multisource/multifeature remote sensing data. The flowchart of the proposed framework is shown in Fig. 1. Consider an initial small set of labeled samples extracted from various sources, noted as $L = \{(\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^P), y_i\}_{i=1}^N$, where y_i is the label of multisource data $(\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^P)$, P is the number of sources, and N is the total number of labeled samples. The goal is to select a series of examples from a set of unlabeled samples $U = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^P\}_{i=1}^Q$ and add them to the training set after labeling (in a practical framework, this labeling would typically be undertaken by a human analyst, e.g., via photo-interpretation or collection of ground reference information). It is desired to obtain good system performance by inducting as few training samples as possible that add the most information. Q is the total number of unlabeled samples, and satisfies $Q \gg N$. The proposed framework can be subdivided into four steps: (1) multiple features are extracted from multisource data; (2) an ensemble of probabilistic MKL

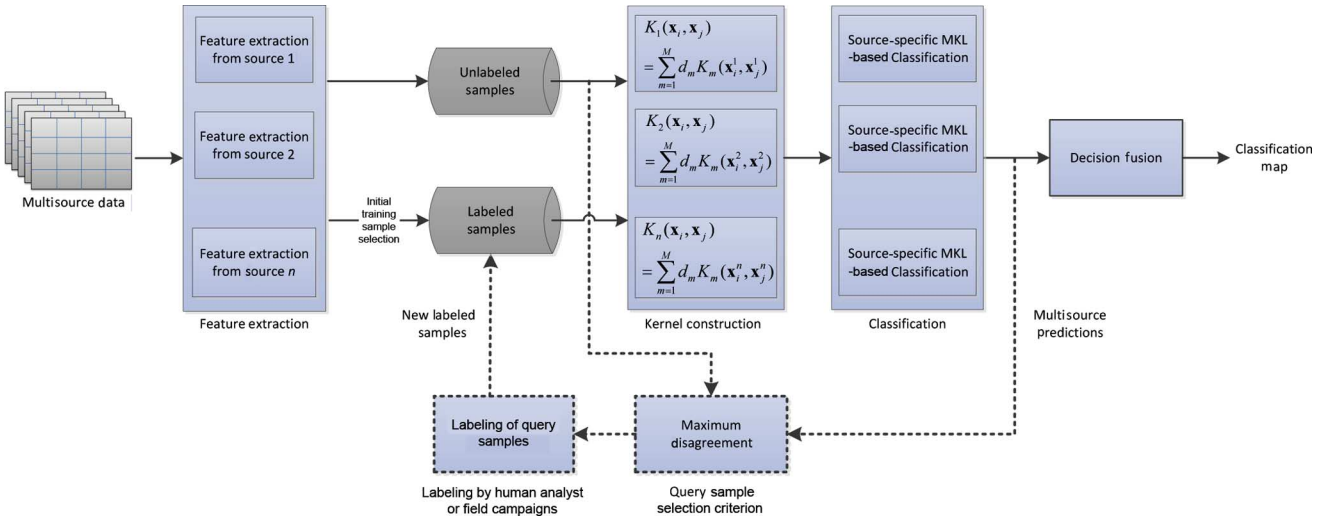


Fig. 1. Flowchart of the proposed framework.

classifiers is implemented to optimize the kernel for each source or feature set; (3) a maximum disagreement-based AL strategy is used to select the most informative samples; (4) after the learning is accomplished, a decision fusion strategy is applied to the posterior probabilities computed by the MKL classifiers to obtain a final classification map. In this paper, a soft fusion strategy—a linear opinion pool (LOP) [26] is studied, and compared with the widely used majority voting (MV) method. A detailed description of the steps is presented in the following subsections.

B. Probabilistic Multiple Kernel Classifier

In most kernel-based learning methods [27], [28], performance is greatly affected by the choice of kernel function and related kernel hyper-parameters. The standard SVM only utilizes a single kernel function with fixed parameters, which necessitates model selection for good classification performance. Besides, using a fixed kernel may introduce bias, since different sources of data may have different representations of the phenomena of interest, and hence the similarity should not be measured via the same kernel function. For such situations, multiple kernel learning methods have been proposed, which typically apply a weighted linear combination of basis kernels instead of selecting one specific kernel function for learning. In the multisource scenario, for a specific source p, the combined kernel function K between two pixels $\mathbf{x}_i^p$ and $\mathbf{x}_j^p$ can be represented as

$$\begin{aligned} K(\mathbf{x}_i^p,\mathbf{x}_j^p) &= \sum_{m=1}^{M} d_m K_m(\mathbf{x}_i^p,\mathbf{x}_j^p) \\ \text{s.t. } \mathrm{d_m} \geq 0 \quad &\text{and} \quad \sum_{m=1}^{M} d_m = 1 \end{aligned} \tag{1}$$

where M is the number of candidate basis kernels representing different kernel parameters, K_m is the m-th basis kernel and d_m is the associated weight. Weights can be estimated through cross-validation, which is computationally demanding when the number of basis kernels (i.e., feature sets or data sources) is large. An alternative strategy, which we adopt in this paper, is based on the SimpleMKL algorithm [25]. It optimizes the weights automatically in a learning problem. Based on the SVM optimization problem, the SimpleMKL learning problem is expressed as

$$\min_{\mathbf{d}} J(\mathbf{d}), \quad \text{s.t. } d_m \geq 0, \quad \text{and} \quad \sum_{m=1}^{M} d_m = 1$$

$$J(\mathbf{d}) = \begin{cases} \min\limits_{\mathbf{w},b,\xi} \frac{1}{2} \sum\limits_{m=1}^{M} \frac{1}{d_m} \|\mathbf{w}_m\|^2 + C \sum\limits_{i=1}^{N} \xi_i \\ \text{s.t. } y_i \left(\sum\limits_{m=1}^{M} \langle \mathbf{w}_m, \Phi_m(\mathbf{x}_i^p) \rangle + b \right) \geq 1 - \xi_i \\ \xi_i \geq 0 \quad \forall i = 1,2,\ldots,N \end{cases} \tag{2}$$

where $\Phi_m(\mathbf{x}_i^p)$ is the kernel mapping function of $\mathbf{x}_i^p$, $\mathbf{w}_m$ is the weight vector of the mth decision hyperplane, C is the regularization parameter controlling the generalization capabilities of the classifier, and ξ_i is a positive slack variable.

Similar to standard SVM, SimpleMKL can also be represented in its dual form as

$$\begin{aligned} \max \Bigg\{ L(\alpha_i,\alpha_j) &= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \\ &\quad \times \sum_{m=1}^{M} d_m K_m(\mathbf{x}_i^p,\mathbf{x}_j^p) \Bigg\} \end{aligned} \tag{3}$$

$$\text{s.t.} \begin{cases} \sum\limits_{i=1}^{N} \alpha_i y_i = 0 \\ \alpha_i,\ \alpha_j \in [0,C] \quad \forall i,j = 1,2,\ldots,N \\ d_m \geq 0, \quad \text{and} \quad \sum\limits_{m=1}^{M} d_m = 1 \end{cases}$$

where α_i and α_j are Lagrange multipliers. The kernel weight d_m can be optimized via a simple gradient descent approach by updating it along the direction of the negative gradient

of $L(\alpha_i, \alpha_j)$. The gradient of the objective function can be computed as

$$\frac{\partial L}{\partial d_m} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K_m(\mathbf{x}_i^p, \mathbf{x}_j^p),\ m = 1, 2, \ldots, M. \tag{4}$$

Then d is updated by using a search scheme as

$$\mathbf{d} \leftarrow \mathbf{d} + \gamma\mathbf{D} \tag{5}$$

where γ is the step length, D is the descent direction of $L(\alpha_i, \alpha_j)$, and $\mathbf{d} = [d_1, d_2, \ldots, d_M]^{\mathrm{T}}$ is the kernel weight vector. Following this optimization, SimpleMKL provides a predicted label for each test sample. However, many applications require a posterior class probability instead of a specific label. Platt proposed an approach to approximate the posterior class probabilities $P(y = 1|\mathbf{x})$ by a sigmoid function which is commonly used in single-kernel SVM implementations [29], [30]. In this paper, we implement this approach for our MKL framework in a similar way.

C. Maximum Disagreement-Based Ensemble AL

Ensemble AL is based on a committee of learners, in which each member of the committee is learned on a subset of the samples or of the feature space. Diversity is important for building a robust ensemble [9], [31], as it ensures that each subset contains additional information to improve the learner relative to the other subsets. In previous work with hyperspectral imagery, these conditions were created via a multiview approach, in which the original set of spectral features was partitioned into disjoint subsets, i.e., different views [9]. A similar approach can be adopted in our case, in which different sources or different types of features can be assigned to a different view of the data.

In an AL framework, the choice of criterion function used to select samples from the unlabeled set is crucial. In the context of ensemble AL, we adopt the maximum disagreement criterion, which demonstrated its capabilities in identifying the most informative samples across multiple views [9]. It is based on two successive steps.

Let us suppose that the estimated label of a sample $\mathbf{x}_i^p \in U$ from source/feature-set p is obtained by learning a classification function, and $\hat{y}_i^p = f(\mathbf{x}_i^p)$. For each sample i, we define a symmetric matrix D_i, which measures the disagreement between each pair of predictions $(\hat{y}_i^1, \hat{y}_i^2, \ldots, \hat{y}_i^P)$. Each element of the matrix is defined as

$$D_i(p, n) = \begin{cases} \Delta(\hat{y}_i^p, \hat{y}_i^n), & \text{if } p \neq n \\ 0, & \text{if } p = n \end{cases} \tag{6}$$

where

$$\Delta(\hat{y}_i^p, \hat{y}_i^n) = \begin{cases} 1, & \text{if } \hat{y}_i^p \neq \hat{y}_i^n \\ 0, & \text{if } \hat{y}_i^p = \hat{y}_i^n. \end{cases} \tag{7}$$

$p, n \in \{1, 2, \ldots, P\}$. Then, the disagreement level of sample i over all sources can be expressed as

$$DL_i = \sum_{p=1}^{P}\sum_{n=1}^{P} D_i(p, n). \tag{8}$$

The maximum disagreement contention set P_{MD} is constructed by selecting unlabeled samples with the maximum uncertainty, i.e., the maximum disagreement level. If the number of samples with the highest disagreement level is less than the batch size, i.e., the number of samples to select, more samples having the maximum disagreement levels are included in P_{MD}. In this way, P_{MD} is always larger than the batch size. Note that because we want to keep the size of the training set small, a small fixed number of samples should be carefully selected in each learning step. However, the samples belonging to the maximum disagreement set are usually characterized by strong redundancy. To limit this, and to select nonredundant samples, a pruning strategy is applied on P_{MD}.

We use weighted voting entropy (WVE) [32] to quantitatively measure the uncertainty of votes over labels provided by each source. For this purpose, we define a $P \times N_c$ weighting matrix $\mathbf{W}$ (where N_c is the number of classes), in which $\mathbf{W}(p, c)$ is the class-specific accuracy for source p and class c. The WVE value of $\mathbf{z}_i \in P_{MD}$ at the τ^{th} query is defined as

$$WVE^{\tau}(\mathbf{z}_i) = -\frac{1}{\log \varpi^{\tau}}\sum_{c=1}^{N_c}\frac{\sigma_c^{\tau}(\mathbf{z}_i)}{\varpi^{\tau}}\log\left(\frac{\sigma_c^{\tau}(\mathbf{z}_i)}{\varpi^{\tau}}\right) \tag{9}$$

where

$$\varpi^{\tau} = \sum_{p=1}^{P}\sum_{c=1}^{N_c}\mathbf{W}^{\tau-1}(p, c) \tag{10}$$

$$\sigma_c^{\tau}(\mathbf{z}_i) = \sum_{p=1}^{P}\mathbf{W}^{\tau-1}(p, c) \times \delta\left(f(\mathbf{z}_i^p)\right) \tag{11}$$

$$\delta\left(f(\mathbf{z}_i^p)\right) = \delta(\hat{y}_i^p) = \begin{cases} 1, & \text{if } \hat{y}_i^p = y_i \\ 0, & \text{if } \hat{y}_i^p \neq y_i \end{cases} \tag{12}$$

$\mathbf{W}^{\tau-1}$ is the weighting matrix from the last query. Following this, only samples with the highest entropy values are selected and inserted into the final set of informative samples

$$P_{WVE}^{\tau} = \left\{\mathbf{z}_j^1, \mathbf{z}_j^2, \ldots, \mathbf{z}_j{}^P : \max WVE^{\tau}(\mathbf{z}_j)\right\}. \tag{13}$$

The combination of the maximum disagreement method with the pruning strategy based on WVE allows us to select samples having the highest disagreement level while simultaneously exhibiting poor classification performance.

D. Decision Fusion Strategy

After the learning is accomplished, each source-specific MKL classifier can output the predicted labels as well as their corresponding posterior probabilities for all samples. Following this, a decision fusion strategy is applied to perform classification per pixel. Decision fusion can occur either at the class label

level, known as hard fusion, or at the posterior probability level, known as soft fusion.

Majority voting—a popular approach to conduct *hard decision fusion*, achieves the final classification decision based on a vote over individual class labels from each classifier in the ensemble. A simple majority voting is given by

$$w = \underset{i\in\{1,2,\ldots,N_c\}}{\arg\max} \; N(i) \tag{14}$$

where w is the class label from one of the N_c possible classes for the test sample, and $N(i)$ is the number of times that the class i predicted by the ensemble of classifiers.

Soft decision fusion makes the use of posterior probabilities for making the final decision. A popular soft decision fusion scheme is a linear opinion pool, which makes the final classification decision by constructing a global membership function by using individual posterior probabilities $p_j(w_i|\mathbf{x})$ of each classifier

$$P(w_i|\mathbf{x}) = \sum_{j=1}^{P} \alpha_j p_j(w_i|\mathbf{x}) \tag{15}$$

$$w = \underset{i\in\{1,2,\ldots,N_c\}}{\arg\max} \; P(w_i|\mathbf{x}) \tag{16}$$

where $\alpha_j \; (j = 1, 2, \ldots, P)$ is the classifier weight, which can either be uniformly distributed over all classifiers, or can be assigned based on the "confidence score" of each classifier. In this paper, we use the uniformly distributed weight for each source classifier.

III. Multiple Feature Extraction

The EnsembleMKL-AL approach proposed in this paper is particularly suited for multisource data. We use the terminology multisource to loosely refer to data obtained via different sensors, or different feature-types derived from the same sensor. Although it can be applied to any multisource scenario, in this paper, we focus on one important remote sensing scenario—fusion of hyperspectral and LiDAR data. Although the spectral signatures derived from hyperspectral images can be directly used in the successive steps of the framework, we extract additional features from the LiDAR data before using it in this framework. In particular, a digital surface model (DSM) and pseudo-waveforms are generated as additional feature streams. Additionally, spatial features are extracted from both sources. The inclusion of spatial information is expected to stabilize per-pixel features, and to provide additional "views" in our ensemble. We choose to include two potentially diverse spatial features—object-based textural features and extended morphological attribute profiles (EMAPs). A brief description of each type of feature set is presented in the following sections.

A. LiDAR Data Processing

Among several data products that can be extracted from discrete return LiDAR data, DSM and pseudo-waveforms are considered in this paper. A DSM is one of the most popular and simple data products that can be generated from the discrete return LiDAR data. In addition to the DSM, we also generate pseudo-waveforms over the same grid structure as used in the DSM and the hyperspectral data by stacking voxels with 1 m vertical dimension on the grid and accumulating points within every voxel. We refer readers to [33] for a detailed description of the pseudo-waveform generation process. Although the pseudo-waveform generation approach is adopted in this study since only discrete return LiDAR data are available over the study area, LiDAR data from advanced full waveform LiDAR systems could be used instead by applying waveform decomposition [34].

B. Object-Based Texture Features

Spatial features are usually extracted by considering a window-based approach, which, however, suffers from the "border-effect"—an issue where the neighborhood includes pixels from multiple objects/thematic classes. This problem can be mitigated by following an object-based approach, in which the neighborhood system is defined in an adaptive way. For this purpose, we build upon our recent work [33], integrating it into the proposed multisource AL strategy. The approach entails three key steps. 1) The original data are subdivided into spatially homogeneous regions using the *HSeg algorithm* [35]. HSeg is a segmentation approach that combines region growing, which produces spatially connected regions, with clustering, which results in groupings based on spectral similarity from spatially disjoint regions. Two main factors influence the merging process: the dissimilarity criterion (DC) and the weighting factor S_{wght}, which ranges from 0 to 1 and sets the relative importance of spatially adjacent regions with respect to those that are nonadjacent. The output is a hierarchy of segmentation maps at different levels of detail. 2) Following this, an unsupervised strategy of pruning is applied to remove subtrees of the hierarchy that are homogeneous with respect to a given homogeneity criterion. In this way, the final segmentation does not represent one of the actual levels of the hierarchy, but incorporates regions potentially selected from different levels. This is accomplished by characterizing each region of the hierarchy in terms of second-order statistics (standard deviation is considered in our specific case). The homogeneity criterion is calculated adaptively for each pixel by computing the standard deviation in a window with size $W = [w_x, w_y]$. 3) Finally, texture features (mean and standard deviation are considered in our work) are extracted from the object-based detected regions.

C. Extended Multiattribute Profiles

Morphological attribute filters have been applied to extract morphological features in many recent remote sensing applications [36], [37]. Profiles are computed by removing the connected components that do not fulfill a specified criterion. The value of an arbitrary attribute $attr$ measured on a component is compared to a given reference value λ, e.g., $T(Com) = attr(Com) > \lambda$. If the criterion is satisfied, then the regions are kept intact; otherwise, they are set to the gray level of a darker or brighter surrounding region. Such attributes can be geometric (e.g., area, shape, length of the perimeter, image moments), or

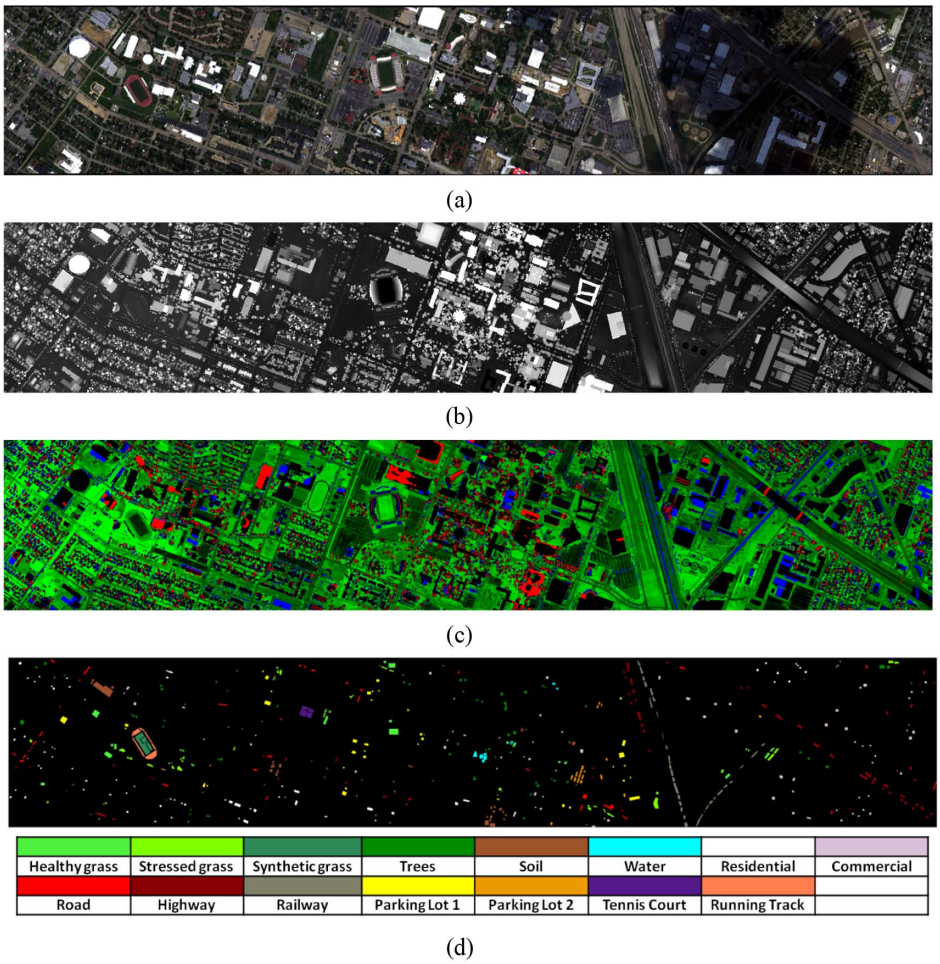


Fig. 2. UH dataset. (a) True-color composite of the hyperspectral data (wavelength R: 640.7 nm, G: 550.2 nm, B: 459.6 nm). (b) LiDAR DSM data. (c) False-color composite of the pseudo-waveform data (elevation R: 10 m, G: 0 m, B:5 m); (d) Ground reference map.

textural (e.g., range, standard deviation, entropy), etc. Attribute profiles (APs) are an extension of the widely used morphological profiles (MPs). Analogous to the definition of the MPs, APs consist of n morphological attribute thickening (ϕ^T) and n attribute thinning (γ^T) operators as given by

$$\mathbf{AP}(f) = \{\phi_n^T(f), \ldots, \phi_1^T(f), f, \gamma_1^T(f), \ldots, \gamma_n^T(f)\} \quad (17)$$

where f is the input image. Each AP can be computed on one of the features from a multivariate image (e.g., the first c principal components of a hyperspectral image), and different APs can be combined as an extended attribute profile (EAP). Also, according to the attributes considered, different morphological information can be extracted from the image, and merged into a single data structure denoted as EMAP

$$\mathbf{EMAP} = \left\{\mathbf{EAP}_1, \mathbf{EAP}'_2, \ldots, \mathbf{EAP}'_n\right\} \quad (18)$$

where each $\mathbf{EAP}$ corresponds to a specific attribute and $\mathbf{EAP}'$ consists of all thickening and thinning operators, excluding the multiple presence of the input image f or the c principle components which have already been included in $\mathbf{EAP}_1$.

IV. Experimental Results

A. Dataset and Extracted Features

The dataset [38] used in this work is a hyperspectral image and discrete return LiDAR data, which were acquired over the University of Houston campus and the neighboring urban area. The hyperspectral data were acquired with the ITRES-CASI 1500 sensor, on June 23, 2012 between 17:37:10 and 17:39:50 UTC. The average altitude of the sensor was 5500 ft, which resulted in 2.5 m spatial resolution data. The hyperspectral imagery consists of 144 spectral bands ranging from 380 to 1050 nm, and was processed (radiometric correction, attitude processing, GPS processing, geo-correction etc.) to yield the final geo-corrected image cube representing at sensor spectral radiance, $SRU = \mu\text{W}/(\text{cm}^2\ \text{sr nm})$. A true color composite of the hyperspectral signatures is shown in Fig. 2(a).

The LiDAR data were acquired using an Optech Gemini sensor on June 22, 2012 between 14:37:55 and 15:38:10 UTC. The 167 kHz laser operates at 1064 nm and records up to four returns. The average height of the sensor at the time of acquisition was 2000 ft above ground level, which resulted in an average point density of 35.38 points/m^2 on the ground. The DSM and pseudo-waveform data were generated from the original LiDAR point cloud as described in Section III-A and are

TABLE I
NUMBER OF LABELED SAMPLES FOR THE UH DATASET

Class	Class name	# samples	Class	Class name	# samples
1	Healthy grass	1251	9	Road	1252
2	Stressed grass	1254	10	Highway	1227
3	Synthetic grass	697	11	Railway	1235
4	Trees	1244	12	Parking Lot 1	1233
5	Soil	1242	13	Parking Lot 2	469
6	Water	325	14	Tennis Court	428
7	Residential	1268	15	Running Track	660
8	Commercial	1244			

shown in Fig. 2(b) and (c) respectively. The pseudo-waveform data are represented in 80 bins, which correspond to the LiDAR aggregated discrete returns within predefined voxels at predetermined elevations above/below the ground. The quantization unit of elevation spacing is 1 m, and the first band corresponds to 9 m below ground elevation, so the 80th band corresponds to 70 m above ground elevation. For this urban area, most objects are located in the range of 0–20 m, so some bands are "empty." From Fig. 2(c), we can see most of the objects are in the green color range, which means that they are close to the ground. The mid-elevation objects are represented as blue, and taller objects are in red. All images are composed of a 349×1905 grid at 2.5 m spatial resolution. The ground reference map is shown in Fig. 2(d). The total number of ground reference samples is 15 029, covering 15 classes of interest. The number of labeled samples for each class is shown in Table I. The classes include several vegetation classes, different types of roads, as well as some urban classes. The class "Grass" is separated into healthy and stressed grass, and synthetic grass. Parking lots are categorized based on whether they have cars, i.e., Parking Lot1 is empty, and Parking Lot2 is filled with cars.

We retain original features (hyperspectral signatures and pseudo-waveforms) in our set of features. In order to extract textural features, the HSeg algorithm was applied to LiDAR pseudo-waveform and hyperspectral data by adopting four-neighborhood connectivity. For both cases, we considered a Spectral Angle Mapper (SAM) based dissimilarity criterion (DC) and fixed the parameter S_{wght} to 0.1. The strategy of segmentation hierarchy pruning was applied by setting the parameter $W = [w_x, w_y]$ equal to $[3, 3]$. The number of segments obtained after pruning LiDAR pseudo-waveform and hyperspectral data were 67 735 and 55 802, respectively.

EMAPs were extracted from LiDAR DSM data and hyperspectral signatures. While EMAP features can be computed directly from the single-band LiDAR DSM data, for hyperspectral signatures they were extracted from the first four principal components which contain 99% of the total variance of the original data. Three APs were computed for both LiDAR DSM and hyperspectral data considering different attributes related to the geometry of a region. The AP associated with area is a surrogate for the scale of the structures in the scene, which is related to the size of the regions. The length of the diagonal of the bounding box is a different measure of the size and geometrical properties of region. The moment of inertia attribute, which models the elongation of the regions, is a measure of the noncompactness of the objects. As in [37], the values of λ used in each EAP are: 1) area of the regions, $\lambda_a = [100, 500, 1000, 5000]$; 2) length of the diagonal of the box bounding the region, $\lambda_d = [10, 25, 50, 100]$; and 3) moment of inertia, $\lambda_d = [0.2, 0.3, 0.4, 0.5]$. Note that the range of *optimal* parameters is expected to be data-dependent. In this work, we experimentally determined these values to be appropriate. Thus, for the LiDAR DSM (hyperspectral) data, each EAP is 9 (36)-dimensional, i.e., it is composed of one (four) APs with nine levels computed on each component. The final EMAP is obtained by stacking the three EAPs into a single data structure and by considering the original LiDAR DSM (principal components) just one time.

TABLE II
SOURCE/FEATURE TYPES WITH CORRESPONDING NUMBER OF FEATURES

Feature type	Source 1: Hyperspectral	Source 2: LiDAR
Orig	144	26 (pseudo-waveform)
EMAP	100	25
Texture	288	52

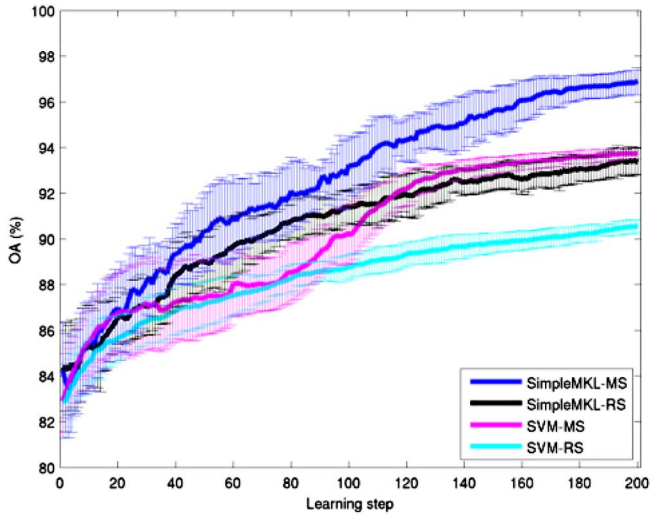


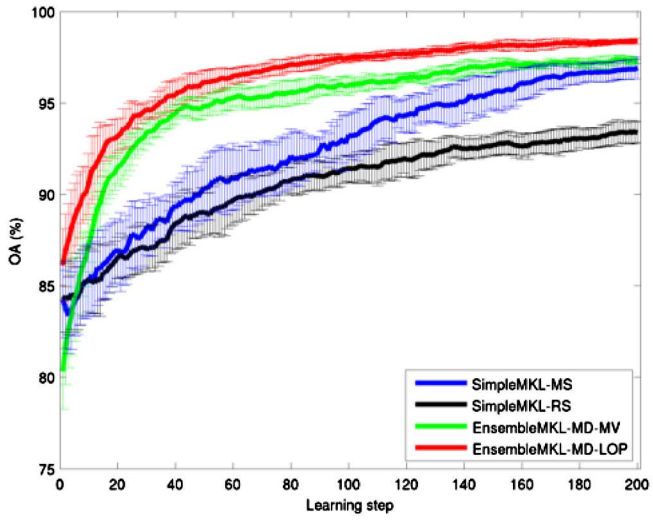
Fig. 3. OA achieved on the UH dataset for SimpleMKL and SVM methods. RS: random sampling; and MS: margin sampling.

We summarize the source/feature types with the corresponding number of features in Table II. In the rest of the paper, if not indicated differently, all the six source/feature types are considered.

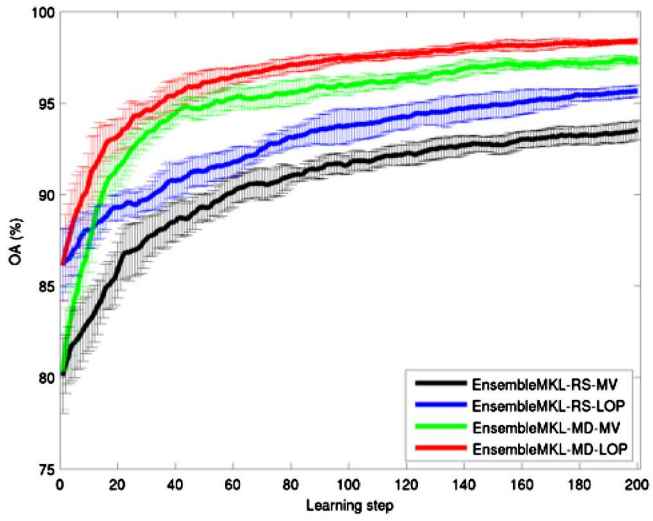
B. Experimental Setting

We present results of three experiments to demonstrate the efficacy of the proposed approach. First, we compare SimpleMKL-AL with standard single kernel SVM-AL and verify that MKL is a suitable classifier for multiple feature AL. In the second set of experiments, we compare the proposed EnsembleMKL-AL system to SimpleMKL-AL to investigate the benefit of using ensemble classifiers. Finally, we quantify the efficacy of multiple features, especially the morphological and textual features, utilized in AL.

From the available labeled data, half of the samples were selected randomly as our query set. The remaining pixels constituted the test set. Ten randomly sampled splits were used, and



(a)



(b)

Fig. 4. OA achieved on the UH dataset for (a) SimpleMKL and EnsembleMKL and (b) EnsembleMKL-RS and EnsembleMKL-MD methods. RS: random sampling; MS: margin sampling; and MD: maximum disagreement.

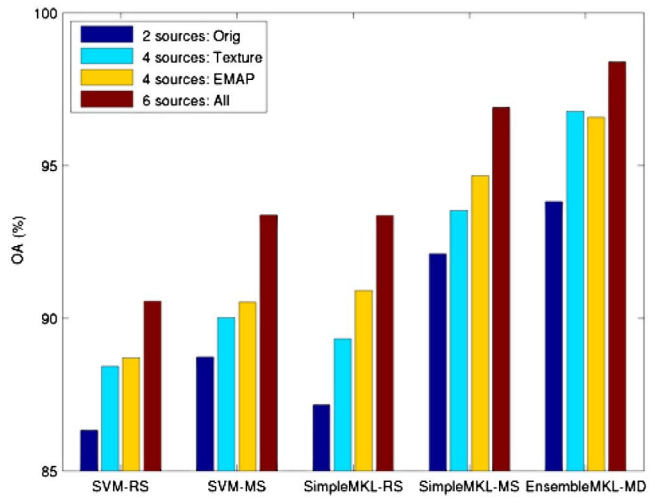


Fig. 5. OA achieved on the UH dataset using a different number of sources at the final AL step.

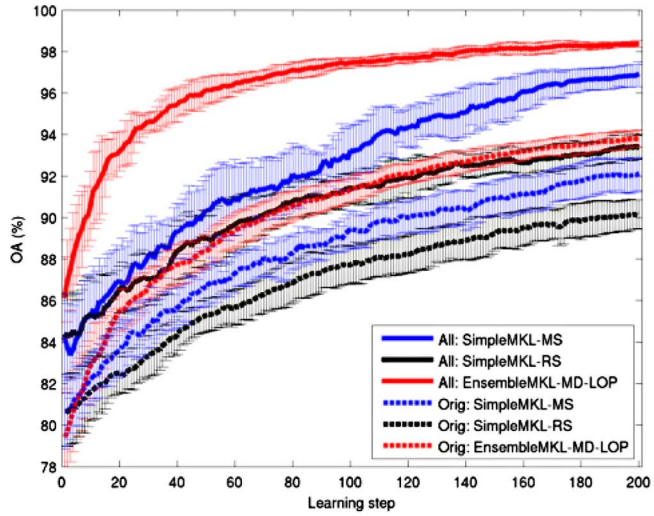


Fig. 6. OA achieved on the UH dataset for SimpleMKL and EnsembleMKL methods by considering different number of sources. Orig: 2 sources; All: 6 sources.

the average results over these random splits are reported in all experiments. The initial training set contained 20 samples for each class randomly selected from the query set. At each learning step, five samples were selected from the candidate pool and added to the training set based on learning system specific query criteria. For the single-classifier learning system, the criterion was margin sampling (MS) and the baseline was random sampling (RS). For the ensemble classifier learning system, we employed maximum disagreement (MD) for AL, and a final decision was made based on majority voting (MV) or linear opinion pool (LOP) fusion.

All experiments were conducted using an RBF kernel function with relative width parameter σ. For the standard SVM, this parameter was estimated before the learning process by applying kernel alignment [39] to the initial training set. Starting from 0.05 and with a step size of 0.05, the alignment of the kernel was maximized for $\sigma = 0.9$. For the MKL-based experiments, we did not select a specific kernel parameter; instead, we defined a set of different values as candidate input parameters. In a multisource scenario, we can build several basis kernels with different values of σ for each source; however, the number of parameters should be kept small to reduce the computational complexity and memory requirements. In particular, four base kernels with $\sigma = [0.2, 0.5, 1, 1.5]$ were considered for all sources. This range of values was found to be reasonable after applying kernel alignment to the initial training set of each source. For all classifiers (i.e., standard SVM and MKL), the penalty parameter C was selected by cross-validation in the range of $[2^{-1}, \ldots, 2^{15}]$.

C. Experimental Results and Discussion

1) Comparison of SimpleMKL-AL and SVM-AL: The first experiment compares results obtained by the SimpleMKL-AL and SVM-AL algorithms, to investigate the potential of MKL-AL in processing a large number of features obtained from different sensors using different spatial feature extraction

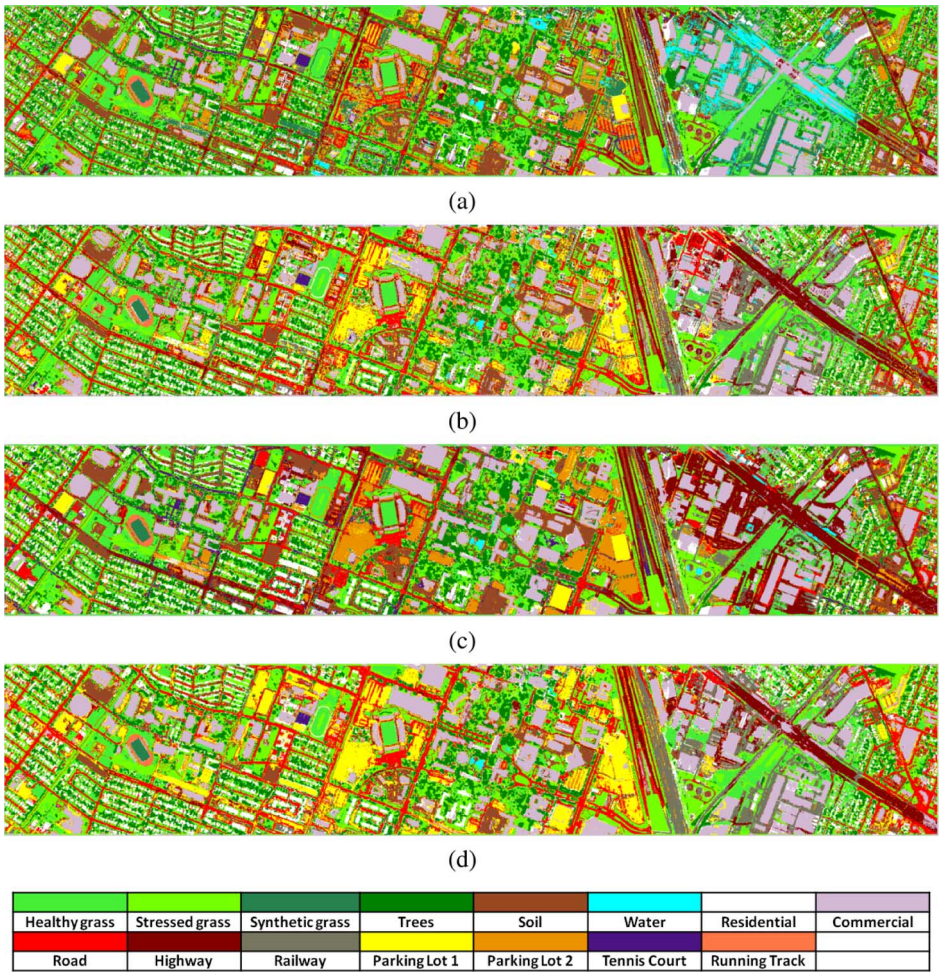


Fig. 7. Classification maps obtained at the final AL step. (a) SimpleMKL-MS: 2 sources. (b) EnsembleMKL-MD-LOP: 2 sources. (c) SimpleMKL-MS: 6 sources. (d) EnsembleMKL-MD-LOP: 6 sources.

strategies. The learning curves of the different AL methods are shown in Fig. 3.

In general, the SimpleMKL-AL methods are superior to the standard SVM-AL methods for both MS and RS query criteria throughout the learning processes. Initially, when the training samples are randomly selected and the number is the same for each class, the benefit of SimpleMKL-AL learners is not obvious. For the SimpleMKL-AL learners, the overall average accuracy is 84.24%, compared to 82.92% of SVM-AL learners. With increasing learning steps, SimpleMKL learners start to show advantage, especially for the MS strategy. This is because SimpleMKL is able to optimize a combination of kernels that jointly maximize the sum of margins, which are crucial to determine the hyperplane between classes. At the 200th learning step, the overall accuracies for SimpleMKL-MS and SimpleMKL-RS are 96.90% and 93.35% respectively, compared to 93.37% and 90.55% for SVM-MS and SVM-RS, respectively. We can hence conclude that SimpleMKL has greater potential for handling a larger number of features extracted from different sources than standard SVM.

2) Comparison of Ensemble and Single AL system: In order to adapt a classifier specific for each data source and improve the total performance of AL, we develop a source-specific MKL-AL algorithm using ensemble classifiers. As noted previously, MS is a good strategy for single-classifier AL system, but it is not suitable for an ensemble system, because it cannot exploit the potentially diverse information across different sources. In the second part of the experiments, the proposed MD-based EnsembleMKL method is compared with the single-classifier MS-based SimpleMKL method. For EnsembleMKL, two decision fusion strategies (i.e., MV and LOP) are applied after each learning step to generate the classification map and assess the classification accuracies. For completeness, the RS criterion is also considered. The obtained results are reported in Fig. 4(a) and (b).

In Fig. 4(a), results from the EnsembleMKL-MD methods are compared to those from the single-classifier-based SimpleMKL-AL methods. The proposed method achieved higher overall accuracies in general. Among all learning strategies, the EnsembleMKL-MD-LOP performs the best with an overall accuracy of 86.13% at the beginning with 20 samples each class and 98.38% after 200 iterations. EnsembleMKL-MD-MV learns no better than the SimpleMKL-MS learner at the first 5 steps, but it starts to improve after that. Both ensemble methods show significant improvements compared to the baseline SimpleMKL-RS. The increase of the accuracy is 5.00%

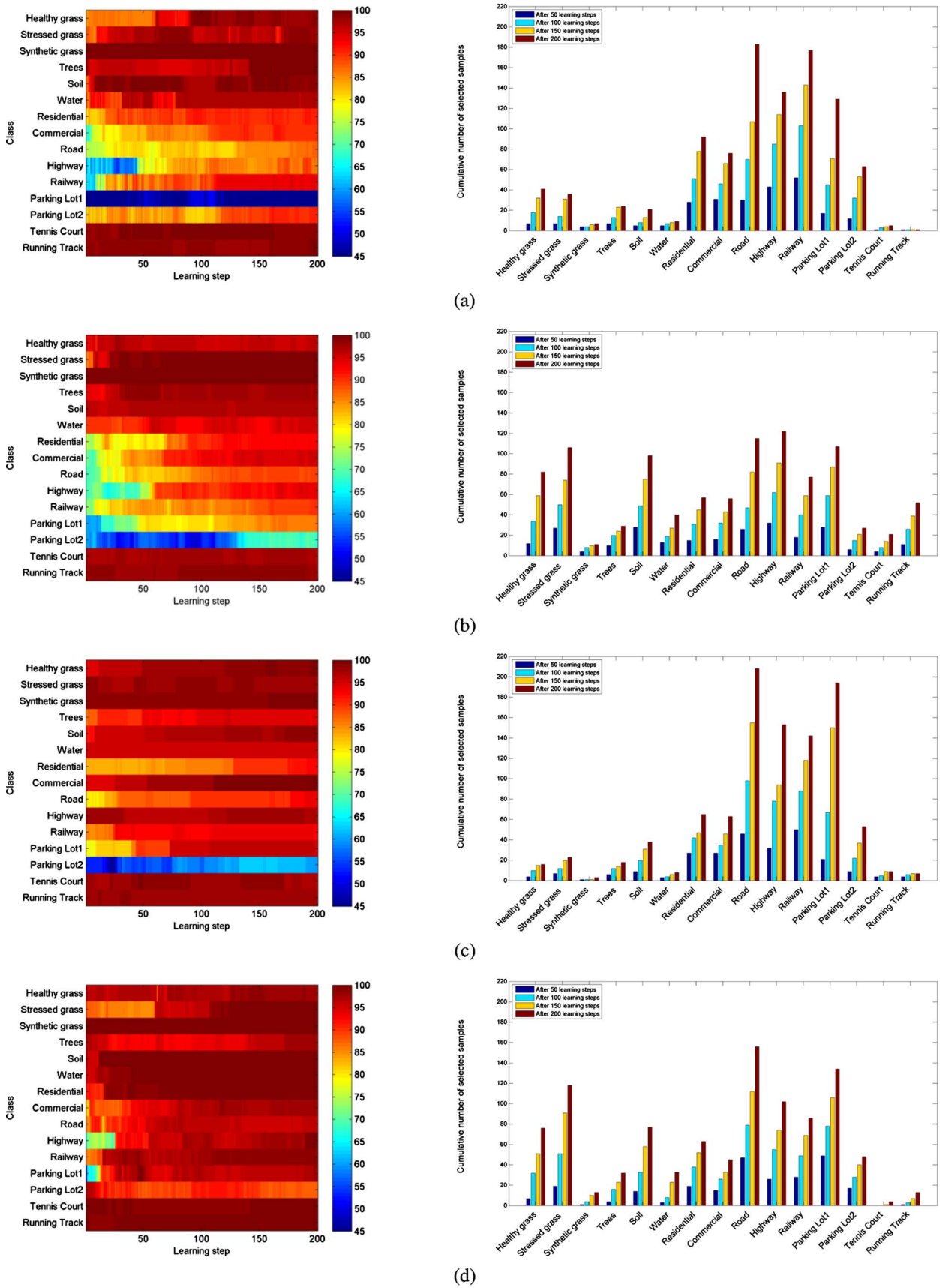
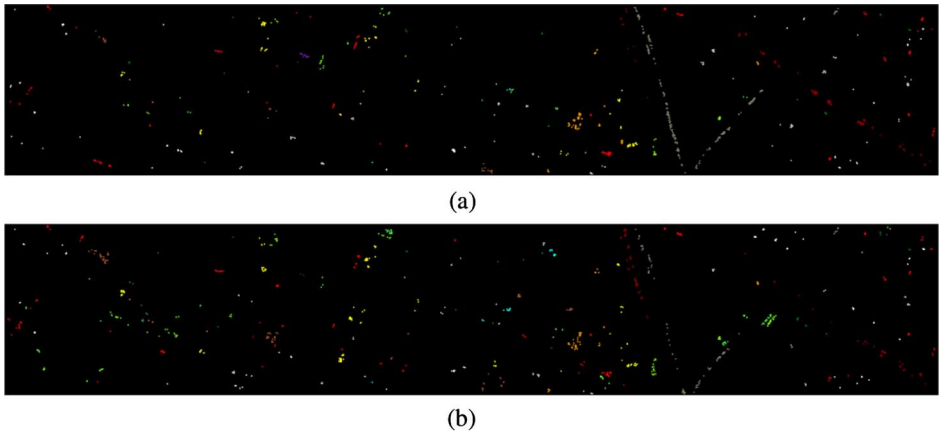


Fig. 8. Class specific accuracies (left) and cumulative number of selected samples (right) at different learning steps. (a) SimpleMKL-MS: 2 sources. (b) EnsembleMKL-MD-LOP: 2 sources. (c) SimpleMKL-MS: 6 sources. (d) EnsembleMKL-MD-LOP: 6 sources.



(a) (b)

Fig. 9. Spatial locations of newly selected samples at the final AL step for (a) SimpleMKL-MS and (b) EnsembleMKL-MD-LOP. All six sources were used.

TABLE III
CLASS ACCURACIES (%), OVERALL ACCURACIES (OA%), AND STANDARD DEVIATIONS (STD.%) OF DIFFERENT AL METHODS

	Original features					All features				
Class index	SVM RS	SVM MS	SimpleMKL RS	SimpleMKL MS	EnsembleMKL MD-LOP	SVM RS	SVM MS	SimpleMKL RS	SimpleMKL MS	EnsembleMKL MD-LOP
1	96.87	97.75	97.96	97.65	98.28	97.27	98.80	98.10	**99.22**	98.12
2	96.76	96.74	97.21	98.29	98.60	97.45	97.37	98.87	98.14	**99.69**
3	96.06	97.06	97.75	97.06	98.13	98.23	99.39	98.27	99.38	**100**
4	94.29	95.88	94.86	95.81	95.43	96.21	96.16	97.24	94.29	**98.04**
5	96.69	97.72	97.72	98.04	97.44	97.15	98.63	98.41	98.72	**99.52**
6	92.15	94.87	93.51	98.08	98.08	93.15	95.36	93.21	95.51	**99.36**
7	89.04	91.83	89.77	91.93	93.67	90.11	92.83	90.31	92.41	**99.84**
8	87.89	88.98	89.16	95.08	96.19	90.77	90.94	94.31	**99.68**	97.94
9	80.87	88.99	82.88	90.15	84.33	88.59	93.53	91.17	94.18	**96.28**
10	86.53	86.98	87.28	89.51	96.18	93.05	96.12	92.28	97.30	**99.68**
11	90.45	89.89	85.88	89.45	88.07	90.74	92.16	91.95	93.28	**98.66**
12	73.94	73.24	70.00	59.83	90.17	75.06	77.39	73.19	95.59	**98.14**
13	64.53	79.62	77.36	84.53	65.09	78.18	83.31	79.77	65.52	**87.92**
14	95.23	96.23	95.32	95.32	98.58	97.56	98.73	96.62	96.68	**99.53**
15	94.13	95.44	95.55	96.55	99.42	96.55	96.01	98.93	97.97	**99.71**
OA	86.32	88.72	87.16	92.10	93.81	90.55	93.37;	93.35	96.90	**98.38**
Std.	0.49	0.62	0.69	0.8	0.41	0.27	0.24	0.58	0.60	**0.15**

and 3.95%, respectively, for LOP and MV at the final step. Furthermore, it is much faster for EnsembleMKL learners to reach a high accuracy and converge than SimpleMKL learner. In Fig. 4(b), we compare the MD and RS criteria in conjunction with EnsembleMKL method. It is evident for both decision fusion strategies (i.e., MV and LOP) that the MD-based AL strategy produces higher accuracies than RS. Therefore, the MD criterion, which has previously been shown to work well as a multiview method for single-source hyperspectral data [9], has been demonstrated to be also suitable for this multisource scenario.

3) Comparison of the number of sources: In this section, we compare the performance of different AL strategies using different sources and consequently different numbers of features. Specifically, four scenarios are considered. 1) Orig: two sources represented by the original hyperspectral and LiDAR pseudo-waveform data. 2) Texture: texture features extracted from hyperspectral and LiDAR pseudo-waveform in addition to the original data for a total of four sources. 3) EMAP: EMAP features extracted from hyperspectral and LiDAR DSM in addition to the original data for a total of four sources. 4) All: all the six sources.

A comparison of the overall accuracies at final AL step, i.e., by considering 1300 training samples, using different numbers of sources is shown in Fig. 5. OA increases as more sources and more diverse features are included in the system. It thus justifies our assumption that a combination of different types of features can provide complementary information of land-cover which is useful for classification.

Specially, Fig. 6 shows the improvements of using all features compared to the original features during the learning. From the results, it is clearly shown that the performance is much better when all the features are included in both single and ensemble AL systems. The improvement is 4%–5% for AL learners with all features compared to that with original features.

4) Class specific analysis: Classification maps obtained at the final step of different AL approaches are reported in Fig. 7. The class accuracies and statistical number of samples selected from each class by different MKL-AL strategies are shown

in Fig. 8. There is a general trend of accuracy increase for most of the classes when more sources and more classifiers are involved. From the results in Fig. 8(a) and (c), most of the samples selected by SimpleMKL-MS are from the classes with low accuracies at the beginning, such as Class 9: Road, Class 10: Highway, and Class 12: Parking Lot 1. When samples from these classes are added to the training set, there is an obvious increase of accuracy during the learning. For example, in Fig. 8(c), the accuracy of Class 9 is 74.49% at the beginning when only 20 randomly selected samples are used. As the learning progresses, the accuracy increases to 84.47% and 95.59% after 100 and 200 iterations, and the corresponding number of selected samples from Class 9 is 98 and 208 respectively. For EnsembleMKL-AL in Fig. 8(b) and (c), the selection of samples from different classes seems to be more uniform than in SimpleMKL-AL. As a result, it increases the accuracies of most classes and thus overall accuracy as well. The spatial locations of newly selected samples at the final AL step for SimpleMKL-MS and EnsembleMKL-MD-LOP methods are shown in Fig. 9. From the results, we found that the selection of samples from different classes using Ensemble MKL-AL seems to be more uniform than using SimpleMKL-AL. This could be the reason that Ensemble MKL-AL converges much faster and has a higher accuracy.

In Fig. 8(b), morphological features and object-based texture features are added to the single-classifier MKL-AL system, indicating that the combination of spatial features can improve the classification performance of most classes, especially for Road, Highway, Railway, and both Parking Lots. Comparing class-specific accuracies in Fig. 8(c) and (d), it is clear that the ensemble system with multiple features can further improve the accuracies of most classes. The most significant effect of bringing multiple features in this ensemble system is the improvement in discrimination between the residential and commercial area as well as different types of roads. This is due to different morphological and textures present in these urban structures. To be more specific, the residential and commercial buildings are typically differentiated by their shapes and sizes, and these properties may be directly related to different scales of structuring elements in EMAPs. The railway and highway classes are morphologically similar, but are constructed with different materials and thus texture features can help to differentiate these classes to some extent. Table III summarizes the class specific accuracies, overall accuracy, and standard deviations for different AL approaches at the final learning step.

5) Computational overhead: We conclude the experimental analysis by empirically evaluating the computational complexity associated with the different methods investigated in this paper. All the experiments were implemented in MATLAB R2012a on a Linux system with twelve 3.2 GHz Intel(R) processors and 32 GB RAM. The SimpleMKL toolbox [40] was adopted for implementing single kernel SVM and MKL approaches. The total processing time, which includes model selection, training phase and sample selection, for running the 200 steps of the AL process and by considering the six different sources was 9.35×10^3, 7.88×10^3, and 8.19×10^3 s for SVM-MS, SimpleMKL-MS, and EnsembleMKL-MD-LOP, respectively.

V. Conclusion

Multisource remote sensing data require a large quantity of labeled data to train robust classifiers—in practice, this is difficult and often an expensive endeavor. In this paper, we presented an ensemble multiple kernel-based AL system that incorporates different types of features extracted from multisensor remote sensing data for classification. Data from different sources provide the necessary diversity which is a crucial point for constructing the classifier committee in AL. This framework provides a new way to exploit multisensor, multifeature remote sensing datasets for image classification.

In particular, we adopted a maximum disagreement principle for ensemble AL. Inside the learning system, probabilistic MKL was applied as an adaptive classifier, since it can implicitly adapt the kernel to the data by learning appropriate weights of predetermined kernels, eliminating the need to re-tune SVMs at each step of the AL process. Finally, a decision fusion strategy was performed to make a final decision on the probabilistic output and obtain the final classification map. The experiments validated the efficacy of the proposed framework and provided the following conclusions. 1) MKL is a more effective and appropriate classifier for multisource AL compared to the standard SVM classifier. 2) Ensemble classifiers improve the performance of traditional AL substantially for this multisource data. The proposed EnsembleMKL-AL system greatly outperforms the SimpleMKL-AL approach in terms of overall and class-specific accuracies. 3) The inclusion of different types of sources and features, including spectral features from hyperspectral image, elevational information from LiDAR data, morphological profiles, and object-based textural features from multisensor data can provide a truly "multiview" approach to AL and image classification for a more robust classification of the scene. We note how the maximum disagreement AL criterion adopted in this paper works on the estimated labels, while posterior probabilities are considered to fuse the predictions given by the different classifiers and generate the final classification map. Further improvements of the proposed solution can be potentially obtained by considering posterior probabilities in both parts of the process.

Acknowledgment

The authors would like to thank the IEEE GRSS Data Fusion Technical Committee for providing the UH multisensor datasets, the NSF Funded National Center for Airborne Laser Mapping (NCALM) at the University of Houston for acquiring the data, and X. Zhou, M. Cui, A. Singhania, and J. Fernandez for preprocessing the data.

References

[1] B. Settles, "Active learning literature survey." Madison, WI, USA: University of Wisconsin, 2010, vol. 52, pp. 55–66.

[2] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 606–617, Jun. 2011.

[3] M. Crawford, D. Tuia, and H. Yang, "Active learning: Any value for classification of remotely sensed data?" in *Proc. IEEE*, vol. 101, no. 3, pp. 593–608, Mar. 2013.

[4] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proc. 7th Int. Conf. Mach. Learn. (ICML)*, 2000, pp. 839–846.

[5] P. Mitra, B. Uma Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recog. Lett.*, vol. 25, no. 9, pp. 1067–1074, Jul. 2004.

[6] E. Pasolli, F. Melgani, and Y. Bazi, "Support vector machine active learning through significance space construction," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 431–435, May 2011.

[7] H. L. Yang, Y. Zhang, S. Prasad, and M. Crawford, "Multiple kernel active learning for robust geo-spatial image analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2013, pp. 1218–1221.

[8] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 287–294.

[9] W. Di and M. M. Crawford, "View generation for multiview maximum disagreement based active learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5, pp. 1942–1954, May 2012.

[10] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.

[11] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.

[12] T. Luo *et al.*, "Active learning to recognize multiple types of plankton," in *Proc. IEEE Int. Conf. Pattern Recog. (ICPR)*, vol. 3, 2004, pp. 478–481.

[13] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery, "SVM active learning approach for image classification using spatial information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 2217–2233, Apr. 2014.

[14] A. Stumpf, N. Lachiche, J.-P. Malet, N. Kerle, and A. Puissant, "Active learning in the spatial domain for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2492–2507, May 2014.

[15] D. Tuia and J. Munoz-Mari, "Learning user's confidence for active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 872–880, Feb. 2013.

[16] E. Pasolli, F. Melgani, N. Alajlan, and N. Conci, "Optical image classification: A ground-truth design framework," *IEEE Trans. Geoscie. Remote Sens.*, vol. 51, no. 6, pp. 3580–3597, Jun. 2013.

[17] A. Liu, G. Jun, and J. Ghosh, "Spatially cost-sensitive active learning." in *SDM*. SIAM, 2009, pp. 814–825.

[18] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[19] J. Li, P. Reddy Marpu, A. Plaza, J. Bioucas-Dias, and J. A. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.

[20] G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, May 2004.

[21] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, Dec. 2006.

[22] D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski, "Learning relevant image features with multiple-kernel classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3780–3791, Oct. 2010.

[23] Y. Gu *et al.*, "Representative multiple kernel learning for classification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 7, pp. 2852–2865, Jul. 2012.

[24] G. Thoonen, Z. Mahmood, S. Peeters, and P. Scheunders, "Multisource classification of color and hyperspectral images using color attribute profiles and composite decision fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 510–521, Apr. 2012.

[25] A. Rakotomamonjy *et al.*, "Simplemkl," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.

[26] S. Prasad and L. M. Bruce, "Decision fusion with confidence-based weight assignment for hyperspectral target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1448–1456, May 2008.

[27] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[28] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[29] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, Mar. 1999.

[30] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on platts probabilistic outputs for support vector machines," *Mach. Learn.*, vol. 68, no. 3, pp. 267–276, Oct. 2007.

[31] I. Muslea, S. Minton, and C. A. Knoblock, "Active learning with multiple views," *J. Artif. Intell. Res.*, vol. 27, pp. 203–233, Oct. 2006.

[32] W. Di and M. M. Crawford, "Multi-view adaptive disagreement based active learning for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2010, pp. 1374–1377.

[33] J. Jung, E. Pasolli, S. Prasad, J. C. Tilton, and M. M. Crawford, "A framework for land cover classification using discrete return lidar data: Adopting pseudo-waveform and hierarchical segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 2, pp. 491–502, Feb. 2014.

[34] J. Jung and M. M. Crawford, "Extraction of features from lidar waveform data for characterizing forest structure," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 492–496, May 2012.

[35] J. Tilton, Y. Tarabalka, P. Montesano, and E. Gofman, "Best merge region-growing segmentation with integrated nonadjacent region object aggregation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4454–4467, Nov. 2012.

[36] M. Dalla Mura, J. Atli Benediktsson, L. Bruzzone, and B. Waske, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

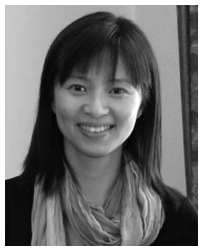
[37] M. Dalla Mura, A. Villa, J. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.

[38] UH Dataset. [Online] Available: http://hyperspectral.ee.uh.edu/?page_id=459.

[39] N. Shawe-Taylor and A. Kandola, "On kernel target alignment," *Adv. Neural Inf. Process. Syst.*, vol. 14, pp. 367–373, 2002.

[40] SimpleMKL Toolbox. [Online] Available: http://asi.insa-rouen.fr/enseignants/arakoto/code/mklindex.html

Yuhang Zhang (S'12) received the B.S. and M.S. degrees in information and communication engineering from Harbin Institute of Technology, Harbin, China, in 2011 and 2013, respectively. She is currently pursuing the Ph.D. degree in electrical and computer engineering at the University of Houston, Houston, TX, USA. She is a Graduate Research Assistant with the Hyperspectral Image Analysis Group, Electrical and Computer Engineering Department, University of Houston. Her research interests include machine learning, signal and image processing with applications in hyperspectral remote sensing, and brain–machine interfaces.



Hsiuhan Lexie Yang (S'10) received the M.S. degree in geomatics from the National Taiwan University, Taipei City, Taiwan, in 2007, and the Ph.D. degree in civil engineering from Purdue University, Lafayette, IN, USA, in 2014. She is currently a Postdoctoral Researcher with the School of Civil Engineering, Purdue University, West Lafayette, IN, USA. Her research interests include statistical pattern recognition, algorithm developments for geospatial data analysis, manifold learning for hyperspectral data analysis, machine learning, and its related applications to remote sensing data.

Saurabh Prasad (S'05–M'09–SM'14) received the B.S. degree in electrical engineering from Jamia Millia Islamia, New Delhi, India, in 2003, the M.S. degree in electrical engineering from Old Dominion University, Norfolk, VA, USA, in 2005, and the Ph.D. degree in electrical engineering from Mississippi State University, Starkville, MS, USA, in 2008. Currently, he is an Assistant Professor with the Electrical and Computer Engineering Department, University of Houston (UH), Houston, TX, USA. His research interests include statistical pattern recognition, adaptive signal processing, and kernel methods for medical imaging, optical, and synthetic aperture radar remote sensing. Dr. Prasad is an active Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and the *Elsevier Pattern Recognition Letters*. He was the recipient of Geosystems Research Institutes Graduate Research Assistant of the Year Award in May 2007, and the Office-of-Research Outstanding Graduate Student Research Award in April 2008 at Mississippi State University. In July 2008, he received the Best Student Paper Award at the IEEE International Geoscience and Remote Sensing Symposium 2008 held in Boston, MA, USA. In October 2010, he received the State Pride Faculty Award from Mississippi State University for his academic and research contributions. In 2014, he received the NASA New Investigator (Early Career) Award.

Edoardo Pasolli (S'08–M'13) received the M.Sc. degree in telecommunications engineering and the Ph.D. degree in information and communication technologies from the University of Trento, Trento, Italy, in 2008 and 2011, respectively. He was a Postdoctoral Fellow with the Intelligent Information Processing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, from December 2011 to September 2012, and with the Computational and Information Sciences and Technology Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA, from October 2012 to October 2013. He is currently a Postdoctoral Fellow with the School of Civil Engineering, Purdue University, West Lafayette, IN, USA. His research interests include processing and recognition techniques applied to remote sensing images and biomedical signals.

Jinha Jung received the B.S. and M.S. degrees in civil engineering from Seoul National University, Seoul, South Korea, and the Ph.D. degree in civil engineering from Purdue University, West Lafayette, IN, USA, in 2003, 2005, and 2011, respectively.

He worked as a Postdoctoral Research Associate with the Institute for Environmental Science and Policy, University of Illinois at Chicago, Chicago, IL, USA, and with the School of Civil Engineering, Purdue University. He is currently an Assistant Professor with the School of Science and Engineering, Texas A&M University-Corpus Christi, Corpus Christi, TX, USA. His research interests include advanced LiDAR data analysis for interdisciplinary research leveraging and remote sensing and geospatial science.

Melba Crawford (M'90–SM'05–F'08) received the B.S. and M.S. degrees in civil engineering from the University of Illinois, Urbana, IL, USA, in 1970 and 1973, respectively, and the Ph.D. degree in systems engineering from Ohio State University, Columbus, OH, USA in 1981. He is a Professor of Civil Engineering, Electrical and Computer Engineering, and Agronomy with Purdue University, Lafayette, IN, USA, where she is the Director of the Laboratory for Applications of Remote Sensing and the Chair of Excellence in Earth Observation. She also serves as the Associate Dean for Research with the College of Engineering. She has authored more than 150 publications in scientific journals, conference proceedings, and technical reports. His research research interests include methods for analysis of multisensor remotely sensed data, including classification, unmixing, active learning, and knowledge transfer, and the application of these methodologies. Dr. Crawford is the current President of the IEEE Geoscience and Remote Sensing Society and an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She has also served as Executive Vice President of the Society, Vice President for Conferences and Symposia, Vice President for Professional Activities, and Education Director of GRSS.