Class-Specific Sparse Multiple Kernel Learning for Spectral–Spatial Hyperspectral Image Classification

Tianzhu Liu, *Student Member, IEEE*, Yanfeng Gu, *Member, IEEE*, Xiuping Jia, *Senior Member, IEEE*, Jón Atli Benediktsson, *Fellow, IEEE*, and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—In recent years, many studies on hyperspectral image classification have shown that using multiple features can effectively improve the classification accuracy. As a very powerful means of learning, multiple kernel learning (MKL) can conveniently be embedded in a variety of characteristics. This paper proposes a class-specific sparse MKL (CS-SMKL) framework to improve the capability of hyperspectral image classification. In terms of the features, extended multiattribute profiles are adopted because it can effectively represent the spatial and spectral information of hyperspectral images. CS-SMKL classifies the hyperspectral images, simultaneously learns class-specific significant features, and selects class-specific weights. Using an L_1-norm constraint (i.e., group lasso) as the regularizer, we can enforce the sparsity at the group/feature level and automatically learn a compact feature set for the classification of any two classes. More precisely, our CS-SMKL determines the associated weights of optimal base kernels for any two classes and results in improved classification performances. The advantage of the proposed method is that only the features useful for the classification of any two classes can be retained, which leads to greatly enhanced discriminability. Experiments are conducted on three hyperspectral data sets. The experimental results show that the proposed method achieves better performances for hyperspectral image classification compared with several state-of-the-art algorithms, and the results confirm the capability of the method in selecting the useful features.**

***Index Terms*—Classification, extended multiattribute profile (EMAP), group lasso, hyperspectral images, multiple kernel learning (MKL).**

I. Introduction

RECENTLY, due to advances in remote sensing technologies, hyperspectral imaging sensors are able to capture hundreds of narrow spectral channels with a very high spatial resolution, which has led to a better identification of relatively small structures [1], [2]. Due to the high spatial resolution, the geometrical features of the structures in a scene have a great perceptual significance that can be directly exploited to model the objects in the scene. This can also increase the discriminability between different thematic classes.

To produce satisfactory classification results, one typically needs to consider the integration of spatial and spectral features. The spatial characteristics of the objects in an image can be modeled with several approaches, such as Markov random fields [3], [4], segmentation [5]–[7], 3-D Gabor feature extraction [8], superpixel [9]–[11], morphological component analysis [12], [13], adaptive sparse representation [14], [15] and so on. A widely used technique for extracting spatial features is based on mathematical morphology. Among all the operators belonging to this framework, morphological connected operators [16] proved to be suitable for extracting spatial information while preserving the geometrical characteristics of the structures in the image (i.e., without distorting the borders). A morphological profile (MP) is constructed based on the repeated use of opening and closing by reconstruction with a structuring element (SE) of increasing size (i.e., a sequence of multiscale connected operators), applied to a scalar image. MPs simultaneously remove some spatial details and preserve the geometrical characteristics of the other regions. Pesaresi and Benediktsson [17] used morphological transformations to build the so-called MP. In [18], the MP generated by morphological opening and closing operations was used for classifying a Quickbird panchromatic image. The standard opening and closing along with white and black top hats [19] and opening and closing by reconstruction were computed, and the resulting features were classified using a support-vector-machine (SVM) classifier [20]. In [21], the concept of MPs was successfully extended to handle hyperspectral images; the authors reduce the high dimensionality of the data by principal component analysis (PCA) and compute the profiles on the first principal components (PCs) extracted, leading to the definition of extended MPs (EMPs).

Some studies have been conducted to assess the capability of SEs with different shapes for the extraction of spatial information. MPs computed with a compact SE (e.g., square, disk, etc.) can be considered for modeling the size of the objects in the image. In [22], one MP is built using disk-shaped SEs for extracting the smallest size of the structures, whereas the other employs linear SEs for characterizing the object's maximum size. In our previous study, multiple SEs with different scales are employed to generate multi-EMPs to present

Manuscript received March 9, 2016; revised July 8, 2016; accepted August 7, 2016. Date of publication September 1, 2016; date of current version September 30, 2016. This work was supported in part by the National Science Fund for Excellent Young Scholars under Grant 61522107, by the Natural Science Foundation of China under Grant 61371180, by the China Aerospace Science and Technology Corporation–Harbin Institute of Technology Joint Center for Technology Innovation Fund under Grant CASC-HIT15-1C03, and by the Smart Sea Technical Innovation Foundation, CSSC-SERI.

T. Liu and Y. Gu are with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: guyf@hit.edu.cn).

X. Jia is with the School of Engineering and Information Technology, University of South Wales, Australian Defence Force Academy, Canberra, ACT 2600, Australia.

J. A. Benediktsson is with the Engineering Research Institute, University of Iceland, 107 Reykjavik, Iceland.

J. Chanussot is with the GIPSA-Lab, Grenoble Institute of Technology, 38402 Grenoble, France.

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/TGRS.2016.2600522

0196-2892 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

spatial–spectral information [23]. Although MP is a powerful technique for the extraction of spatial information, the concept has a few limitations: 1) The shape of SEs is fixed; and 2) SEs are unable to characterize information related to the gray-level characteristics of the regions. To avoid this limitation, morphological attribute filters (AFs) have been proposed as the generalization of the MP, which provides a multilevel characterization of an image by using the sequential application of a morphological AF [24]. AFs are connected operators that process an image by considering only its connected components. For binary images, the connected components are simply the foreground and background regions present in the image. To deal with grayscale images, the set of connected components can be obtained by considering the image to be composed by a stack of binary images generated by thresholding the image at all its gray-level values [25]. Thus, they process the image without distorting or inserting new edges but only by merging existing flat regions [19]. AFs include in their definition the morphological operators based on geodesic reconstruction [16]. Moreover, they are a flexible tool, since they can perform processing based on many different types of attributes. For example, they can be purely geometric (e.g., area, diagonal of the objects, image moments, and shape factors) or textural (e.g., standard deviation and entropy). In [26], the definition of the extended attribute profile (EAP) and the extended multiattribute profile (EMAP) were proposed, which rely on the application of the APs to hyperspectral data and to a straightforward further extension to a multiattribute scenario, respectively. As the EMAPs can provide richer descriptions of the spatial information contained in the scene, in this paper, we combine EMAP and the spectral features to classify the hyperspectral images. However, the EMAPs, which are built on different attributes and multiple degrees, significantly increase the dimensionality of the data being analyzed. Furthermore, the information contained in a profile is, in general, intrinsically redundant [27]. Therefore, it is important to integrate the EMAP features into a classifier that can learn the features in a selective way. As a result, choosing the most suitable attribute and range of threshold values for extracting the information on the geospatial objects certainly becomes an important and complex task.

While the use of spatial and spectral features becomes more practical for classification tasks, how to properly integrate those features is still one of the main research topics. Kernel methods have attracted the attention of many researchers. In [28], ideal regularization is proposed to utilize the label information of a data set for learning kernels. In [29], fuzzy robust kernel entropy component analysis is proposed to induce the fuzzy item to optimize the kernel entropy components, which makes the transformed data more robust. In particular, SVMs [30]–[32] have shown reasonable performances in terms of accuracy for classifying high dimensional data when a limited number of training samples are available. A more flexible learning model using multiple kernels instead of one, which is known as multiple kernel learning (MKL), has been proposed to correspond to different notions of similarity or information from multiple sources and represent differences between features [33]. The conventional MKL framework combines multiple features by constructing base kernels for each type of feature. The optimal weights for each base kernel are determined by MKL, and they indicate the contribution of the associated features. Several variants of MKL for feature fusion have also been proposed to improve the performances in terms of accuracy. A sample-screening MKL method is proposed to employ a boosting strategy for screening the limited training samples under the MKL framework [34]. In [35], an MKL method using low-rank nonnegative matrix factorization for classification of hyperspectral imagery is presented. In [36], a novel MKL model is proposed for urban classification to integrate heterogeneous features (HF-MKL) from spectral images and lidar data. In [37], a discriminative MKL method is proposed for spectral image classification to learn an optimal combined kernel from predefined basic kernels by maximizing separability in the reproducing kernel Hilbert space. A composite kernel integrates the measurement of both spectral and textural features and achieves better performances than spectral features only in terms of accuracy [38]–[40]. Rule-based MKL (RBMKL) generates the kernel via summation or multiplication of the base kernels [41]. Representative MKL (RMKL) learns a kernel matrix to capture the most variation of the original base kernels [42]. Although these MKL methods can improve the performances for many classification tasks in terms of accuracy, they need to predetermine the parameters for each kernel, and this parameter selection procedure either requires prior knowledge or results in increased computational complexity due to the need to perform cross-validation.

In this paper, we concentrate on embedding EMAP features in the class-specific sparse MKL (CS-SMKL) to classify hyperspectral data sets. In our approach, PCA is first performed on the original data sets, and then spatial features, i.e., EMAPs, are obtained by processing the PCs with different degrees of multiple AFs. For kernel learning, the base kernels are Gaussian kernels associated with EMAPs; L_1-norm of the kernel weights, also known as the simplex constraint, is used in our methods because it can lead to a sparse solution, i.e., only a few base kernels among many carry significant weights. We propose a class-specific algorithm that automatically learns an efficient feature set for the classification of any two classes. In the solving process, we adopt an efficient optimization method using the equivalence between group lasso and MKL [43] to obtain a closed-form solution for updating the kernel weights at each iteration [44]. Experimental results illustrate that the proposed CS-SMKL method provides good performances in terms of accuracy on three real high-spatial-resolution hyperspectral data sets.

In summary, the contributions of this paper are as follows:

1) We propose a CS-SMKL method to well fuse multiscale EMAPs for spatial–spectral joint classification of hyperspectral images.

   MKL provides a very effective means of learning and can conveniently be embedded in a variety of characteristics. Many previous works have demonstrated that the multiscale EMAPs are very effective features. In this paper, we propose a new multifeature fusion method for classification to integrate the multiscale EMAP features via the CS-SMKL framework. Through such an integration, the redundancy of the multiscale EMAP features

$$\mathrm{AP}(f) = \prod\nolimits_i : \left\langle \begin{array}{ll} \prod_i = \prod_{\phi^{T'_\lambda}}, \text{ with } \lambda = (n-1+i) & \forall \lambda \in [1, n] \\ \prod_i = \prod_{\gamma^{T'_\lambda}}, \text{ with } \lambda = (i-n-1) & \forall \lambda \in [n+1, 2n+1] \end{array} \right. \tag{3}$$

is well removed, and the associated weights of optimal base kernels are determined to improve the classification performances.

2) We propose a class-specific manner for the kernel learning in the classification algorithm.

   In hyperspectral image classification, the effectiveness of spatial–spectral features is class dependent. Therefore, it is more reliable to conduct learning for each individual class pair [45]. The CS-SMKL method, in which the sparse constraint is introduced via group lasso, can learn a compact set of features, and the proposed class-specific way can search the most effective features for the classification of any two classes, leading to the interpretation of the class-specific classification model. In other words, our method is able to select the class-specific weights for different attributes of MPs and remove the redundancy of those features, thus further improving the classification performances using the same input features.

The remainder of this paper is organized as follows. In Section II, EMAP is reviewed as related work. Section III details the proposed method, including the variational equivalence between MKL and group lasso, the class-specific kernel learning manner, and the framework of our proposed method. Experimental setting and results for three real hyperspectral data sets are presented in Section IV. Finally, Section V concludes this paper.

II. Extended Profiles With AFs

Here, the morphological AFs are presented at first. Subsequently, the concept of the AP is briefly recalled. Finally, the EMAP is described.

A. Morphological AFs

Morphological AFs are connected filters, and they are morphological adaptive filters introduced by Breen and Jones [16]. They process the input image by removing the connected components that do not fulfill a given criterion. The criterion could evaluate any attribute extracted from the regions. The great flexibility in defining the attribute leads to an improved capability in modeling the spatial information with respect to operators based on fixed SEs. An attribute can be any measure computable on the regions of the connected components present in the image.

Let us consider a numerical function f and a binary criterion T that acts on the sections $h_k\ (f)$ of f at successive thresholds $k_1 < k_2 < k_3$. We may have $Th_{k_2}(f) = \phi$, while $Th_k(f) \neq \phi$ for $k = k_1, k_3$. The successive transformations do not decrease as k increases. Therefore, they cannot be considered the stack of sections of a function. The simplest way to force the decrease in the sequence is to replace the transform $Th_k\ (f)$ by the union of all the transforms from the top section, i.e., by $T^*h_k(f) = \cup\{Th_p(f), p \geq k\}$, which is equal to the section at level k of the function transform $\gamma^T(f)$

$$\gamma^T(f)(x) = \max\{k : x \in Th_k(f)\}. \tag{1}$$

Analogously, as a dual counterpart, the grayscale thickening transformation can be formulated by

$$\phi^T(f)(x) = \min\{k : x \in Th_k(f)\}. \tag{2}$$

B. APs

APs are a multilevel decomposition of the input image based on AFs [24]. As for MPs, APs have to be cumulative functions. This is an important condition because it leads to achieving a progressively increased simplification of the image when the filters values are increased. In the case of openings by reconstruction, it occurs automatically when the size of the SE increases. When considering AFs, it is only verified for increasing criteria. Thus, to guarantee this property for all AFs, the family of criteria T_i taken into account must be formally ordered, so that $i \leq j \Rightarrow T_i \subseteq T_j \Rightarrow \gamma^{T_i} \geq \gamma^{T_j}$. This is different from the increasingness property, which involves two input functions and one criterion, i.e., $f \leq g \Rightarrow \gamma^T(f) \leq \gamma^T(g)$ (a condition that is not fulfilled for thickening and thinning transformations).

Analogously to an MP, an AP can be defined as a concatenation of a thickening AP $\prod_{\phi^{T'}}$ and a thinning AP $\prod_{\gamma^{T'}}$ (3) shown at the top of the page, where λ is considered the reference and defines the degree of filtering. Being $T' = \{T_1, T_2, \ldots, T_n\}$ the set of ordered criteria, for $T_i, T_j \in T'$, and $j \geq i$, the relation $T_i \subseteq T_j$ holds.

According to the attribute considered, different information can be extracted from the image. For example, if an increasing attribute is considered (e.g., the area of the regions), the AP performs analysis based on the scale of the structures in the scene. Instead, if a measure of the homogeneity of the gray-level values of the pixels (which is, usually, nonincreasing) belonging to each region is considered an attribute, it is possible to gather information on the texture.

C. EAP and EMAP

Analogously to the definition of EMPs, we can compute the APs on the c PCs extracted from the original hyperspectral data. This leads to the definition of the EAP as follows:

$$\mathrm{EAP} = \{\mathrm{AP}(\mathrm{PC}_1), \mathrm{AP}(\mathrm{PC}_2), \ldots, \mathrm{AP}(\mathrm{PC}_c)\}. \tag{4}$$

Moreover, since APs are created by different attributes to extract different information from the scene, the idea of the EAPs is further evolved to the EMAPs. The EMAP merges different EAPs in a single data structure. An EMAP composed of S different EAPs can be easily formulated as

$$\text{EMAP} = \left\{\text{EAP}_{a_1}, \text{EAP}'_{a_2}, \ldots, \text{EAP}'_{a_S}\right\} \tag{5}$$

where a_i is a generic attribute, and $\text{EAP}' = \text{EAP}\backslash\{\text{PC}_1, \ldots, \text{PC}_c\}$. The latter relation is necessary for avoiding the multiple presence of the c PCs since the original PCs are present in each EAP. The following attributes have been widely used in the literature to produce EMAP:

1) area of the region (related to the size of the regions);
2) diagonal of the box bounding the regions;
3) moment of inertia (as an index for measuring the elongation of the regions);
4) standard deviation (as an index for showing the homogeneity of the regions).

III. Proposed Method

Here, the variational equivalence between MKL and group lasso is presented at first, followed by the class-specific kernel learning strategy and the proposed algorithm. Then, the framework of our method is described.

A. Connection Between MKL and Group Lasso

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{n\times d}$ denote the collection of n training samples that are in a d-dimensional space. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in \{-1, +1\}^n$ denote the binary class labels for the data points in $\mathbf{X}$. MKL is often cast into the following optimization problem:

$$\min_{f\in H_{\boldsymbol{\eta}}} \left[\frac{1}{2}\|f\|^2_{H_{\boldsymbol{\eta}}} + C\sum_{i=1}^{n} \ell\left(y_i f(x_i)\right)\right] \tag{6}$$

where $H_{\boldsymbol{\eta}}$ is a reproducing kernel Hilbert space parameterized by $\boldsymbol{\eta}$, and $\ell(\cdot)$ is a loss function. $H_{\boldsymbol{\eta}}$ is endowed with kernel function $\kappa(\cdot,\cdot;\boldsymbol{\eta}) = \sum_{j=1}^{m} \eta_j \kappa_j(\cdot,\cdot)$.

When the Hinge loss is employed, the dual problem of MKL [46] is equivalent to

$$\min_{\boldsymbol{\eta}\in\Omega} \max_{\boldsymbol{\alpha}\in\Delta} \left[\mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})^T\left(\sum_{j=1}^{m}\eta_j\mathbf{K}_j\right)(\boldsymbol{\alpha}\circ\mathbf{y})\right] \tag{7}$$

where Ω is the domain of $\boldsymbol{\eta}$, and Δ is the domain of $\boldsymbol{\alpha}$. $\mathbf{1}$ is a vector of all ones, $\{\mathbf{K}_j\}_{j=1}^m$ is a group of base kernel matrices associated with H'_j, and $\circ$ defines the elementwise product between two vectors. The domain Δ is usually defined as $\Delta = \{\boldsymbol{\alpha} \in \mathbb{R}^n : \boldsymbol{\alpha}^T\mathbf{y} = 0, \mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{C}\}$, where $\boldsymbol{\eta} \in \Omega$ lies in a simplex, i.e., $\Omega = \{\boldsymbol{\eta} \in \mathbb{R}^m_+ : \sum_{j=1}^m \eta_j = 1, \eta_j \ge 0\}$, which we call the L_1-norm of kernel weights. Most MKL methods fall in this category.

The group lasso estimate is defined as the solution to

$$\frac{1}{2}\left\|\mathbf{y} - \sum_{j=1}^{m}\mathbf{K}_j\eta_j\right\|^2 + \Lambda\sum_{j=1}^{m}\|\eta_j\|_1 \tag{8}$$

where $\Lambda \ge 0$ is a tuning parameter. Bakin proposed (8) as an extension of the lasso for selecting groups of variables [40]. To show the connection between MKL and group lasso, we first transform (6), of which the optimization problem is equivalent to the following optimization problem

$$\min_{\boldsymbol{\eta}\in\Omega} \min_{\{f_j\in H_j\}_{j=1}^m} \left[\frac{1}{2}\sum_{j=1}^{m}\eta_j\|f_j\|^2_{H_j} + C\sum_{i=1}^{n}\ell\left(\sum_{j=1}^{m} y_i\eta_j f_j(x_i)\right)\right]. \tag{9}$$

It is important to note that problem in (9) is nonconvex; therefore, we cannot deploy the standard approach to convert the problem in (9) into its dual form. To transform (9) into (6), first, we rewrite $C\ell(z) = \max_{\boldsymbol{\alpha}\in[\mathbf{0},\mathbf{C}]} \boldsymbol{\alpha}(1-z)$. Then, the optimization problem in (9) can be transformed as follows:

$$\min_{\boldsymbol{\eta}\in\Omega} \min_{\{f_j\in H_j\}_{j=1}^m} \left[\frac{1}{2}\sum_{j=1}^{m}\eta_j\|f_j\|^2_{H_j} + \max_{\boldsymbol{\alpha}\in[\mathbf{0},\mathbf{C}]^n}\sum_{i=1}^{n}\alpha_i\left(1 - \sum_{j=1}^{m} y_i\eta_j f_j(x_i)\right)\right]. \tag{10}$$

Since the problem is convex in f_j and concave in $\boldsymbol{\alpha}$, we can then switch the minimization of f_j with the maximization of $\boldsymbol{\alpha}$

$$\min_{\boldsymbol{\eta}\in\Omega} \max_{\boldsymbol{\alpha}\in[\mathbf{0},\mathbf{C}]^n} \left[\sum_{i=1}^{n}\alpha_i + \min_{\{f_j\in H_j\}_{j=1}^m}\sum_{j=1}^{m}\eta_j\left(\frac{1}{2}\|f_j\|^2_{H_j} - \sum_{i=1}^{n}\alpha_i y_i f_j(x_i)\right)\right]. \tag{11}$$

By taking the minimization of f_j, we have

$$f_j(x) = \sum_{i=1}^{n}\alpha_i y_i \kappa_j(\mathbf{x}_i, \mathbf{x}). \tag{12}$$

Then, the resulting optimization problem becomes

$$\min_{\boldsymbol{\eta}\in\Omega} \max_{\boldsymbol{\alpha}\in[\mathbf{0},\mathbf{C}]^n} \left[\mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\sum_{j=1}^{m}\eta_j(\boldsymbol{\alpha}\circ\mathbf{y})^T\mathbf{K}_j(\boldsymbol{\alpha}\circ\mathbf{y})\right] \tag{13}$$

which is exactly the same dual problem of (6) as shown in (7).

To further show the connection between group lasso and MKL, we need to decouple the interaction between weights $\boldsymbol{\eta}$ and classification functions f_j, $j = 1, \ldots, m$. Based on the result of Theorem 1, we define

$$\tilde{f}_j = \eta_j f_j \tag{14}$$

We then rewrite the problem in (9) as

$$\min_{\boldsymbol{\eta}\in\Omega} \min_{\{\tilde{f}_j\in H_j\}_{j=1}^m} \left[\frac{1}{2}\sum_{j=1}^{m}\frac{1}{\eta_j}\|\tilde{f}_j\|^2_{H_j} + C\sum_{i=1}^{n}\ell\left(y_i\sum_{j=1}^{m}\tilde{f}_j(x_i)\right)\right]. \tag{15}$$

By taking the minimization over $\boldsymbol{\eta}$, we obtain

$$\eta_j = \frac{\|\tilde{f}_j\|_{H_j}}{\sum_{j=1}^{m}\|\tilde{f}_j\|_{H_j}}, \quad j = 1, \ldots, m. \tag{16}$$

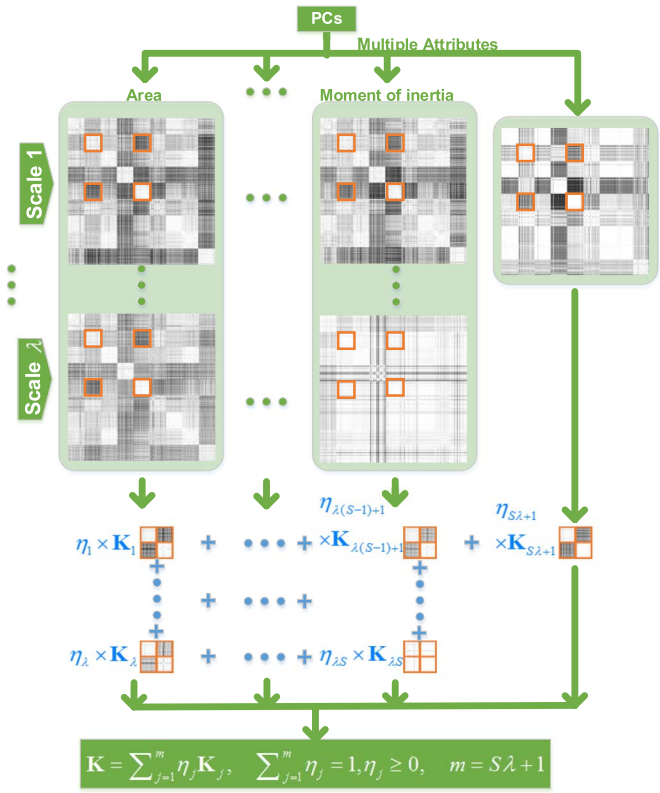


Fig. 1. Class-specific kernel learning (taking classes 2 and 5 as examples).

Based on (12) and (14), $\|\tilde{f}_j\|_{H_j}$ can be calculated as

$$\|\tilde{f}_j\|_{H_j}^2 = \eta_j^2 (\boldsymbol{\alpha} \circ \mathbf{y})^T \mathbf{K}_j (\boldsymbol{\alpha} \circ \mathbf{y}). \tag{17}$$

To rewrite (9) in another form that is clearly related to group lasso, (16) is substituted into the optimization problem (15), and the following optimization problem can be easily obtained:

$$\min_{\{f_j \in H_j\}_{j=1}^m} \left[\frac{1}{2} \left(\sum_{j=1}^{m} \|f_j\|_{H_j} \right)^2 + C \sum_{i=1}^{n} \ell \left(y_i \sum_{j=1}^{m} f_j(x_i) \right) \right] \tag{18}$$

where $f_j \in H_j$ for $j = 1, 2, \ldots, m$. The regularizer in the above can be regarded as a group-lasso regularizer for f [47]. It states that the classical formulation of MKL essentially uses a group-lasso-type regularizer, which is not obvious from the optimization problem in (6). For more detailed information, please check [44].

B. Class-Specific Kernel Learning

The proposed class-specific way via the one-versus-one learning strategy can search the most effective features for the classification of any two classes, leading to the better accuracy than those simple feature fusion methods. In other words, our method is able to select the class-specific weights for different attributes of MPs and remove the redundancy of those features, thus further improving the classification accuracy using the same input features. As shown in Fig. 1, when classifying one class pair (take classes 2 and 5 as examples), first, we find their position coordinates according to the label of training samples, and then the associate kernel $\mathbf{K}$ is extracted from the base kernels via the corresponding location. After that, CS-SMKL is used to optimal the class-specific weights. We summarize the proposed method discussed in the above into Algorithm 1, and NumClass denotes the number of categories.

Algorithm 1 The CS-SMKL

1: for $p = 1$: (NumClass $- 1$)
2: for $q = (p + 1)$: NumClass
3: Extract class-specific $\mathbf{K}$ from base kernels.
4: Initialize $\boldsymbol{\eta}^0 = 1/m$
5: **repeat**:
6: Solve the dual problem of SVM with $\mathbf{K} = \sum_{j=1}^{m} \eta_j \mathbf{K}_j$ to obtain the optimal solution $\boldsymbol{\alpha}$
7: Calculate $\|f_j\|_{H_j}$ and η_j according to (17) and (16), respectively
8: **until**: Convergence
9: end
10: end

C. Framework of the Proposed Method

The flowchart of the proposed framework is illustrated in Fig. 2. First, PCA was performed on the original data set, and multiple AFs with different degrees were carried out on the extracted PCs to obtain the EMAP features. Then, base kernels associated with EMAPs were constructed, and the final kernel was acquired through the CS-SMKL algorithm. Finally, the optimal kernel was embedded into an SVM to complete the classification of hyperspectral images. Fig. 1 shows the construction of the kernel used in the SVM. Obviously, there are S kinds of attributes, and under each attribute, λ scales of AFs are adopted; therefore, we get the EMAP which contains one PC and $S\lambda$ EAPs as (5). Each EAP is associated with one base kernel, and the PCs are also relevant to one base kernel. Therefore, when combined, we have $S\lambda + 1$ base kernels. Then, we extract the class-specific kernels from each base kernel, and we have $\boldsymbol{\eta} = [\eta_1, \eta_2, \ldots, \eta_m]$, $m = S\lambda + 1$, which is associated with the extracted class-specific kernels as shown in Fig. 1. Then, the optimal kernel is obtained by the linear combination of these extracted kernels, which is constrained by the criteria $\sum_{j=1}^{m} \eta_j = 1, \eta_j \geq 0$. The criteria can enforce the sparsity at the group/feature level and automatically learn a compact feature set for classification purposes. More precisely, our CS-SMKL determines the optimal weights of base kernels and results in improved classification performances.

IV. Experimental Results and Discussions

Here, we report the results of the experiments to evaluate the effectiveness of the proposed method, compared with the existing kernel learning methods.

A. Data Description

In our experiments, three high-spatial-resolution hyperspectral image data sets were used.

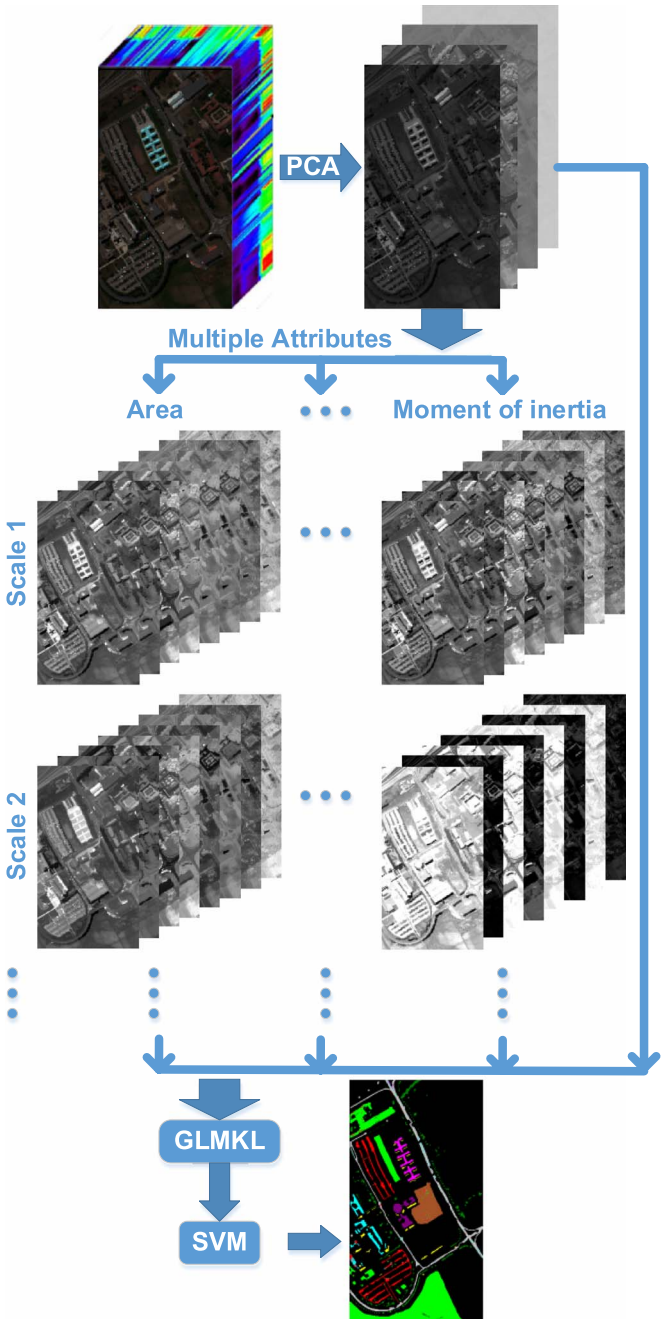


Fig. 2. General architecture of the proposed classification system.

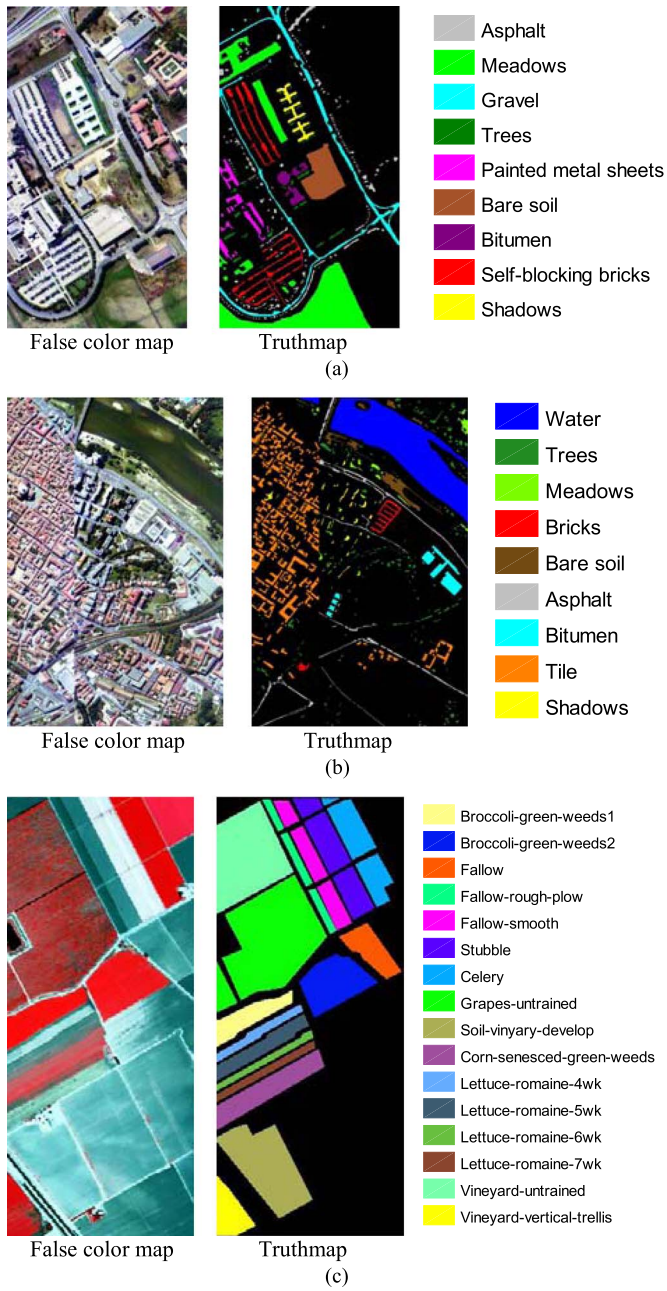


Fig. 3. Ground reference maps for the three data sets. (a) Pavia University. (b) Pavia Center. (c) Salinas.

Two images of an urban area were acquired with the Reflective Optics System Imaging Spectrometer (ROSIS-03) optical sensor [48]. The flight over the city of Pavia, Italy, was operated by the Deutschen Zentrum für Luft- und Raumfahrt (DLR, German Aerospace Agency) within the context of the HySens project, managed and sponsored by the European Union. According to specifications, the ROSIS-03 sensor provides 115 bands with a spectral coverage ranging from 0.43 to 0.86 μm. The spatial resolution is 1.3 m per pixel. The two data sets are as follows.

1) *University Area*: This hyperspectral image with 610 $\times$ 340 pixels took place near the Engineering School, University of Pavia, Pavia, Italy. Twelve channels were removed due to noise [38]. The remaining 103 spectral channels were processed. There are 43 923 labeled samples in total and nine classes of interest. Fig. 3(a) presents false color images of this data set.
2) *Center Area*: The second test set was the center of Pavia. The image was originally 1096 $\times$ 1096 pixels. A 381-pixel-wide black band in the left-hand side part of image was removed, resulting in a "two part" image of 1096 $\times$ 715 pixels. Thirteen channels have been removed due to noise. The remaining 102 spectral channels were processed. In total, there are 148 152 labeled samples and nine classes of interest. Fig. 3(b) presents false color images of this data set.

TABLE I
INFORMATION FOR THE THREE CONSIDERED DATA SETS

No.	Pavia University		Pavia Center		Salinas	
	Categories	Samples	Categories	Samples	Categories	Samples
C1	Asphalt	6852	Water	65971	Broccoli-green-weeds_1	2009
C2	Meadows	18686	Tree	7598	Broccoli-green-weeds_2	3726
C3	Gravel	2207	Meadow	3090	Fallow	1976
C4	Trees	3436	Brick	2685	Fallow-rough-plow	1394
C5	Painted metal sheets	1378	Soil	6584	Fallow-smooth	2678
C6	Bare soil	5104	Asphalt	9248	Stubble	3959
C7	Bitumen	1356	Bitumen	7287	Celery	3579
C8	Self-blocking bricks	3878	Tile	42826	Grapes-untrained	11271
C9	Shadows	1026	Shadows	2863	Soil-vinyary-develop	6203
C10					Corn-senesced-green-weeds	3278
C11					Lettuce-romaine-4wk	1068
C12					Lettuce-romaine-5wk	1927
C13					Lettuce-romaine-6wk	916
C14					Lettuce-romaine-7wk	1070
C15					Vineyard-untrained	7268
C16					Vineyard-vertical-trellis	1807
	Total	43923	Total	148152	Total	54129

The third hyperspectral image was acquired by AVIRIS sensor over the Valley of Salinas in Southern California [49]. The spatial resolution is 3.7 m per pixel.

Salinas: It has spatial dimensions of 512 $\times$ 217 pixels. Twenty water absorption bands were removed, and a 200-band image was used for the experiments. The ground reference map was composed of 54 129 pixels and 16 land-cover classes. Fig. 3(c) shows an RGB composition and information of its labeled classes.

More details about these three data sets are listed in Table I.

B. Experimental Setting

In our experiments, first, the data sets were processed by PCA: The eigenvalues were arranged in the descending order and reserved first c PCs which, divided by the eigenvalues of all original PCs, equaled 99%. After that, multiple AFs were carried out on the extracted c PCs. We used all the four kinds of AFs presented in Section II-C. For the Pavia data sets, we apply the same setting in [50]: 1) a, area of the regions, $\lambda_a = [100\,500\ 1000\ 5000]$; 2) d, length of the diagonal of the box bounding the region, $\lambda_d = [10\ 25\ 50\ 100]$; 3) i, first moment invariant of Hu, moment of inertia [51], $\lambda_i = [0.2\ 0.3\ 0.4\ 0.5]$; and 4) s, standard deviation of the gray-level values of the pixels in the regions, $\lambda_s = [20\ 30\ 40\ 50]$. For the Salinas data set, we adopt the following setting: 1) $\lambda_a = [100\ 500\ 1000\ 5000]$; 2) $\lambda_d = [10\ \ 25\ \ 50\ \ 100]$; 3) $\lambda_i = [0.1\ \ 0.15\ \ 0.2\ \ 0.25]$; and 4) $\lambda_s = [15\ 20\ 25\ 30]$.

Three parameters are introduced in the proposed algorithm. They are the scale σ of base kernels, the threshold ε which controls the sparsity of the kernels, and the tuning parameter C. The kernel scale is recommended to be set as 2 as there is no need to pay great attention to the scale parameter adjustment and it can be implicitly transformed into the process of optimizing kernel weights in the MKL framework. The threshold ε can usually be set smaller than $1/m$ (m denotes the number of base kernels used on average). It was set to 0.001 in the tests conducted in this paper. The tuning parameter C was experimentally set to 1000.

Our proposed CS-SMKL is compared with eight benchmark methods including standard SVM using original spectral features, standard SVM using stacked EMAP features, sparse-representation-based classification (SRC) [52], collaborative representation-based classification (CRC) [49], RBMKL, RMKL, spatial–spectral combined kernel learning (CKL), and multiple feature learning (MFL) [53].

Spe-SVM It has a standard SVM using all the original spectral features and a single Gaussian kernel function with a fixed-scale parameter $\sigma = 2$.

SRC All the original spectral features are used, and the regularization parameter is set as the optimal parameter.

CRC All the original spectral features are used, and the regularization parameter is set as the optimal parameter.

EMAP-SVM All the EMAP features are stacked as a single vector of features, and it has a standard SVM using the stacked features and a single Gaussian kernel function with a fixed-scale parameter $\sigma = 2$.

CS-SMKL All the EMAP features are used in this method, and a Gaussian kernel function with a fixed-scale parameter $\sigma = 2$ is used to build 17 base kernels corresponding to the 16 EAPs and PCs, respectively. The base kernels carry the corresponding scale information because the EMAP features are obtained by different degrees of AFs.

RBMKL The setting is the same as CS-SMKL, and the mean rule is adopted.

RMKL The setting is the same as CS-SMKL.

CKL All the EMAP features are used. The base kernels consist of two parts, 16 Gaussian kernels corresponding to the 16 EAPs, and 20 Gaussian kernels representing the differences between extracted spectral features (PCs) at different scales ranging from 0.1 to 2, whereas the step size of increment is 0.1, $\mu = 0.4$ is used to weight the spectral kernel and spatial kernel, i.e., $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \mu\mathbf{K}(\mathbf{x}_i^{\mathrm{Spe}}, \mathbf{x}_j^{\mathrm{spe}}) + (1-\mu)\mathbf{K}(\mathbf{x}_i^{\mathrm{Spa}}, \mathbf{x}_j^{\mathrm{spa}})$.

MFL All the EMAP features are used in this method.

TABLE II
OA (%) OF THE THREE DATA SETS

Data Sets	Pavia University			Pavia Center			Salinas		
Classifier	Number of training samples			Number of training samples			Number of training samples		
	10	20	30	10	20	30	10	20	30
Spe-SVM	70.02±2.75	74.69±1.94	76.42±1.53	93.48±0.76	94.83±0.44	95.41±0.54	84.04±2.39	87.43±1.06	88.58±0.50
SRC	69.79±2.87	73.37±2.57	76.33±0.91	94.06±0.55	95.30±0.30	95.85±0.50	84.95±1.33	86.15±0.73	86.19±1.81
CRC	61.98±4.15	64.00±1.94	65.12±2.64	93.06±0.95	93.79±0.39	93.99±0.53	83.76±2.21	85.62±0.75	86.33±2.06
EMAP-SVM	79.82±7.83	88.20±3.81	89.74±3.62	94.83±0.85	95.77±1.16	96.75±1.16	85.56±3.14	88.78±0.82	90.92±0.66
CS-SMKL	**91.36**±2.54	**94.35**±1.22	**95.65**±1.41	**96.01**±0.32	**97.02**±0.45	**97.85**±0.24	**89.23**±2.28	**91.60**±1.01	**93.18**±0.57
RBMKL	87.62±2.51	92.95±1.45	94.71±1.25	***95.59***±0.58	***96.46***±0.66	***97.49***±0.43	87.35±2.30	***90.77***±1.34	***91.82***±0.42
RMKL	85.03±3.09	91.86±1.53	93.86±1.38	94.28±0.63	95.13±1.06	96.59±0.86	86.83±2.16	90.21±1.19	91.28±0.52
CKL	78.85±3.23	86.73±1.83	91.52±1.47	95.14±0.72	96.34±0.46	97.10±0.42	86.19±1.36	88.07±1.05	89.98±0.67
MFL	***89.00***±2.83	***93.06***±0.99	***94.80***±1.14	90.28±3.55	95.86±0.53	96.67±0.59	***87.94*** ±2.70	90.21±1.26	91.53 ±1.30

C. Evaluation of Classification Performances

Conclusions were drawn from analyzing the classification accuracy assessed with the following measures:

a) overall accuracy (OA), which is the number of correctly classified samples out of the total number of tested samples in percentage;
b) a kappa coefficient of agreement, which is the percentage of agreement corrected by the amount of agreement that could be expected due to chance alone [54].

These criteria were used to compare classification results and were computed by using the confusion matrix. Furthermore, the statistical significance of differences was computed by using McNemar's test, which is based upon the standardized normal test statistics [55], i.e.,

$$Z = \frac{Y_{12} - Y_{21}}{\sqrt{Y_{12} + Y_{21}}} \tag{19}$$

where Y_{12} indicates the number of samples classified correctly by classifier 1 and incorrectly by classifier 2. The difference in accuracy between classifiers 1 and 2 is said to be statistically significant if $|Z| > 1.96$. The sign of Z indicates whether classifier 1 is more accurate than classifier 2 $(Z > 0)$ or vice versa $(Z < 0)$. This test assumes related testing samples and thus is adapted to our situation since the training and testing sets were the same for each experiment.

The number of training samples for all the data sets has been set to 10, 20, and 30 per class. All the results reported are the mean and standard deviation of ten experiments with randomly selected training and test sets.

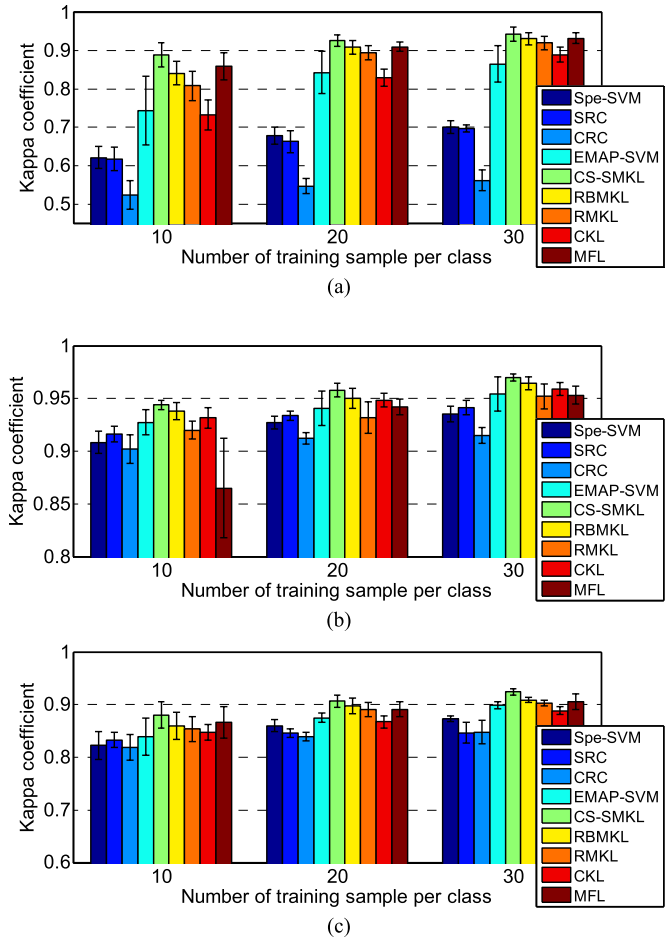
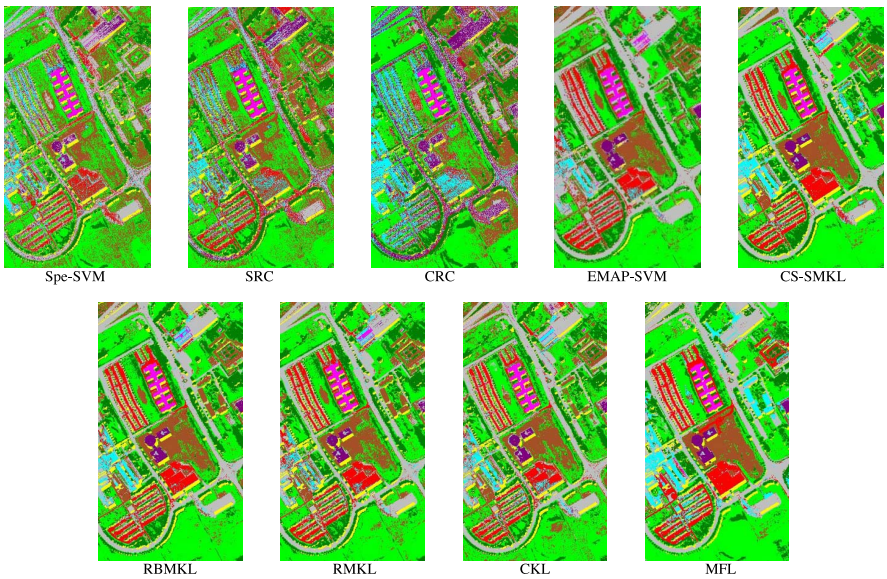


Fig. 4. Kappa coefficient of three data sets. (a) Pavia University. (b) Pavia Center. (c) Salinas.

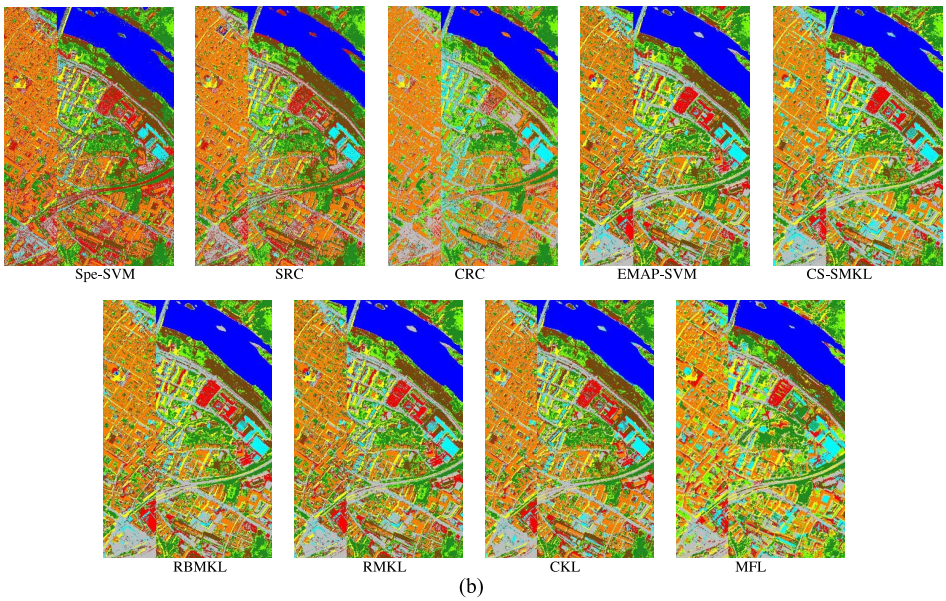
D. Experimental Results and Analysis

a) Analysis of the overall classification performances: The CS-SMKL method, whose classification results are given in bold letters, performs best among all the methods tested in terms of classification accuracy for all the data sets, as shown in Table II. These results demonstrate that, by properly weighting the importance of the features, we can construct efficient and compact kernel machines encoding the relationships in the observed data. The box plot of the kappa coefficients versus the number of the training samples per class is shown in Fig. 4. The highest kappa coefficients of CS-SMKL illustrate that CS-SMKL is the most accurate classifier among these methods, which is in line with the conclusions drawn from Table II. For all these three data sets, the standard deviation becomes smaller as the number of training samples increases, which means that the larger the training set is, the more stable CS-SMKL is.

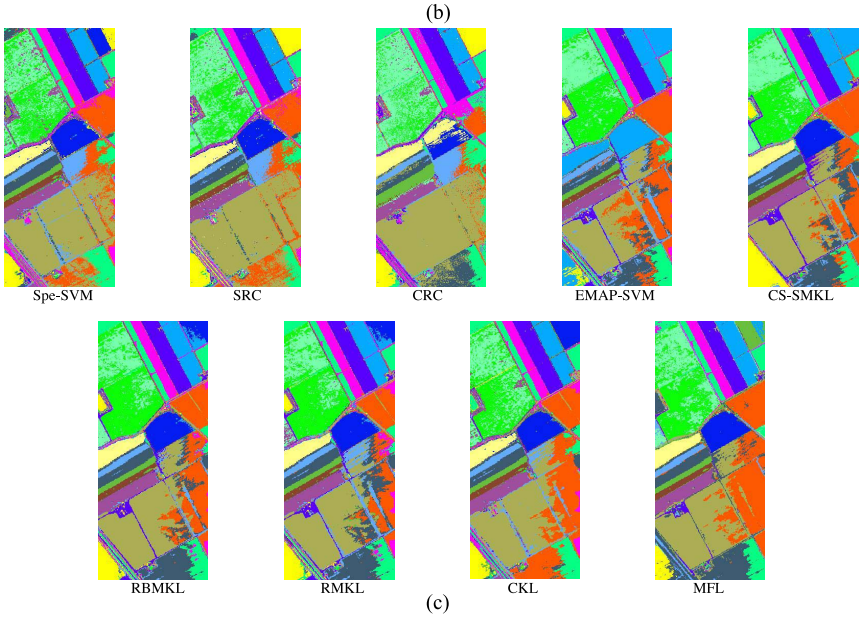
The classification maps and confusion matrices for different approaches are given in Figs. 5 and 6, respectively. For the Pavia University data set, we can see that the improvement acquired by CS-SMKL is between 0.94% and 3.74%. It is also shown in Figs. 5(a) and 6(a) that, for the classes which are easy to be incorrectly classified such as Gravel, Bare soil, and



(a)



(b)



(c)

Fig. 5. Classification maps of three data sets for different approaches (Size of training samples: ten per class). (a) Pavia University. (b) Pavia Center. (c) Salinas.

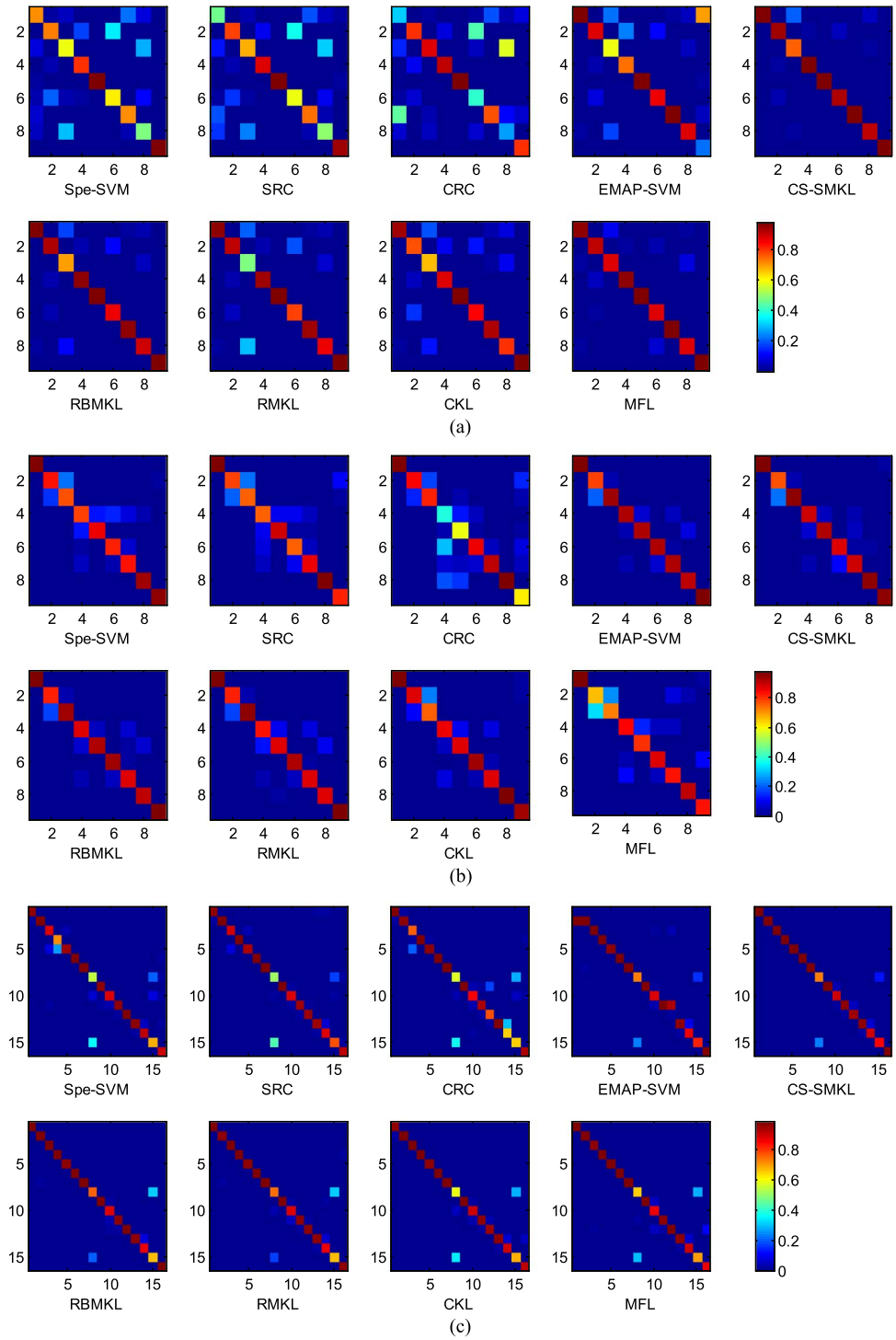


Fig. 6. Confusion matrices showing the classification performances with three data sets (Size of training samples: 30 per class). (a) Pavia University. (b) Pavia Center. (c) Salinas.

Self-blocking bricks, the proposed CS-SMKL outperforms the other classifiers. For the Pavia Center data set, the CS-SMKL obtain comparable classification results with an increment of 0.56%. Fig. 5(b) shows that the wrongly classified pixels in the water area are much less, and it is obvious that the classification results of Tree, Soil, and Bitumen becomes better as shown in Fig. 6(b). Moreover, for the Salinas data set, the improvement is up to 1.88%. Grapes-untrained and Vineyard-untrained are two categories that are easy to be confused, as shown in Figs. 5(c) and 6(c); however, the proposed method offers a great improvement in classification accuracy of these two classes.

The Z-tests between CS-SMKL and other classifiers are shown in Table III. Although the classification accuracy of CS-SMKL is slightly larger than other methods for the Pavia Center data set, all the values of Z-test are much greater than 1.96, which means the proposed CS-SMKL is better.

TABLE III
Z-TEST OF THE THREE DATA SETS

Classifier	Pavia University			Pavia Center			Salinas		
	Number of training samples			Number of training samples			Number of training samples		
	10	20	30	10	20	30	10	20	30
Spe-SVM	85.76	85.35	87.03	40.91	37.12	44.03	34.63	29.08	30.49
SRC	85.69	87.60	86.06	30.95	29.63	36.18	29.31	35.12	44.00
CRC	103.07	109.09	111.38	42.47	48.26	58.88	31.72	36.44	40.70
EMAP-SVM	57.00	39.11	43.59	27.06	28.78	27.16	27.52	23.46	19.67
RBMKL	28.43	15.43	11.84	11.29	16.38	14.84	19.79	9.75	16.75
RMKL	41.44	24.26	18.97	38.09	41.62	35.60	23.63	14.72	21.54
CKL	65.37	50.18	39.13	18.48	16.59	20.17	26.88	28.71	29.82
MFL	16.82	12.03	11.80	74.31	25.43	26.58	10.90	12.59	14.51

TABLE IV
COMPUTATIONAL COST OF THE THREE DATA SETS (S)

Data set	Pavia University			Pavia Center			Salinas		
Classifier	Number of training samples			Number of training samples			Number of training samples		
	10	20	30	10	20	30	10	20	30
Spe-SVM	0.27	0.49	0.78	0.88	1.53	2.23	1.13	1.87	2.87
SRC	17.00	28.51	47.53	47.56	85.70	146.31	27.79	98.06	205.63
CRC	15.17	30.84	53.01	117.18	154.34	240.48	46.46	118.37	259.94
EMAP-SVM	0.28	0.48	0.71	0.91	1.51	2.33	1.07	1.88	2.84
CS-SMKL	5.17	9.12	13.53	15.56	26.97	41.47	18.97	42.69	49.61
RBMKL	0.49	0.85	1.80	1.49	3.87	5.82	1.26	3.02	4.49
RMKL	0.85	1.63	3.23	1.88	3.73	6.01	1.76	3.44	5.65
CKL	0.58	1.09	1.64	1.97	3.47	5.94	1.77	3.17	4.79
MFL	0.20	0.21	0.22	0.43	0.45	0.47	0.24	0.32	0.39

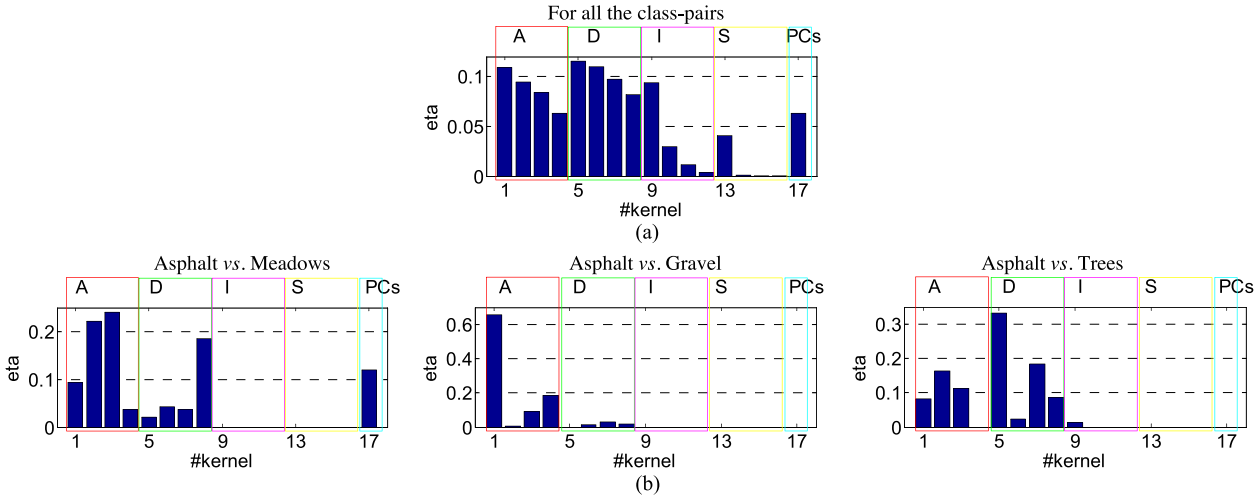


Fig. 7. Weights $\boldsymbol{\eta}$ determined for each base kernel and the corresponding feature type. (a) Fixed set of kernel weights selected by RMKL. (b) Kernel weights selected for three different class pairs by CS-SMKL.

The complexity of the proposed method is measured by the computation time of the training and test phases. The complexity of the proposed method is measured by the computation time of the training and test phases. Note that the experiments were performed on a computer with Intel Core i5-4460 CPU 3.2 GHz and 16-GB RAM. Table IV shows the time cost of all the methods. According to Table IV, we can see that the proposed method is faster than SRC and CRC but is slower than the others. The reason is that the proposed method selects class-specific feature sets, which means that the kernel weights need to be optimized for each class pair, whereas other MKL methods use a fixed set of kernel weights for all the class pairs. For the RMKL method, the computational complexity for solving kernel weights is $\mathrm{O}((m \times n \times T)^2)$, where m denotes the number of base kernels, n denotes the number of training samples for each class, and T denotes the number of categories. For RBMKL and CKL, the computational complexity is $O(m)$. According to (16), the computational complexity of CS-SMKL is $\mathrm{O}(m \times I \times (2n \times T)^2$, where I denotes the iterations when achieving convergence. The higher computational cost, however, provides the best classification accuracy.

b) Analysis of selected features and classification ability using class-specific kernel learning: To visualize the contribution of each feature type in these MKL methods, we plot the kernel weights of the base kernels for RMKL and CS-SMKL in Fig. 7(a) and (b), respectively. We only list three one-against-one classifiers of the Pavia University data set (Asphalt vs Meadows, Asphalt vs Gravel, and Asphalt vs Trees) for

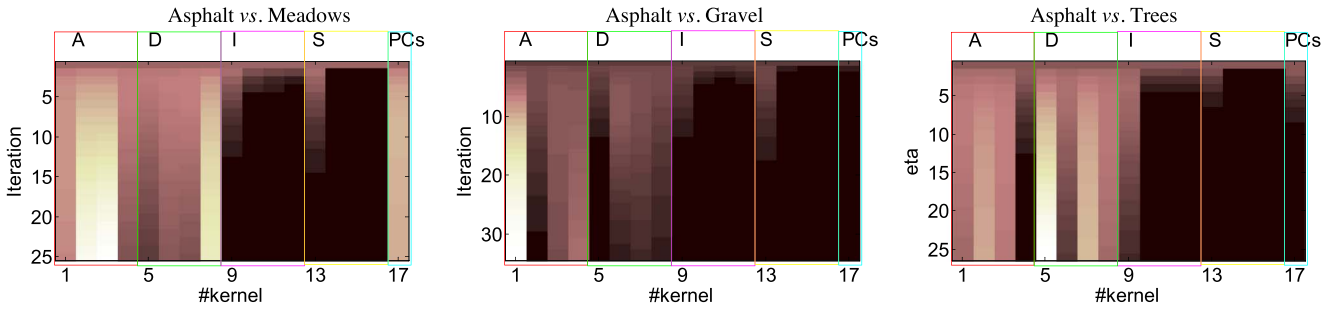


Fig. 8. Optimization of the weights $\boldsymbol{\eta}$.

TABLE V
OA (%) OF THE THREE DATA SETS

Data Sets	Pavia University			Pavia Center			Salinas		
Classifier	Number of training samples			Number of training samples			Number of training samples		
	10	20	30	10	20	30	10	20	30
M-CS-SMKL	***89.95***±4.50	***93.72***±2.42	***94.94***±2.25	95.34±2.35	96.27±0.58	***97.55***±0.38	***88.46***±2.47	90.15±1.25	***92.45***±1.25
CS-SMKL	**91.36**±2.54	**94.35**±1.22	**95.65**±1.41	**96.01**±0.32	**97.02**±0.45	**97.85**±0.24	**89.23**±2.28	**91.60**±1.01	**93.18**±0.57
RBMKL	87.62±2.51	92.95±1.45	94.71±1.25	***95.59***±0.58	***96.46***±0.66	97.49±0.43	87.35±2.30	***90.77***±1.34	91.82±0.42

simplicity. Note that we did not plot the kernel weights for RBMKL and CKL methods because the mean rule is adopted, and the weights of base kernels are equal. RMKL used the same kernel weights as shown in Fig. 7(a) for all the class pairs. In Fig. 7(b), we see that our CS-SMKL selected different sparse base kernel sets for different class pairs. Therefore, the fixed set of kernels selected using RMKL [see Fig. 7(a)] is unable to accommodate the need for individual cases. Taking the categories of Asphalt vs Meadows as an example, the spectral features (PCs) are important for this class pair. For the CS-SMKL, it only selected very few base kernels for classification purposes, whereas the kernel weight for the spectral features is very high. However, these corresponding kernel weights in RMKL are much lower. For class pairs Asphalt vs Gravel and Asphalt vs Trees, CS-SMKL did not select the base kernel corresponding to PCs, whereas RMKL still select the 17th base kernel. In Fig. 7(b), it is also clearly shown that, for all these three class pairs, CS-SMKL did not select the base kernels associate with attribute S (standard deviation), but RMKL selected all of them, which means that when classifying different class pairs, the EAPs which are obtained by different attributes and degrees of AFs produce different effectiveness in the classification process. This is an example that shows that our CS-SMKL provides more flexibility in selecting kernels (features) for improved classification.

To visualize the chosen features, Fig. 8 shows the iterative optimization of the weights using CS-SMKL. Starting by a uniform configuration of weights ($\eta_j = 1/m = 0.0588$, $j = 1, \ldots, m$), for the first class pair, the 17th kernel corresponding to PCs is given a strong weight after several iterations. The area attribute and diagonal attribute are also selected in the following steps. This way of selecting features matches the classification results; although the spectral features serve to discriminate man-made Asphalt from natural classes Meadows, the proper assignment to the area and diagonal attribute which measure the shape of the regions largely improves results. For the second classifier, the spectral features have not been selected because the original spectrum of these two classes is very similar and provides less information to discriminate between each of the two classes. The area and diagonal attributes almost all have been retained. Since the spectral information provides a limited usefulness, the spatial information becomes very important for the classification task. The area and the length of the diagonal of the bounding box are increasing attributes that are useful to perform a multiscale analysis of the data. Then, it is easier to tell Asphalt from Gravel. For the class pair Asphalt vs Trees, the area attribute and diagonal attribute are selected; in particular, the fifth kernel is given a strong weight. The ninth kernel associated with the moment-of-inertia attribute is also selected as a purely geometric descriptor and can be employed to extract information on the geometry of the regions, regardless of their scale. We can see from the reference map that the shape of these two classes are very different and using spatial features only is enough two separate them apart.

Finally, we embed the mean rule into CS-SMKL, expressed as M-CS-SMKL, which means that we first used the mean rule to deal with the base kernels under the same attribute. In our method, we got four mean kernels; superadding the base kernel related to PCs, we inputted five kernels to the CS-SMKL. The classification results are shown in Table V. All the results in the same column are acquired by using the same training samples. We can see from Table V that M-CS-SMKL obtained higher classification accuracy than RBMKL on Pavia University and Salinas data sets; the improvements can be 2.33% and 1.11%, respectively. While on the Pavia Center data set, M-CS-SMKL obtained comparable results. The results mean that the selected features are very effective and have very low redundancy, i.e., CS-SMKL selects a compact set of features. Meanwhile, the accuracy of M-CS-SMKL is lower than the proposed CS-SMKL, illustrating that CS-SMKL provides an efficient way to combine these significant features and select the class-specific weights.

V. Conclusion

In this paper, a new classification method is proposed for the analysis of hyperspectral images. Since MKL provides a very effective means of learning, a CS-SMKL classifier combined with the use of EMAP is developed. The proposed framework allows embedding a variety of characteristics in the classifier. Expanding the feature spaces with a number of information diversities (different attributes, different scales, different components), Extended Morphological Attribute Profiles provide excellent classification performances, but with a high redundancy of information. CS-SMKL can remove the redundancy of EMAP effectively to learn a compact set of features and select the class-specific weights, leading to remarkable discriminability. The experimental results on three different hyperspectral data sets, corresponding to different contexts (urban, semi-urban, and agricultural) and different spectral and spatial resolutions, demonstrate that the proposed method CS-SMKL offers excellent performances and accurate classification maps, and the efficiency of CS-SMKL is further demonstrated. The combined use of EMAP and CS-SMKL can handle difficult-to-separate classes with irregular class boundaries. In the future, we will further focus on integration of sparse-MKL-based feature learning and sparse-representation-based classifier for spatial–spectral hyperspectral image classification and feature interpretation.

Acknowledgment

The authors would like to thank Prof. P. Gamba for providing the Pavia data sets. The authors also would like to thank Prof. W. Li and Prof. H. Zhang for providing the corresponding codes of SRC and CRC.

References

[1] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral–spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[2] P. Ghamisi, M. D. Mura, and J. A. Benediktsson, "A survey on spectral–spatial classification techniques based on attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2335–2353, May 2015.

[3] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.

[4] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral–spatial classification of hyperspectral images based on hidden Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2565–2574, May 2014.

[5] P. Ghamisi, M. S. Couceiro, F. M. Martins, and J. A. Benediktsson, "Multilevel image segmentation approach for remote sensing images based on fractional-order Darwinian particle swarm optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2382–2394, May 2014.

[6] P. Ghamisi, M. Couceiro, M. Fauvel, and J. A. Benediktsson, "Integration of segmentation techniques for classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 342–346, Jan. 2014.

[7] J. Li, H. Zhang, and L. Zhang, "Supervised segmentation of very high resolution images by the use of extended morphological attribute profiles and a sparse transform," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1409–1413, Aug. 2014.

[8] Z. Zhu, S. Jia, S. He, Y. Sun, Z. Ji, and L. Shen, "Three-dimensional Gabor feature extraction for hyperspectral imagery classification using a memetic framework," *Inf. Sci.*, vol. 298, pp. 274–287, Mar. 2015.

[9] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson, "Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6663–6674, Dec. 2015.

[10] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral–spatial classification of hyperspectral images with a superpixel-based discriminative sparse model," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4186–4201, Aug. 2015.

[11] G. Zhang, X. Jia, and J. Hu, "Superpixel-based graphical model for remote sensing image mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 5861–5871, Nov. 2015.

[12] Z. Xue, J. Li, L. Cheng, and P. Du, "Spectral–spatial classification of hyperspectral data via morphological component analysis-based image separation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 70–84, Jan. 2015.

[13] X. Xu, J. Li, X. Huang, M. Dalla Mura, and A. Plaza, "Multiple morphological component analysis based decomposition for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 3083–3102, May 2016.

[14] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral–spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, Dec. 2014.

[15] F. Wei, S. Li, L. Fang, X. Kang, and J. A. Benediktsson, "Hyperspectral image classification via shape-adaptive joint sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 2, pp. 556–567, Feb. 2016.

[16] E. J. Breen and R. Jones, "Attribute openings, thinnings, and granulometries," *Comput. Vis. Image Understand.*, vol. 64, no. 3, pp. 377–389, Nov. 1996.

[17] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.

[18] M. Chini, N. Pierdicca, and W. Emery, "Exploiting SAR and VHR optical images to quantify damage caused by the 2003 Bam earthquake," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 1, pp. 145–152, Jan. 2009.

[19] P. Soille, *Morphological Image Analysis, Principles and Applications*, 2nd ed. New York, NY, USA: Springer-Verlag, 2003.

[20] M. K. D. Tuia, F. Pacifici, and W. Emery, "Classification of very high spatial resolution imagery using mathematical morphology and support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3866–3879, Nov. 2009.

[21] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

[22] R. Bellens, S. Gautama, L. Martinez-Fonte, W. Philips, J. Chan, and F. Canters, "Improved classification of VHR images of urban areas using directional morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 2803–2813, Oct. 2008.

[23] Y. Gu, T. Liu, X. Jia, J. A. Benediktsson, and J. Chanussot, "Nonlinear multiple kernel learning with multi-structure-element extended morphological profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3235–3247, Jun. 2016.

[24] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "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.

[25] N. Bouaynaya and D. Schonfeld, "Theoretical foundations of spatially variant mathematical morphology Part II: Gray-level images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 837–850, May 2008.

[26] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *Int. J. Remote Sens.*, vol. 31, no. 22, pp. 5975–5991, Nov. 2010.

[27] M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone, "Classification of hyperspectral images with extended attribute profiles and feature extraction techniques," in *Proc. IEEE IGARSS*, Jul. 2010, pp.76–79.

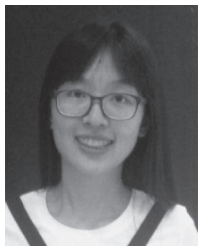
[28] B. Pan, J. Lai, and L. Shen, "Ideal regularization for learning kernels from labels," *Neural Netw.*, vol. 56, pp. 22–34, Aug. 2014.

[29] J. Shi, Q. Jiang, R. Mao, M. Lu, and T. Wang, "FR-KECA: Fuzzy robust kernel entropy component analysis," *Neurocomputing*, vol. 149, pp. 1415–1423, Feb. 2015.

[30] 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.

[31] J. A. Gualtieri and R. F. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Proc. SPIE AIPR Workshop, Adv. Comput.-Assisted Recognit.*, 1998, vol. 3584, pp. 221–232.

[32] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Evaluation of kernels for multiclass classification of hyperspectral remote sensing data," in *Proc. ICASSP*, May 2006, pp. II-813–II-816.
[33] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. Int. Conf. Mach. Learn.*, 2004, p. 6.
[34] Y. Gu and H. Liu, "Sample-screening MKL method via boosting strategy for hyperspectral image classification," *Neurocomputing*, vol. 173, pp. 1630–1639, Jan. 2016.
[35] Y. Gu, Q. Wang, H. Wang, D. You, and Y. Zhang, "Multiple kernel learning via low-rank nonnegative matrix factorization for classification of hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2739–2751, Jun. 2015.
[36] Y. Gu, Q. Wang, X. Jia, and J. A. Benediktsson, "A novel MKL model of integrating LiDAR data and MSI for urban area classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5312–5326, Oct. 2015.
[37] Q. Wang, Y. Gu, and D. Tuia, "Discriminative multiple kernel learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3912–3927, Jul. 2016, doi: 10.1109/TGRS.2016.2530807.
[38] 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.
[39] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. L. Rojo-Alvarez, and M. Martinez-Ramon, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1822–1835, Jun. 2008.
[40] D. Tuia, F. Ratle, A. Pozdnoukhov, and G. Camps-Valls, "Multisource composite kernels for urban-image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 88–92, Jan. 2010.
[41] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[42] Y. Gu, C. Wang, D. You, Y. Zhang, S. Wang, and Y. Zhang, "Representative multiple-kernel learning for classification of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 7, pp. 2852–2865, Jul. 2012.
[43] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, pp. 49–67, 2006.
[44] Z. Xu, R. Jin, H. Yang, I. King, and M. R. Lyu, "Simple and efficient multiple kernel learning by Group Lasso," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 1–8.
[45] M. A. Hossain, X. Jia, and J. A. Benediktsson, "One-class oriented feature selection and classification of heterogeneous remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 4, pp. 1606–1612, Apr. 2016.
[46] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, Aug. 2004.
[47] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *J. Mach. Learn. Res.*, vol. 9, pp. 1179–1225, Aug. 2008.
[48] P. Gamba, "A collection of data for urban area characterization," in *Proc. IGARSS*, Anchorage, AK, USA, 2004, pp. 69–72.
[49] S. Jia, Z. Zhu, L. Shen, and Q. Li, "A two-stage feature selection framework for hyperspectral image classification using few labeled samples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 3, pp. 1023–1035, Apr. 2014.
[50] M. Dalla Mura, A. Villa, J. A. 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.
[51] M. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.
[52] W. Li and Q. Du, "A survey on representation-based classification and detection in hyperspectral remote sensing imagery," *Pattern Recognit. Lett.*, to be published.
[53] J. Li *et al.*, "Multiple feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1592–1606, Mar. 2015.
[54] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis: An Introduction*. New York, NY, USA: Springer-Verlag, 1999.
[55] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, May 2004.



Tianzhu Liu (S'14) was born in Harbin, China, in 1990. She received the B.S. and M.S. degrees from the Harbin Institute of Technology, Harbin. She is currently working toward the Ph.D. degree in hyperspectral image classification and scene classification with the School of Electronics and Information Engineering, Harbin Institute of Technology.

Yanfeng Gu (M'06) received the Ph.D. degree in information and communication engineering from Harbin Institute of Technology (HIT), Harbin, China, in 2005.

Then, he joined as a Lecturer with the School of Electronics and Information Engineering, HIT. He was appointed as an Associate Professor with the same institute in 2006; meanwhile, he was enrolled in the first Outstanding Young Teacher Training Program of HIT. From 2011 to 2012, he was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA, USA. He is currently a Professor with the Department of Information Engineering, HIT. He is the author of more than 60 peer-reviewed papers and four book chapters, and is the inventor or coinventor of seven patents. His research interests include image processing in remote sensing, machine learning and pattern analysis, and multiscale geometric analysis.

Dr. Gu serves as an Associate Editor for *Neurocomputing* and as a Peer Reviewer for several international journals such as the IEEE TRANSACTION ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTION ON INSTRUMENT AND MEASUREMENT, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and *IET Electronics Letters*. He received funding from the Natural Science Foundation of China for the Excellent Young Scholars in 2015.

Xiuping Jia (M'93–SM'03) received the B.Eng. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 1982 and the Ph.D. degree in electrical engineering from The University of New South Wales, Sydney, Australia, in 1996.

Since 1988, she has been with the School of Information Technology and Electrical Engineering, The University of New South Wales, Australian Defence Force Academy, Canberra Campus, Australia, where she is currently a Senior Lecturer. She is also a Guest Professor with Harbin Engineering University, Harbin, China, and an Adjunct Researcher with China National Engineering Research Center for Information Technology in Agriculture. She is a coauthor of the remote sensing textbook titled *Remote Sensing Digital Image Analysis* (Springer-Verlag, 3rd ed., 1999; 4th ed. 2006). Her research interests include remote sensing, image processing, and spatial data analysis.

Dr. Jia serves as an Editor for the *Annals of GIS* for remote sensing topic and as a Subject Editor for the *Journal of Soils and Sediments*. She also serves as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Jón Atli Benediktsson (F'04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984 and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

On July 1, 2015, he became the Rector of the University of Iceland. From 2009 to 2015, he was the Pro Rector of the Science and Academic Affairs and the Professor of electrical and computer engineering with the University of Iceland. He is a cofounder of the biomedical start-up company Oxymap. His research interests include remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in those fields.

Dr. Benediktsson was the 2011–2012 President of the IEEE Geoscience and Remote Sensing Society (GRSS) and has been on the GRSS Administrative Committee since 2000. He served as the Editor-in-Chief for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) from 2003 to 2008 and as an Associate Editor for TGRS since 1999, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003, and IEEE ACCESS since 2013. He is on the Editorial Board of the PROCEEDINGS OF THE IEEE and the International Editorial Board of the *International Journal of Image and Data Fusion*. He served as the Chair of the Steering Committee of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (J-STARS) from 2007 to 2010. He received the Stevan J. Kristof Award from Purdue University in 1991 as an outstanding graduate student in remote sensing, the Icelandic Research Council's Outstanding Young Researcher Award in 1997, the IEEE Third Millennium Medal in 2000, the yearly research award from the Engineering Research Institute of the University of Iceland in 2006, the Outstanding Service Award from the IEEE GRSS in 2007, and the IEEE/VFI Electrical Engineer of the Year Award in 2013. He was a corecipient of the University of Iceland's Technology Innovation Award in 2006, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Paper Award in 2012, the IEEE GRSS Highest Impact Paper Award in 2013, and the International Journal of Image and Data Fusion Best Paper Award in 2014. He is a member of the Association of Chartered Engineers in Iceland (VFI), Societas Scinetiarum Islandica, Tau Beta Pi, and the 2014 IEEE Fellow Committee, and a Fellow of SPIE.

Jocelyn Chanussot (M'04–SM'04–F'12) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995 and the Ph.D. degree from Savoie University, Annecy, France, in 1998.

In 1999, he was with the Geography Imagery Perception Laboratory for the Delegation Generale de l'Armement (DGA - French National Defense Department). Since 1999, he has been with Grenoble INP, where he was an Assistant Professor from 1999 to 2005, an Associate Professor from 2005 to 2007, and is currently a Professor of signal and image processing. He is conducting his research at the Grenoble Images Speech Signals and Automatics Laboratory (GIPSA-Lab). From 2014 to 2015, he was a Visiting Professor with the University of California, Los Angeles, CA, USA. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA, KTH Royal Institute of Technology, Stockholm, Sweden, and National University of Singapore, Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland. His research interests include image analysis, multicomponent image processing, nonlinear filtering, and data fusion in remote sensing.

Dr. Chanussot served as the founding President of the IEEE Geoscience and Remote Sensing Society (GRSS) French Chapter (2007–2010) which received the 2010 IEEE GRSS Chapter Excellence Award. He was a member of the IEEE GRSS Administrative Committee (2009–2010), in charge of membership development. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS). He was the Chair (2009–2011) and Cochair of the GRS Data Fusion Technical Committee (2005–2008). He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006–2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He was an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (2005–2007) and for *Pattern Recognition* (2006–2008). Since 2007, he has been an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. Since 2011, he has been the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. In 2013, he was a Guest Editor for the PROCEEDINGS OF THE IEEE and in 2014, a Guest Editor for the IEEE SIGNAL PROCESSING MAGAZINE. He was the corecipient of the NORSIG 2006 Best Student Paper Award, the IEEE GRSS 2011 and 2015 Symposium Best Paper Awards, the IEEE GRSS 2012 Transactions Prize Paper Award, and the IEEE GRSS 2013 Highest Impact Paper Award. He is a member of the Institut Universitaire de France (2012–2017).