Ideal Regularized Composite Kernel for Hyperspectral Image Classification

Jiangtao Peng, Hong Chen, Yicong Zhou, *Senior Member, IEEE*, and Luoqing Li

***Abstract*—This paper proposes an ideal regularized composite kernel (IRCK) framework for hyperspectral image (HSI) classification. In learning a composite kernel, IRCK exploits spectral information, spatial information, and label information simultaneously. It incorporates the labels into standard spectral and spatial kernels by means of the ideal kernel according to a regularization kernel learning framework, which captures both the sample similarity and label similarity and makes the resulting kernel more appropriate for specific HSI classification tasks. With the ideal regularization, the kernel learning problem has a simple analytical solution and is very easy to implement. The ideal regularization can be used to improve and to refine state-of-the-art kernels, including spectral kernels, spatial kernels, and spectral-spatial composite kernels. The effectiveness of the proposed IRCK is validated on three benchmark hyperspectral datasets. Experimental results show the superiority of our IRCK method over the classical kernel methods and state-of-the-art HSI classification methods.**

***Index Terms*—Composite kernel (CK), hyperspectral image (HSI) classification, ideal kernel, regularization.**

I. Introduction

HYPERSPECTRAL remote sensors capture digital images in hundreds of narrow and continuous spectral bands spanning the visible to infrared spectrum [1]. The rich spectral information makes hyperspectral images (HSIs) being applied in different fields, such as military, agriculture, and mineralogy. Among all these applications, classification is a very important topic. Various HSI classification methods have been developed in the past decades [2]–[6]. Traditional HSI classification methods usually discriminate and classify the pixels by measuring the similarity among different spectral curves implicitly based on assumption that HSI samples in the same class have similar spectral characteristics. The key to success for these classification methods is to learn an accurate similarity metric between samples.

In order to learn a desirable similarity metric, kernel functions and kernel methods are introduced into the HSI classification and have shown good classification performance [7], [8]. Kernel methods can solve the high-dimensional HSI classification problem effectively and are easy to measure the linear/nonlinear relations between hyperspectral samples in reproducing kernel hilbert space (RKHS) [8]. In the HSI classification, there are mainly three kinds of kernels: spectral kernels, spatial kernels, and spectral-spatial composite kernels.

Considering each spectral pixel as a sample or a pattern, classical kernels in machine learning can be set as spectral kernels to measure the similarities between different spectral pixels. The commonly used spectral kernels are Gaussian radial basis function (RBF), polynomial and linear kernels [8]. Taking into account spectral meaning and behavior, spectral-angle-based kernel was proposed to tackle the variation of spectral energy [9]. Considering that the useful information for classification is not equally distributed across bands, a spectrally weighted kernel was proposed to highlight the informative spectral bands [10]. Considering that the discriminant capability of single kernel is insufficient, representative multiple kernel learning (RMKL) and discriminative multiple kernel learning (DMKL) algorithms were proposed [11], [12]. The RMKL finds the max-variance kernel by learning the linear combination of basis kernels [11], whereas the DMKL learns an optimal combined kernel by maximizing separability in the RKHS [12]. Exploiting a large amount of unlabeled samples for kernel regularization locally, a bagged or cluster kernel encoding the similarity between unlabeled samples was proposed [13]. Similarly, graph Laplacian kernel was proposed for the semisupervised HSI classification [14].

Spectral kernels are constructed based on the spectral information, whereas spatial kernels use the spatial information. As spatial neighboring pixels usually belong to the same class, local spatial features extracted from the spatial neighborhood can be used to represent neighboring pixels, and then used to generate the spatial kernel. The local spatial features can be the mean or standard deviation features extracted from a squared neighborhood [15], median features extracted from an adaptive morphological neighborhood [16], and morphological profiles (MPs) [17]. Among all these features, mean feature is the most commonly used one. Rather than using the mean feature that computes the average of the spatially neighboring pixels in the

Manuscript received May 24, 2016; revised September 4, 2016 and October 9, 2016; accepted October 23, 2016. Date of publication November 14, 2016; date of current version March 22, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 41501392, Grant 11371007, and Grant 11671161, in part by the Natural Science Foundation of Hubei under Grant 2015CFB327, in part by the Fundamental Research Funds for the Central Universities (Program 2014PY025, 2662015PY046), in part by the Macau Science and Technology Development Fund under Grant FDCT/016/2015/A1, and in pary by the Research Committee at University of Macau under Grant MYRG2014-00003-FST and Grant MYRG2016-00123-FST. (*Corresponding author: Jiangtao Peng*).

J. Peng and L. Li are with the Faculty of Mathematics and Statistics, Hubei Key Laboratory of Applied Mathematics, Hubei University, Wuhan 430062, China (e-mail: pengjt1982@126.com; lilq@live.cn).

H. Chen is with the College of Science, Huazhong Agricultural University, Wuhan 430070, China (e-mail: chenh@mail.hzau.edu.cn).

Y. Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@umac.mo).

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.2016.2621416

original space, the spatial feature used in mean map kernel [18] or mean filtering kernel [19] is the average of the spatially neighboring pixels in the kernel space. The mean map kernel or mean filtering kernel measures distances between sets of pixels in the high-dimensional feature space. Furthermore, given each spatially neighboring pixel a weight, a soft mean map kernel or neighborhood filtering kernel can be obtained [18], [19]. Similar to the mean map kernel that considers all neighbors in a spatial window, a region kernel measuring the similarity of different pixel regions was proposed [20]. Rather than computing average distances as in mean map kernel, graph kernel was proposed to estimate higher order deviations of all neighboring samples in the feature space [21].

HSIs have both the spectral and spatial characteristics. By exploiting the complementary discriminant information in the spatial and spectral domains, the composite kernel (CK) method is commonly used to perform the joint spatial–spectral classification. Joint consideration of the spectral and spatial textural information, four different composite kernels are proposed [15], including stacked kernel, direct summation kernel, weighted summation kernel, and cross-information kernel. Similarly, sample-cluster composite kernels [18], spatial and spectral activation-function-based composite kernels [22], generalized composite kernels [17], and point-region composite kernels [20], have been proposed for the spectral–spatial classification of HSIs. Besides the composite kernels, the multiple kernel learning methods also combine the rich spatial and spectral features and achieve good performance [23], [24].

However, almost all of the aforementioned kernel-based methods learn the standard kernels from the samples alone (i.e., spectral and spatial samples, or supervised and unsupervised samples), without considering the labels of a dataset. In fact, the label information can be used for the kernel learning and to refine the standard kernels. Exploiting the labels explicitly, an ideal kernel is constructed [25]. It assigns the sample pair with a kernel value 1 if they belong to the same class, and a kernel value 0 if they belong to different classes. The ideal kernel incorporates label similarities, and is usually used for kernel parameter selection [26]. Based on the ideal kernel, an ideal regularization strategy is recently proposed to learn a data-dependent kernel from the labels and shows good performance [27], [28].

In this paper, we propose an ideal regularized (IR) composite kernel (IRCK) framework for spatial–spectral classification of HSIs. In IRCK, we consider spectral and spatial kernels as initial standard kernels, and employ an ideal regularization strategy to refine the initial kernels by incorporating the labels into standard spectral and spatial kernels. Finally, the regularized spatial and spectral kernels are combined to form a CK for the HSI classification.

The main contribution of this paper is that it develops a uniform ideal regularization framework to improve the existing spectral kernels, spatial kernels, and spatial–spectral composite kernels. Although the ideal regularization kernel learning method was proposed in [27], it is the first time to employ it for the HSI classification. It should be noted that the ideal regularization framework in [27] employs both the label information and geometric structure information from unlabeled samples, and the experiments are performed mainly on transductive and semisupervised settings. In [27], the unlabeled samples should be used to produce a manifold regularization term in the ideal regularization framework and the IR methods do not always show the best results for general pattern classification problems. However, in our framework, the unlabeled samples and geometric structure information are not needed and the proposed IRCK method shows consistently better results than other kernel methods for the HSI classification. Due to the spatial local similarity of HSI, the label information incorporated in the ideal regularization model can largely improve the classification performance in the large homogeneous regions. In addition, using the spectral similarity, spatial similarity, and label similarity simultaneously, the proposed IRCK method is much effective and outperforms the classical kernel methods and several state-of-the-art spatial–spectral classification methods.

The rest of this paper is organized as follows. The related work on kernel methods is introduced in Section II. In Section III, the proposed IR kernel method is described. The experimental results and analysis are provided in Section IV. Finally, Section V gives a summary of our work.

II. Related Work

A. Spectral Kernel

Given a set of HSI training sample points, $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$ with $\mathbf{x}_i \in R^d$, a spectral kernel measures the spectral similarity between sample points. In the HSI classification, the commonly used kernels are linear kernel, polynomial kernel, and Gaussian RBF kernel [7], [8], [15]. For example, the Gaussian RBF kernel between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ can be expressed as

$$K^w(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_w^2}\Big) \tag{1}$$

where σ_w is the width of the spectral RBF kernel.

B. Spatial Kernel

Given a pixel $\mathbf{x}_i$, we can extract a local spatial feature vector $\mathbf{x}_i^s$ from its spatial neighborhood $N(\mathbf{x}_i) = \{\mathbf{x}_{i0}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{in}\}$ ($\mathbf{x}_{i0} \triangleq \mathbf{x}_i$) [15]. For two pixels $\mathbf{x}_i$ and $\mathbf{x}_j$, the spatial RBF kernel between them is expressed as

$$K^s(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big(-\frac{\|\mathbf{x}_i^s - \mathbf{x}_j^s\|^2}{2\sigma_s^2}\Big) \tag{2}$$

where σ_s is the width of the spatial RBF kernel. The spatial mean map kernel is defined as follows [18], [19]:

$$\begin{aligned} K^m(\mathbf{x}_i, \mathbf{x}_j) &= \Big\langle \frac{1}{1+n}\sum_{p=0}^{n}\phi(\mathbf{x}_{ip}), \frac{1}{1+n}\sum_{q=0}^{n}\phi(\mathbf{x}_{jq})\Big\rangle \\ &= \frac{1}{(1+n)^2}\sum_{p=0}^{n}\sum_{q=0}^{n}\langle\phi(\mathbf{x}_{ip}), \phi(\mathbf{x}_{jq})\rangle \\ &= \frac{1}{(1+n)^2}\sum_{p=0}^{n}\sum_{q=0}^{n}K(\mathbf{x}_{ip}, \mathbf{x}_{jq}) \end{aligned} \tag{3}$$

where ϕ is a nonlinear feature map.

C. Spectral–Spatial CK

By exploiting the complementary discriminant information in spatial-domain and spectral-domain, the CK method is commonly used to perform the spatial–spectral classification [15], [20], [22]. A typical CK used in SVM-CK is the weighted summation kernel:

$$K^{ws}(\mathbf{x}_i, \mathbf{x}_j) = (1-\mu)K^w(\mathbf{x}_i, \mathbf{x}_j) + \mu K^s(\mathbf{x}_i, \mathbf{x}_j) \quad (4)$$

where μ is a combination coefficient balancing the spatial and spectral similarity information.

Similarly, a weighted combination of spectral kernel and spatial mean map kernel is represented as

$$K^{wm}(\mathbf{x}_i, \mathbf{x}_j) = (1-\mu)K^w(\mathbf{x}_i, \mathbf{x}_j) + \mu K^m(\mathbf{x}_i, \mathbf{x}_j). \quad (5)$$

III. IR Kernel

A. Ideal Kernel

The ideal kernel [25], [29] is defined as

$$T(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1, & y_i = y_j, \\ 0, & y_i \neq y_j. \end{cases} \quad (6)$$

The ideal kernel leads to a perfect classification inspired from an “oracle”: two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ should be considered as “similar” (with kernel value 1) if and only if they belong to the same class ($y_i = y_j$) [25], [29]. In other words, the ideal kernel incorporates the label information and reflects the similarity between labels.

B. Ideal Regularization

The standard spectral and spatial kernels measure the similarity between samples. However, these kernels do not carry any label information of the given data. The label information can be used to refine the standard kernels and to obtain new kernels more suitable for specific classification tasks [27].

In order to embed the label information into a standard kernel K_0 and to learn a desirable kernel K, an IR kernel learning framework is proposed [27], [28]:

$$\min_{K \succeq 0} D(K, K_0) + \gamma \Omega(K) \quad (7)$$

where $D(\cdot,\cdot)$ denotes the divergence between the matrices, $\Omega(\cdot)$ is a regularization term, γ is a tradeoff parameter, and $K \succeq 0$ means K is a symmetric positive semidefinite matrix. The divergence can be chosen as the von Neumann divergence:

$$D(K, K_0) = \mathrm{tr}(K \log K - K \log K_0 - K + K_0) \quad (8)$$

where $\mathrm{tr}(A)$ denotes the trace of matrix A. The regularization term can be defined as $\Omega(K) = -\mathrm{tr}(KT)$, which encodes the label information of the given data samples [27]. Then, the ideal regularization framework (7) can be rewritten as

$$\min_{K \succeq 0} \mathrm{tr}(K \log K - K \log K_0 - K + K_0) - \gamma \mathrm{tr}(KT). \quad (9)$$

Setting the derivative with respect to K to zero for the objective function in (9), it gets $\log K - \log K_0 - \gamma T = 0$. It follows that

$$K^* = \exp(\log K_0 + \gamma T) = K_0 \odot \exp(\gamma T) \quad (10)$$

where $\odot$ denotes the dot product between two matrices.

The kernel (10) can be directly extended to new samples that are never encountered before, this is so-called out-of-sample extensions. Denote $S = K_0^{-1}(K^* + K_0)K_0^{-1}$, the kernel between new points $\mathbf{s}$ and $\mathbf{t}$ is computed as [27], [30]

$$K(\mathbf{s}, \mathbf{t}) = -K_0(\mathbf{s}, \mathbf{t}) + \sum_{i,j=1}^{\ell} S(i,j) K_0(\mathbf{s}, \mathbf{x}_i) K_0(\mathbf{x}_j, \mathbf{t}). \quad (11)$$

Remark 1: The Taylor expansion of (10) is

$$K^* = K_0 + \gamma K_0 \odot T + \frac{\gamma^2}{2!} K_0 \odot T^2 + \cdots. \quad (12)$$

The first term on the right-hand side of the equation is the original kernel, and the rest of the terms are regularized kernels. It demonstrates that the IR kernel can be a linear combination of the original kernel and regularized kernels. When $\gamma = 0$, the IR kernel is reduced to the original kernel. When γ is very small, only the first order regularization term $K_0 \odot T$ plays a role in the ideal regularization. Because T equals to 1 only for sample pairs belonging to the same class, ideal regularization enhances the kernel similarity or kernel function values on sample pairs in the same class. In other words, ideal regularization exploits the sample similarity in K_0, and meanwhile uses the label similarity in T to increase the intraclass similarity.

Remark 2: From (10), we can see that the IR kernel is a dot product between the original kernel and an exponential ideal kernel. The computation of ideal kernel and exponential ideal kernel are relatively simple, so the implement of ideal regularization algorithm is very easy.

C. Ideal Regularized CK

In order to learn an ideal spectral kernel K^w and an ideal spatial kernel K^s and hence an ideal spectral–spatial kernel $K^{ws} = (1-\mu)K^w + \mu K^s$, we propose the following IRCK optimization framework:

$$\begin{aligned} \min_{K^w, K^s \succeq 0} \; & D(K^w, K^{w0}) + D(K^s, K^{s0}) + \gamma \Omega(K^{ws}) \\ = \; & \mathrm{tr}(K^w \log K^w - K^w \log K^{w0} - K^w + K^{w0}) \\ & + \; \mathrm{tr}(K^s \log K^s - K^s \log K^{s0} - K^s + K^{s0}) \\ & - \; \gamma\big((1-\mu)\mathrm{tr}(K^w T) + \mu \mathrm{tr}(K^s T)\big). \end{aligned} \quad (13)$$

The optimal solution of (13) is

$$K^w = K^{w0} \odot \exp(\gamma(1-\mu)T) \quad (14)$$

$$K^s = K^{s0} \odot \exp(\gamma \mu T). \quad (15)$$

And the composite IR kernel is

$$\begin{aligned} K^{ws} &= (1-\mu)K^w + \mu K^s \\ &= (1-\mu)K^{w0} \odot \exp(\gamma(1-\mu)T) \\ &\quad + \mu K^{s0} \odot \exp(\gamma \mu T). \end{aligned} \quad (16)$$

Algorithm 1: SVM with IRCK for the HSI classification.

```
Input: Training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$, regularization parameter $\gamma$.
1 Compute standard spectral kernel $K^{w0}$ and spatial kernel $K^{s0}$ (or $K^{m0}$) based on samples.
2 Compute the ideal kernel $T$ based on labels.
3 Compute the ideal regularized spectral kernel $K^{w}$, spatial kernel $K^{s}$ (or $K^{m}$), and composite kernel $K^{ws}$ (or $K^{wm}$).
4 Extend the ideal regularized kernels to new test samples according to (11).
5 Set the composite kernel on new samples as:
  $K^{ws}(\mathbf{s}, \mathbf{t}) = (1-\mu)K^{w}(\mathbf{s}, \mathbf{t}) + \mu K^{s}(\mathbf{s}, \mathbf{t})$.
6 Perform SVM classification based on the ideal regularized composite kernels.
Output: The prediction label for each sample.
```

If the spatial kernel is the mean map kernel, the CK is

$$K^{wm} = (1-\mu)K^{w} + \mu K^{m}. \tag{17}$$

Incorporating the label information into the spectral and spatial kernels, the proposed IRCK SVM algorithm is summarized in Algorithm 1.

D. Representer Theorem of IRCK-Based SVM

Here, we provide the representer theorem of IRCK SVM.

Theorem 1: Given the training samples $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ and the ideal kernel $T(\mathbf{x}, \mathbf{x}')$ defined in (6), the minimizer of SVM with IRCK admits an expansion:

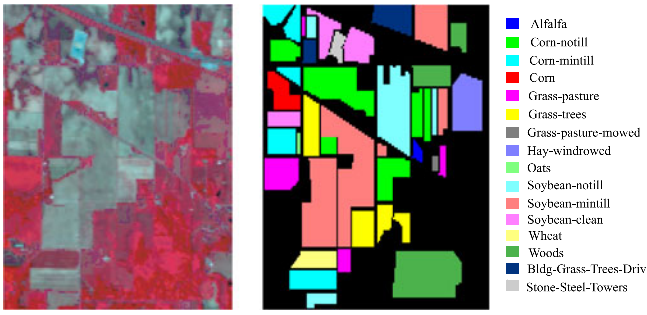
$$f(\mathbf{x}) = (1-\mu)\sum_{i=1}^{\ell}\alpha_i K^{\omega}(\mathbf{x}_i, \mathbf{x}) + \mu\sum_{i=1}^{\ell}\alpha_i K^{s}(\mathbf{x}_i, \mathbf{x}) \tag{18}$$

where $K^{\omega}(\mathbf{x}_i, \mathbf{x})$ and $K^{s}(\mathbf{x}_i, \mathbf{x})$ are the IR spectral and spatial kernels in (14) and (15), respectively.

Remark 3: The proof of the theorem is given in Appendix A. From the Theorem 1, the predictive function (18) can be expanded as

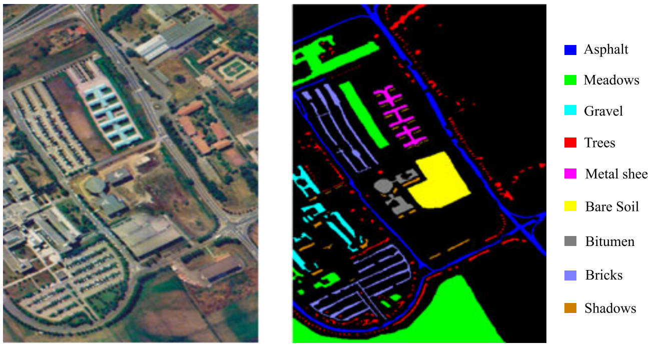
$$\begin{aligned} f(\mathbf{x}) = &\sum_{i=1}^{\ell}\alpha_i\Big((1-\mu)K^{\omega 0}(\mathbf{x}_i, \mathbf{x}) + \mu K^{s0}(\mathbf{x}_i, \mathbf{x})\Big) \\ &+ \gamma\sum_{i=1}^{\ell}\alpha_i\Big((1-\mu)[K^{\omega 0}\odot T](\mathbf{x}_i, \mathbf{x}) \\ &+ \mu[K^{s0}\odot T](\mathbf{x}_i, \mathbf{x})\Big) + \cdots \end{aligned} \tag{19}$$

The additional terms in $f(\mathbf{x})$ enrich the representation ability of IR kernel SVM. That is, the predictive function with IR kernel usually has better approximation ability than the traditional SVM with data independent kernel. When $\gamma = 0$, the representer theorem is consistent with the standard SVM with the spectral and spatial kernels. Moreover, the predictor only uses the spectral information as $\mu = 0$, and just uses the spatial data as $\mu = 1$.



(a) (b)

Fig. 1. Indian Pines dataset. (a) RGB composite image of three bands 50, 27, and 17. (b) Ground-truth map.



(a) (b)

Fig. 2. University of Pavia dataset. (a) RGB composite image of three bands 60, 30, and 2. (b) Ground-truth map.

IV. Experimental Results

A. Datasets

Three public HSI datasets are used in the experiments.[1]

1) Indian Pines: Acquired by the AVIRIS sensor in 1992. The image scene contains 145 × 145 pixels and 220 spectral bands, where 20 channels were discarded because of atmospheric affection. There are 16 classes in the data. The total number of samples is 10 249 ranging from 20 to 2455 in each class. The false color composition of bands 50, 27, and 17 and the ground-truth map are shown in Fig. 1.

2) Pavia University: Acquired in 2001 by the ROSIS instrument over the city of Pavia, Italy. This image scene corresponds to the University of Pavia and has the size of 610 × 340 pixels and 115 spectral bands. After discarding noisy and water absorption bands, 103 bands are retained. The data contain nine ground-truth classes. The false color composition of bands 60, 30, and 2 and the ground-truth map are shown in Fig. 2.

3) Botswana: Acquired by NASA EO-1 satellite over the Okavango Delta, Botswana in May 31, 2001 [31]. The image scene has the size of 1476 × 256 pixels. By discarding water absorption and noisy bands, 145 bands are retained. The data contain 3428 samples from 14 identified classes. The false color

[1] http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

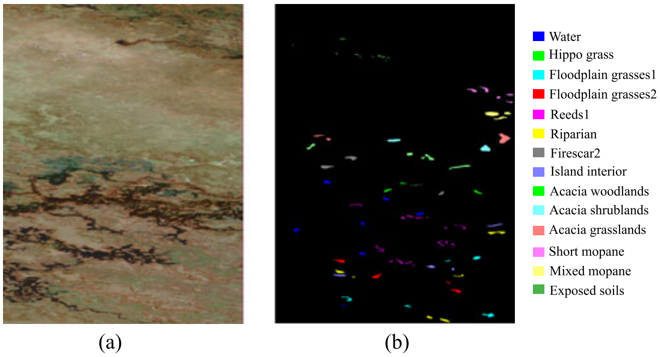


Fig. 3. Botswana dataset. (a) RGB composite image of three bands 51, 149, and 31. (b) Ground-truth map.

composition of bands 51, 149, and 31 and the ground-truth map are shown in Fig. 3.

B. Results

1) Comparison With Kernel Methods: We first compare the proposed IR kernel method with classical kernel methods. Five classical SVM classification methods are considered, including spectral SVM (K^{ω}), spatial SVM (K^{s}), spectral–spatial SVM (SVM with CK, SVM-CK, $K^{\omega s}$) [15], SVM with mean map kernel (K^{m}), and SVM with composite mean map kernel ($K^{\omega m}$) [18]. Imposing ideal regularization on SVM kernels, the corresponding IR SVM methods are compared. The classification performance is assessed on the testing set by the overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ). All data are normalized to have a unit ℓ_2 norm. CK refers to the spatial and spectral weighted summation kernel. Gaussian kernel is used in all SVM algorithms, and LIBSVM software is used to implement SVM [32]. For the spatial-based methods, 9×9 neighborhood window is used.

We investigate the performance of the proposed IRCK methods under different numbers of labeled samples per class. We randomly choose $M = 15, 20, 25, 30, 35, 40$ samples from each class to form the training set, respectively (For the class less than M samples, half of total samples are chosen). The remaining samples consist of testing set.

The classification overall accuracies, average accuracies, and κ coefficients under different numbers of training samples for three datasets are shown in Tables I–III, respectively. From results in these tables, we can conclude:

1) The proposed IRCK methods show a significant improvement over the existing spectral, spatial, and spectral–spatial kernel methods in the case of different numbers of labeled samples. The improvement of IR kernel over the corresponding original kernel demonstrates that the ideal regularization can enhance the kernel discriminant ability.
2) The IR composite mean map kernel ($K^{\omega m}$-IR) provides the best classification results. Compared with the original composite mean map kernel ($K^{\omega m}$–Ori), $K^{\omega m}$–IR increases the OA over different number of labeled samples by 3.7% in average for Indian Pines, by 2.2% in average for Pavia University, and by 0.85% in average for Botwana dataset. Compared with the commonly used benchmark SVM-CK algorithm ($K^{\omega s}$-Ori), $K^{\omega m}$-IR increases the OA by 8.2%, 5.4%, and 1.1% in average for three datasets, respectively.
3) The proposed IRCK is effective in the case with limited training samples. When the number of labeled samples is limited, the kernel similarity measured by samples is insufficient to reflect the class discrepancy. In this case, the label similarity in ideal kernel can assist the sample similarity to obtain a reliable metric and desirable classification result.
4) For the spectral SVM (K^{ω}), the corresponding IR kernel has little or no improvements over the original kernel, especially for the University of Pavia dataset. However, for the spatial SVM (K^{s}), the corresponding IR kernel largely improves the original kernel. Because the ideal regularization enhances the kernel similarity between samples in the same class and the spectral kernel similarity is usually less accurate than spatial kernel similarity, ideal regularization on spectral kernel is less effective than that on spatial kernel.
5) Although the improvement on spectral kernel is marginal, the improvement on spectral–spatial CK is more significant than that on spatial kernel. It demonstrates that the IRCK is not simply a combination of IR spectral and spatial kernels. In an ideal regularization CK framework (16) or (17), the spectral kernel, spatial kernel, and CK are improved simultaneously.

Figs. 4 and 5 show the classification maps for the Indian Pines and University of Pavia datasets in the case of 40 labeled samples per class for training, respectively. The maps correspond to the classification results using different original kernels and IR kernels. It can be seen that the IR kernel methods show relatively better results than the original kernel methods in terms of consistent classification results with little noise.

2) Comparison With Spatial–Spectral Classification Methods: Furthermore, we compare the proposed IR kernel method with some state-of-the-art spatial–spectral classifiers:

1) Spatial–spectral preprocessing and feature extraction methods: Gabor-filtering-based KELM (Gabor-KELM) and multihypothesis-based KELM (MH-KELM) [33];
2) probability-based postprocessing methods: edge-preserving filtering (EPF) [34], and maximizer of the posterior marginal by loopy belief propagation (MPM-LBP) [35];
3) spatial-spectral sparse coding models: joint sparse representation (JSR) [36] and JSR with nonlocal weight (JSR-NLW) [37], spatial aware dictionary learning (SADL) [38];
4) deep learning method: deep belief network (DBN) [39], [40]; and
5) multiple kernel learning method: discriminative MKL (DMKL) [12].

In this experiment, we randomly choose 1%, 3%, 5%, 7%, and 9% labeled samples per class for training for three datasets, respectively (For the class with extremely limited samples, at least three samples are chosen). The remaining labeled samples are used for testing. The classification OAs are shown in Tables IV–VI. It can be seen that the proposed IR composite

TABLE I
CLASSIFICATION ACCURACIES (%) UNDER DIFFERENT NUMBERS OF LABELED SAMPLES FOR INDIAN PINES DATASET

M	Index	K^{ω}		K^{s}		$K^{\omega s}$		K^{m}		$K^{\omega m}$	
		Ori	IR	Ori	IR	Ori	IR	Ori	IR	Ori	IR
15	OA	60.35±1.53	61.30±1.23	78.33±2.91	82.59±2.04	78.61±2.47	84.98±1.94	82.70±2.08	86.22±1.94	83.31±2.20	**86.94±2.12**
	AA	72.07±1.36	72.68±1.49	87.19±1.36	90.73±0.92	87.80±1.36	92.29±0.78	90.86±1.20	92.62±1.09	91.39±1.31	**93.26±1.20**
	κ	55.62±1.57	56.61±1.33	75.58±3.17	80.36±2.24	75.91±2.71	83.02±2.14	80.47±2.29	84.41±2.15	81.14±2.42	**85.21±2.35**
20	OA	63.13±1.30	64.94±1.38	82.04±0.73	86.55±1.78	82.32±0.87	89.33±1.57	86.94±1.26	91.16±1.71	87.36±1.19	**92.05±1.77**
	AA	74.49±1.50	75.31±1.47	89.95±0.73	93.02±1.02	90.38±0.73	94.87±0.69	93.46±0.53	95.32±0.75	93.80±0.58	**95.95±0.70**
	κ	58.67±1.40	60.66±1.48	79.74±0.78	84.79±1.97	80.05±0.95	87.91±1.76	85.23±1.39	89.96±1.93	85.69±1.32	**90.97±1.99**
25	OA	66.48±1.86	68.04±1.61	83.07±1.55	87.73±1.39	83.66±1.31	89.85±0.85	87.88±1.45	91.53±1.31	88.28±1.35	**92.23±1.21**
	AA	77.02±1.33	78.14±1.22	90.47±1.16	93.70±0.54	91.08±1.01	95.08±0.54	93.86±0.82	95.46±0.71	94.32±0.71	**96.09±0.61**
	κ	62.34±1.99	64.06±1.72	80.82±1.70	86.10±1.54	81.50±1.45	88.46±0.94	86.25±1.62	90.36±1.47	86.69±1.50	**91.15±1.37**
30	OA	67.84±1.62	68.72±1.67	86.21±1.60	89.57±1.36	86.48±1.59	92.13±1.42	90.27±1.66	93.57±1.20	90.81±1.65	**94.45±1.41**
	AA	77.71±1.46	78.47±1.55	92.22±1.24	94.78±0.75	92.80±1.03	96.23±0.65	95.04±0.82	96.56±0.59	95.52±0.82	**97.24±0.65**
	κ	63.80±1.78	64.76±1.81	84.33±1.80	88.16±1.53	84.66±1.79	91.04±1.60	88.93±1.87	92.67±1.36	89.54±1.86	**93.67±1.60**
35	OA	69.67±1.91	71.23±2.06	86.90±1.57	89.88±1.56	87.99±1.27	93.18±0.88	91.87±1.53	94.76±1.04	92.46±1.44	**95.65±0.94**
	AA	79.15±1.13	80.02±1.11	93.36±0.75	94.97±0.67	94.10±0.70	96.87±0.44	96.08±0.74	97.31±0.53	96.53±0.70	**97.87±0.56**
	κ	65.86±2.07	67.53±2.25	85.12±1.75	88.50±1.75	86.35±1.43	92.22±0.99	90.74±1.73	94.02±1.19	91.40±1.62	**95.02±1.07**
40	OA	71.08±0.81	72.40±0.99	88.31±1.36	89.08±3.18	89.27±1.36	94.20±0.66	92.60±1.30	95.18±0.67	93.19±1.09	**96.07±0.67**
	AA	80.12±1.21	81.25±1.23	93.77±1.19	94.01±2.55	94.43±1.05	97.38±0.47	96.37±0.62	97.54±0.39	96.86±0.53	**98.04±0.44**
	κ	67.38±0.88	68.83±1.10	86.69±1.53	87.62±3.53	87.79±1.53	93.37±0.74	91.55±1.47	94.49±0.76	92.16±1.24	**95.50±0.77**

TABLE II
CLASSIFICATION ACCURACIES (%) UNDER DIFFERENT NUMBERS OF LABELED SAMPLES FOR PAVIA UNIVERSITY DATASET

M	Index	K^{ω}		K^{s}		$K^{\omega s}$		K^{m}		$K^{\omega m}$	
		Ori	IR	Ori	IR	Ori	IR	Ori	IR	Ori	IR
15	OA	71.44±2.26	71.36±2.29	85.39±4.06	86.83±4.08	86.95±3.06	89.10±3.67	90.20±3.28	92.21±3.23	90.40±3.55	**92.51±3.33**
	AA	78.75±1.31	78.66±1.33	86.48±1.38	87.90±1.53	88.16±0.85	90.08±0.94	93.32±0.71	94.87±0.87	93.87±0.99	**95.38±0.94**
	κ	63.82±2.39	63.72±2.44	81.11±4.84	82.95±4.85	83.06±3.65	85.84±4.38	87.35±4.00	89.90±3.99	87.60±4.37	**90.30±4.13**
20	OA	72.79±3.32	72.68±3.28	86.68±2.21	88.52±1.92	87.83±2.57	90.70±1.80	88.97±3.05	92.73±2.88	89.45±3.08	**93.25±2.94**
	AA	80.14±1.24	80.01±1.26	87.68±0.92	89.62±0.71	88.99±1.25	91.46±0.83	93.48±1.04	95.80±0.98	93.72±1.19	**96.32±1.02**
	κ	65.55±3.62	65.41±3.58	82.73±2.64	85.07±2.32	84.15±3.20	87.85±2.26	85.85±3.73	90.60±3.59	86.44±3.81	**91.26±3.69**
25	OA	76.49±2.60	76.35±2.64	88.66±1.52	89.60±1.01	89.75±1.84	91.86±1.83	93.12±2.08	95.05±1.70	93.25±2.37	**95.47±1.86**
	AA	81.55±0.70	81.39±0.67	89.00±0.90	89.92±0.92	90.61±0.90	92.51±0.72	95.12±0.72	96.60±0.56	95.19±0.98	**97.05±0.67**
	κ	69.88±2.89	69.71±2.93	85.18±1.91	86.38±1.27	86.58±2.29	89.33±2.30	91.03±2.63	93.52±2.17	91.20±2.98	**94.06±2.40**
30	OA	79.07±1.37	78.84±1.35	89.70±1.61	91.02±1.04	90.87±1.31	93.40±0.74	94.84±1.15	96.58±0.73	95.32±1.51	**97.00±0.73**
	AA	82.72±0.57	82.39±0.59	89.62±1.03	90.99±0.86	91.28±0.96	93.59±0.86	95.99±0.62	97.24±0.53	96.43±0.89	**97.72±0.53**
	κ	72.87±1.71	72.57±1.69	86.50±2.01	88.19±1.34	88.00±1.68	91.31±0.94	93.21±1.48	95.50±0.95	93.84±1.95	**96.04±0.95**
35	OA	78.38±2.42	78.31±2.16	90.88±0.99	91.83±1.09	92.39±0.91	94.66±0.52	94.94±1.36	96.90±1.25	95.45±1.09	**97.32±1.27**
	AA	83.03±0.67	82.76±0.49	90.75±0.89	91.90±0.85	92.42±0.56	94.48±0.53	96.14±0.60	97.58±0.61	96.51±0.59	**98.03±0.60**
	κ	72.17±2.77	72.06±2.45	88.02±1.25	89.26±1.41	89.97±1.17	92.95±0.68	93.35±1.75	95.92±1.62	94.01±1.41	**96.46±1.65**
40	OA	79.13±0.91	78.91±1.05	91.54±0.99	92.20±1.58	93.41±0.87	94.99±0.72	96.10±0.78	97.65±0.65	96.45±0.81	**98.03±0.65**
	AA	83.74±0.57	83.37±0.57	91.26±0.98	92.04±0.95	93.39±0.71	94.94±0.77	96.92±0.53	98.09±0.32	97.26±0.45	**98.54±0.29**
	κ	73.10±1.04	72.81±1.20	88.88±1.29	89.74±2.00	91.30±1.12	93.38±0.94	94.85±1.02	96.89±0.85	95.32±1.06	**97.40±0.86**

mean map kernel ($K^{\omega m}$-IR) shows the best overall performances on three datasets. Compared with DMKL, the proposed IRCK method shows slightly better results on the Indian Pines and University of Pavia datasets, and comparable results on the Botswana dataset. DMKL combines ten basic kernels, which are constructed on the MPs with the diamond structure element of size [3, 5, 7, 9, 11, 13, 15, 17, 19, 21], respectively [12]. While the propose IR composite mean map kernel combines a spectral kernel and a spatial mean map kernel constructed on spatial features in a spatial window of width 9. That is, we only use two kernels whereas DMKL uses ten kernels. It should also be noted that DBN provides bad results in the case of limited training samples.

C. Parameter Analysis

Now, we investigate the sensitivity of the proposed algorithm on different parameters. We take the SVM-CK-IR algorithm as an example, and choose 15 and 100 labeled samples per class to form the training set and testing set. For

TABLE III
CLASSIFICATION ACCURACIES (%) UNDER DIFFERENT NUMBERS OF LABELED SAMPLES FOR BOTSWANA DATASET

M	Index	K^{ω}		K^{s}		$K^{\omega s}$		K^{m}		$K^{\omega m}$	
		Ori	IR	Ori	IR	Ori	IR	Ori	IR	Ori	IR
15	OA	90.39±0.99	91.28±0.92	97.44±0.61	98.68±0.47	97.69±0.62	98.83±0.52	97.96±0.47	99.25±0.24	98.28±0.58	**99.36±0.20**
	AA	91.49±0.91	92.35±0.87	97.62±0.60	98.79±0.38	97.87±0.65	98.94±0.43	98.14±0.45	99.33±0.24	98.48±0.62	**99.42±0.21**
	κ	89.58±1.07	90.55±0.99	97.23±0.66	98.56±0.51	97.50±0.68	98.74±0.56	97.79±0.51	99.19±0.26	98.14±0.63	**99.30±0.22**
20	OA	91.02±0.87	91.57±0.55	98.12±0.60	99.00±0.44	98.31±0.54	99.09±0.45	98.34±0.43	99.44±0.50	98.46±0.46	**99.72±0.18**
	AA	92.13±0.75	92.68±0.45	98.17±0.69	99.05±0.43	98.37±0.55	99.16±0.43	98.45±0.42	99.49±0.47	98.64±0.45	**99.73±0.18**
	κ	90.26±0.94	90.86±0.60	97.96±0.65	98.92±0.48	98.17±0.58	99.02±0.49	98.20±0.47	99.40±0.54	98.33±0.50	**99.70±0.19**
25	OA	92.35±0.60	93.07±0.57	98.66±0.59	99.24±0.39	98.75±0.48	99.40±0.28	99.03±0.38	99.67±0.27	98.94±0.32	**99.76±0.23**
	AA	93.25±0.61	93.96±0.57	98.67±0.65	99.30±0.35	98.79±0.53	99.43±0.28	99.11±0.37	99.69±0.28	99.06±0.32	**99.77±0.24**
	κ	91.70±0.65	92.49±0.62	98.55±0.64	99.18±0.42	98.64±0.52	99.35±0.31	98.94±0.41	99.65±0.29	98.85±0.34	**99.74±0.25**
30	OA	92.92±0.51	93.35±0.51	98.85±0.18	99.42±0.31	99.04±0.16	99.61±0.20	99.07±0.20	99.72±0.25	99.22±0.22	**99.88±0.17**
	AA	93.77±0.43	94.18±0.46	98.85±0.34	99.45±0.34	99.09±0.19	99.66±0.19	99.13±0.32	99.77±0.21	99.32±0.21	**99.90±0.14**
	κ	92.32±0.55	92.79±0.56	98.76±0.19	99.37±0.34	98.96±0.17	99.58±0.22	98.99±0.21	99.70±0.27	99.15±0.23	**99.87±0.18**
35	OA	92.85±0.62	93.35±0.46	98.90±0.45	99.26±0.30	98.91±0.39	99.45±0.29	99.04±0.44	99.67±0.24	99.04±0.30	**99.76±0.25**
	AA	93.75±0.56	94.23±0.48	98.95±0.43	99.30±0.29	99.02±0.32	99.49±0.25	99.17±0.37	99.72±0.20	99.19±0.26	**99.80±0.21**
	κ	92.24±0.68	92.79±0.50	98.81±0.49	99.20±0.32	98.82±0.42	99.41±0.31	98.95±0.48	99.64±0.26	98.96±0.33	**99.74±0.27**
40	OA	93.46±0.65	93.61±0.81	99.27±0.36	99.59±0.23	99.37±0.31	99.64±0.24	99.40±0.32	99.91±0.12	99.41±0.25	**99.94±0.09**
	AA	94.37±0.50	94.46±0.60	99.34±0.32	99.64±0.20	99.45±0.30	99.69±0.20	99.51±0.27	99.93±0.10	99.53±0.20	**99.95±0.08**
	κ	92.90±0.71	93.06±0.87	99.20±0.39	99.56±0.25	99.32±0.34	99.61±0.26	99.35±0.35	99.90±0.13	99.36±0.27	**99.93±0.10**

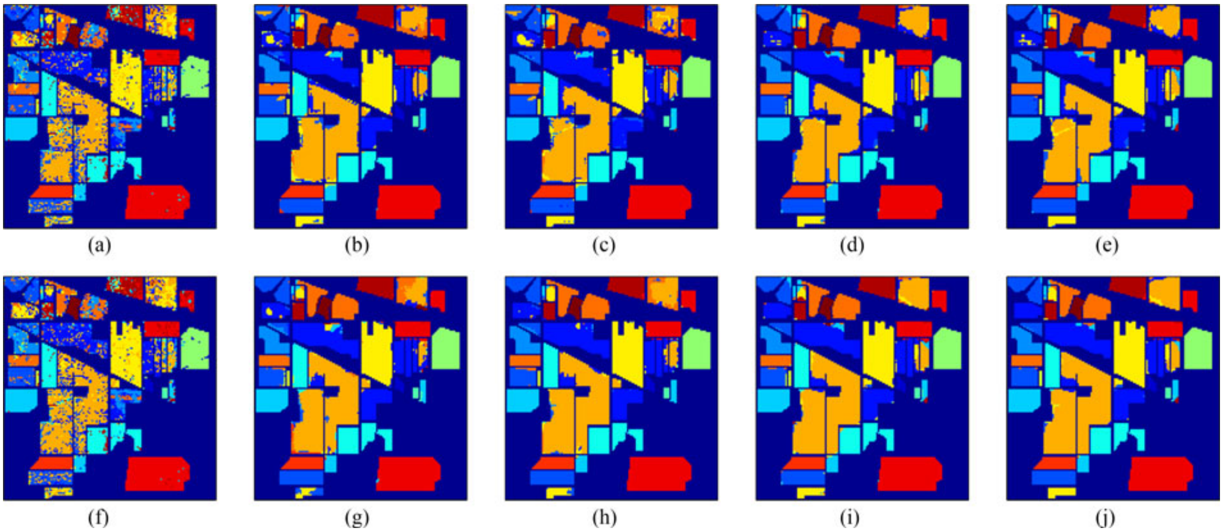


Fig. 4. Classification maps for the Indian Pines dataset. The first and second rows correspond to the original and IR kernels, respectively. (a) K^{ω}-Ori (OA = 71.08%). (b) K^{s}-Ori (OA = 88.31%). (c) $K^{\omega s}$-Ori (OA = 89.27%). (d) K^{m}-Ori (OA = 92.60%). (e) $K^{\omega m}$-Ori (OA = 93.19%). (f) K^{ω}-IR (OA = 72.40%). (g) K^{ω}-IR (OA = 89.08%). (h) $K^{\omega s}$-IR (OA = 94.20%). (i) K^{m}-IR (OA = 95.18%). (j) $K^{\omega m}$-IR (OA = 96.07%).

SVM-CK-IR, there are IR parameter γ and SVM-CK parameters μ, σ_ω, and σ_s. The parameters γ and μ vary in the intervals $\gamma \in [10^{-6}, 10^{-5}, \ldots, 10^{1}]$ and $\mu \in [0, 0.1, \ldots, 1]$. The spectral and spatial Gaussian kernel width σ_ω and σ_s change in the range $\{0.01, 0.05, 0.1, 0.2, \ldots, 1.5\}$. We first investigate the effect of IR parameter γ on SVM-CK-IR. The OAs versus γ are shown in Fig. 6, where the proposed IR kernel method provides stable results over a wide range of regularization parameters. Similarly, we analyze the effect of spatial–spectral combination coefficient μ. Fig. 7 shows the OAs of SVM-CK-IR under different combination coefficients. It can be seen that the proposed SVM-CK-IR is stable when μ is no less than 0.4.

Next, we show the effect of two kernel parameters: spatial and spectral RBF kernel parameters σ_s and σ_ω. The OAs of SVM-CK and SVM-CK-IR versus spatial and spectral kernel parameters are shown in Fig. 8. From the figure, it can be clearly seen that SVM-CK-IR is less sensitive to kernel parameters than the original SVM-CK. The original SVM-CK achieves the best OA at a narrow band whereas the proposed SVM-CK-IR shows good performance over a wide range of spatial and spectral kernel parameters. It demonstrates that the rationality and veracity of kernel function (IR kernel) can reduce the dependence on the kernel parameters. Based on the results, we set the width of spectral kernel K^{ω} and spatial kernel K^{s} as $(1, 0.5)$ for three datasets.

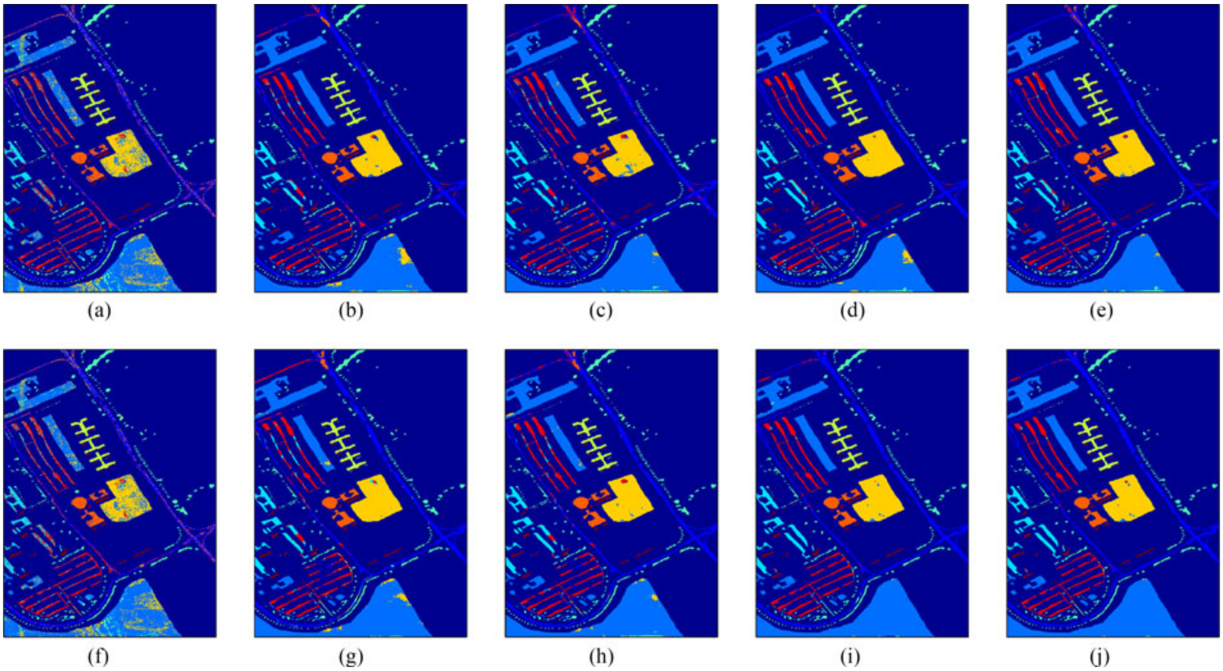


Fig. 5. Classification maps for the Pavia University dataset. The first and second rows correspond to the original and IR kernels, respectively. (a) K^{ω} (OA $=$ 79.13%). (b) K^{s} (OA $=$ 91.54%). (c) $K^{\omega s}$ (OA $=$ 93.41%). (d) K^{m} (OA $=$ 96.10%). (e) $K^{\omega m}$ (OA $=$ 96.45%). (f) K^{ω}-IR (OA $=$ 78.91%). (g) K^{ω}-IR (OA $=$ 92.20%). (h) $K^{\omega s}$-IR (OA $=$ 94.99%). (i) K^{m}-IR (OA $=$ 97.65%). (j) $K^{\omega m}$-IR (OA $=$ 98.03%).

TABLE IV
COMPARISON WITH SPATIAL–SPECTRAL CLASSIFIERS ON INDIAN PINES DATASET

	Gabor-KELM	MH-KELM	EPF	MPM-LBP	JSR	JSR-NLW	SADL	DBN	SVM-CK	DMKL	$K^{\omega m}$-IR
1%	67.96 ± 1.37	78.91 ± 1.32	62.82 ± 6.72	71.09 ± 2.30	67.81 ± 0.85	69.36 ± 0.59	78.80 ± 1.09	46.63 ± 1.34	74.91 ± 2.64	**85.25 ± 2.56**	84.15 ± 2.65
3%	85.80 ± 0.58	92.07 ± 0.63	85.21 ± 3.52	86.04 ± 2.08	83.14 ± 0.76	84.56 ± 0.40	90.73 ± 1.75	56.11 ± 3.10	87.86 ± 0.89	93.67 ± 0.46	**94.67 ± 0.38**
5%	91.68 ± 1.05	95.06 ± 0.37	89.81 ± 0.92	91.55 ± 1.38	88.52 ± 1.42	89.62 ± 1.77	94.01 ± 0.73	64.54 ± 1.94	92.78 ± 0.36	96.11 ± 0.55	**97.40 ± 0.49**
7%	94.66 ± 0.74	96.59 ± 0.39	92.72 ± 0.87	93.46 ± 0.69	92.18 ± 0.47	93.56 ± 0.39	95.54 ± 1.15	69.99 ± 1.18	95.01 ± 0.80	97.31 ± 0.35	**98.22 ± 0.21**
9%	96.73 ± 0.59	97.74 ± 0.21	95.11 ± 0.39	95.29 ± 0.76	94.20 ± 0.51	95.41 ± 0.32	97.45 ± 0.46	74.27 ± 1.56	96.05 ± 0.43	97.95 ± 0.22	**98.72 ± 0.21**

TABLE V
COMPARISON WITH SPATIAL–SPECTRAL CLASSIFIERS ON UNIVERSITY OF PAVIA DATASET

	Gabor-KELM	MH-KELM	EPF	MPM-LBP	JSR	JSR-NLW	SADL	DBN	SVM-CK	DMKL	$K^{\omega m}$-IR
1%	93.46 ± 0.39	95.38 ± 0.72	92.07 ± 3.25	95.60 ± 0.26	80.87 ± 0.84	82.10 ± 0.74	93.10 ± 1.03	64.74 ± 2.56	94.73 ± 0.27	97.80 ± 0.89	**98.20 ± 0.24**
3%	96.87 ± 0.33	98.04 ± 0.16	94.81 ± 0.77	97.90 ± 0.16	84.81 ± 0.40	87.93 ± 0.44	97.31 ± 0.22	78.76 ± 0.29	97.13 ± 0.29	99.43 ± 0.23	**99.62 ± 0.09**
5%	97.62 ± 0.12	98.60 ± 0.06	95.83 ± 0.72	98.32 ± 0.25	87.43 ± 0.67	91.56 ± 0.46	98.71 ± 0.23	85.58 ± 2.64	98.18 ± 0.12	99.63 ± 0.20	**99.72 ± 0.05**
7%	97.99 ± 0.15	98.89 ± 0.16	96.18 ± 1.07	98.58 ± 0.06	89.50 ± 0.45	93.84 ± 0.34	98.94 ± 0.13	85.95 ± 0.48	98.26 ± 0.06	99.80 ± 0.03	**99.83 ± 0.02**
9%	98.25 ± 0.10	99.07 ± 0.08	96.31 ± 1.05	98.62 ± 0.11	91.04 ± 0.27	95.28 ± 0.23	99.04 ± 0.09	88.66 ± 2.05	98.51 ± 0.04	99.86 ± 0.01	**99.87 ± 0.01**

TABLE VI
COMPARISON WITH SPATIAL–SPECTRAL CLASSIFIERS ON BOTSWANA DATASET

	Gabor-KELM	MH-KELM	EPF	MPM-LBP	JSR	JSR-NLW	SADL	DBN	SVM-CK	DMKL	$K^{\omega m}$-IR
1%	83.78 ± 1.75	89.94 ± 0.97	88.12 ± 4.26	87.80 ± 3.00	72.30 ± 2.56	76.61 ± 2.39	88.99 ± 2.96	9.14 ± 0.77	88.36 ± 2.40	**91.88 ± 1.12**	91.83 ± 3.12
3%	93.17 ± 0.86	96.00 ± 0.60	93.01 ± 1.92	93.55 ± 1.16	84.38 ± 1.09	88.26 ± 0.51	94.66 ± 1.88	7.32 ± 2.79	95.75 ± 0.92	96.90 ± 1.72	**97.84 ± 0.91**
5%	95.04 ± 0.77	97.99 ± 0.78	94.67 ± 1.74	96.20 ± 0.91	90.15 ± 1.25	92.88 ± 0.84	97.81 ± 0.55	61.24 ± 9.26	96.74 ± 0.70	**99.38 ± 0.36**	98.93 ± 0.64
7%	96.21 ± 1.38	98.63 ± 1.01	96.54 ± 1.03	97.17 ± 0.77	93.18 ± 1.69	94.95 ± 1.28	98.73 ± 0.45	88.37 ± 3.81	97.69 ± 0.34	99.09 ± 1.03	**99.47 ± 0.28**
9%	97.70 ± 0.16	99.44 ± 0.26	97.00 ± 1.19	97.57 ± 0.61	93.60 ± 1.23	96.37 ± 0.96	99.02 ± 0.33	91.02 ± 2.96	98.38 ± 0.41	**99.83 ± 0.24**	99.59 ± 0.36

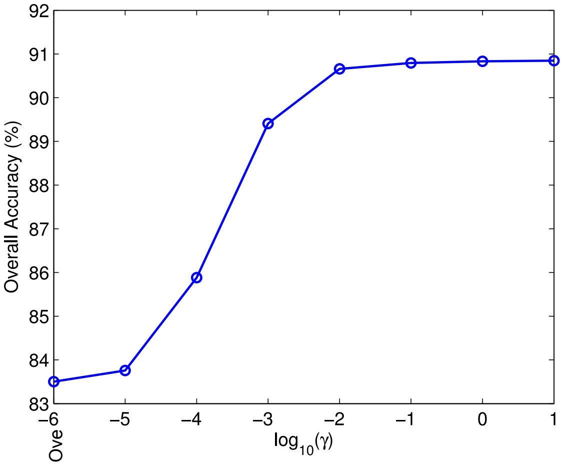


Fig. 6. OA versus IR parameter γ for SVM-CK-IR.

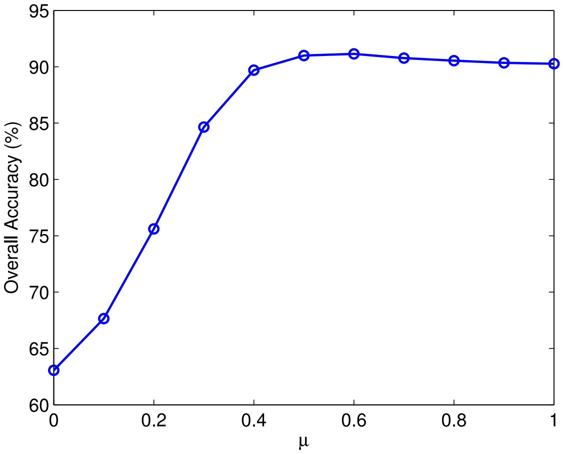


Fig. 7. OA versus combination coefficient μ for SVM-CK-IR.

D. Label Similarity and Class Separability

The kernel measures the similarity between the data in the RKHS. By modifying the standard kernel with ideal regularization, the new regularized kernel embeds both the data similarity and label similarity, which can improve the class separability.

We first use a simulated example to show the effectiveness of label similarity. We choose five samples from "Soybean-mintill" and "Soybean-clean" classes of Indian Pines dataset, respectively. Then, we compute the original weighted summation kernel $K^{\omega s}$-Ori and the corresponding IR kernel $K^{\omega s}$-IR. The kernel values are shown in Fig. 9, where the first five samples belong to "Soybean-mintill" and the last five samples are from "Soybean-clean." Because the ten samples belong to the same Soybean material (two subclasses of Soybean), their differences are very small. It is very difficult to distinguish the subtle difference between them based on the sample similarity (kernel values) as shown in Fig. 9(a). For example, the sample 1 is more similar with sample 9 than samples 3, 4, and 5, although sample 9 comes from a different class. Notwithstanding, when the label similarity is considered, the IR kernel increases the intraclass similarity as shown in Fig. 9(b). From Fig. 9(b), we can see that the samples in $\{1, 2\}$, $\{3, 4, 5\}$ are very similar. Meanwhile, samples 2 and 4 are similar, so the two components $\{1, 2\}$ and $\{3, 4, 5\}$ are connected as a whole and all samples in the first class (samples $\{1, 2, 3, 4, 5\}$) are with higher similarity. Similar results can be obtained for the second class.

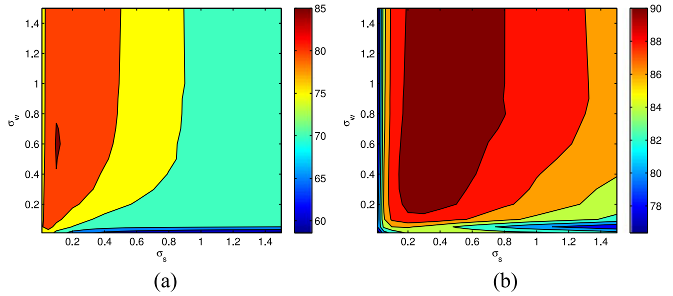


Fig. 8. OA versus spatial and spectral kernel parameters σ_s and σ_ω for SVM-CK (a) and SVM-CK-IR (b). The colorbars on the right of the figures indicate the mapping of OAs into the colormap.

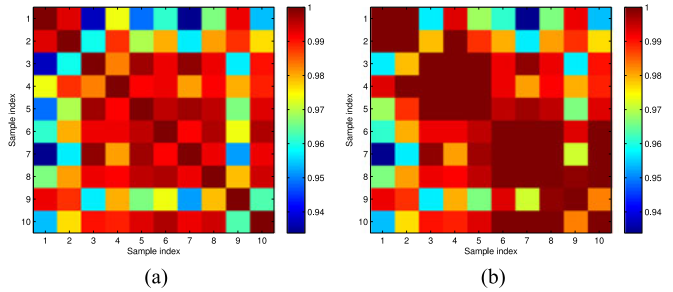


Fig. 9. The Gaussian kernel function values between different samples: (a) Original Gaussian kernel. (b) IR Gaussian kernel. The first five samples "1–5" belong to the class "Soybean-mintill" and the last five samples "6–10" are from the class "Soybean-clean." The colorbars on the right of the figures indicate the mapping of kernel values into the colormap. The more similar the two samples, the higher the kernel values.

In the following, we show the classification class accuracy of the proposed algorithm on three real datasets. For simplicity, we only compare SVM-CK and SVM-CK-IR and show the results in the case of $M = 15$ labeled samples per class for training. The results in Table VII indicate that by incorporating the label information into the standard spatial–spectral CK, the new regularized kernel significantly improves the class separability. In detail, the class accuracy of SVM-CK-IR is higher than that of the original SVM-CK almost for each class and each dataset. This is because the ideal regularization can increase the intraclass similarity and decrease the interclass similarity (if the value 0 changes to -1 in ideal kernel), which is similar to linear discriminant analysis (LDA). Furthermore, the standard derivation of OA of SVM-CK-IR is relatively smaller than that of SVM-CK especially for Indian Pines and Botswana datasets. This shows the proposed IR kernel method is generally more stable than the original kernel method.

E. Computational Time

In Section III-B, we have presented that the computational complexity of IR kernel method is almost the same as that of standard kernel method. Now, we show the computational time of different algorithms under different numbers of labeled samples per class on Indian Pines datasets. The results are shown in Fig. 10, where the computational times of the initial kernel are

TABLE VII
CLASSIFICATION CLASS ACCURACIES OF SVM-CK AND SVM-CK-IR ON THREE DATASETS

Indian Pines			University of Pavia			Botswana		
Class	SVM-CK	SVM-CK-IR	Class	SVM-CK	SVM-CK-IR	Class	SVM-CK	SVM-CK-IR
1	98.06±4.08	99.68±**1.02**	1	83.56±3.34	85.79±**3.23**	1	98.75±2.47	99.22±**1.69**
2	69.50±9.20	79.45±**8.16**	2	87.83±7.63	89.99±8.09	2	100.0±0	100.0±**0**
3	79.29±10.1	85.85±**8.08**	3	79.86±5.78	84.27±7.97	3	98.73±2.11	100.0±**0**
4	96.35±3.37	99.28±**1.09**	4	91.76±4.77	92.39±**4.41**	4	99.90±0.21	100.0±**0**
5	86.82±5.71	90.34±**4.57**	5	99.93±0.12	99.95±**0.09**	5	93.31±4.48	94.25±**3.69**
6	95.83±2.02	98.78±**0.88**	6	87.08±5.02	90.61±5.66	6	94.21±1.91	96.50±**1.85**
7	100.0±0	100.0±**0**	7	93.83±2.77	96.32±**2.42**	7	99.84±0.52	99.92±**0.26**
8	98.40±1.06	99.52±**0.32**	8	80.84±5.18	81.25±7.32	8	99.95±0.17	99.95±**0.16**
9	97.00±4.83	99.00±**3.16**	9	88.79±4.37	89.91±**4.28**	9	96.82±1.28	99.30±**0.60**
10	74.29±6.75	80.58±**6.67**				10	99.91±0.18	100.0±**0**
11	65.91±6.88	74.20±8.15				11	96.55±1.43	98.83±3.47
12	69.07±8.62	86.19±**7.51**				12	94.94±4.14	97.95±**2.65**
13	98.47±1.71	99.68±**0.56**				13	99.29±1.02	99.92±**0.25**
14	89.35±4.77	90.09±4.97				14	98.00±3.34	99.38±**1.21**
15	89.22±9.41	94.37±**5.83**						
16	97.18±3.95	99.62±**0.62**						
AA	87.80±1.36	92.29±**0.78**	AA	88.16±0.85	90.08±0.94	AA	97.87±0.65	98.94±**0.43**

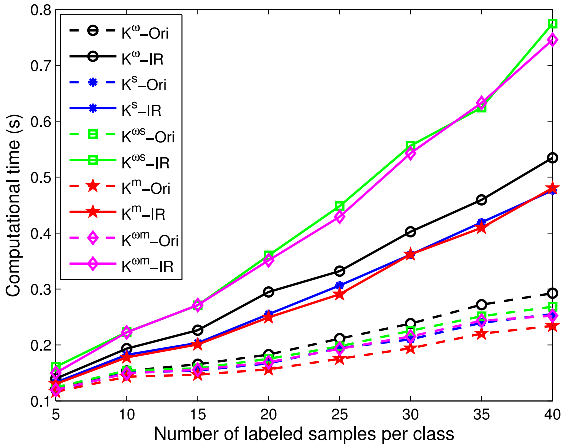


Fig. 10. Computational times of different algorithms on Indian Pines dataset.

not included because the initial kernel is involved in both the standard and regularized kernel methods. It can be seen that the computational time of IR kernel classification method is slightly higher than that of the corresponding original kernel classification method. Notwithstanding, they are in the same order.

V. CONCLUSION

In this paper, we have proposed a new IRCK framework for classification of hyperspectral remote sensing images. Different from the traditional CK methods considering only the sample similarity, the proposed IRCK considers both the sample similarity and label similarity. It incorporates the labels into a standard kernel by means of ideal kernel according to a regularization kernel learning framework. The regularized kernel learning problem is very easy to solve, and suitable for various state-of-the-art kernels. Experimental results have shown that, by exploiting the spectral, spatial, and label information, the proposed IRCK method outperforms state-of-the-art spatial–spectral classification methods.

APPENDIX

A. *Proof of Theorem 1*

Proof. Let $\mathcal{H}_K$ be the RKHS [41], [42] associated with the CK $(1-\mu)K^{\omega}+\mu K^{s}$, and consider the data-dependent space $\text{span}\{(1-\mu)K^{\omega}(\mathbf{x}_i,\cdot)+\mu K^{s}(\mathbf{x}_i,\cdot)\}_{i=1}^{\ell}\subset\mathcal{H}_K$. Let $f\in\mathcal{H}_K$ be the minimizer of IRCK SVM. Then, f can be uniquely decomposed into a component $f_1\in\text{span}\{(1-\mu)K^{\omega}(\mathbf{x}_i,\cdot)+\mu K^{s}(\mathbf{x}_i,\cdot)\}_{i=1}^{\ell}$ and a component f_2 orthogonal to it. Thus, there exist some constants $\{\alpha_i\}_{i=1}^{\ell}$ such that

$$f=f_1+f_2=\sum_{i=1}^{\ell}\alpha_i\Big((1-\mu)K^{\omega}(\mathbf{x}_i,\cdot)+\mu K^{s}(\mathbf{x}_i,\cdot)\Big)+f_2.$$

According to the reproducing property of $\mathcal{H}_K$, we know that for any $\mathbf{x}_j\in\mathcal{L}$

$$\begin{aligned}f(\mathbf{x}_j)&=\Big\langle f,(1-\mu)K^{\omega}(\cdot,\mathbf{x}_j)+\mu K^{s}(\cdot,\mathbf{x}_j)\Big\rangle\\&=\sum_{i=1}^{\ell}\alpha_i\Big((1-\mu)K^{\omega}(\mathbf{x}_i,\mathbf{x}_j)+\mu K^{s}(\mathbf{x}_i,\mathbf{x}_j)\Big)\\&\quad+\Big\langle f_2,(1-\mu)K^{\omega}(\cdot,\mathbf{x}_j)+\mu K^{s}(\cdot,\mathbf{x}_j)\Big\rangle\\&=\sum_{i=1}^{\ell}\alpha_i\Big((1-\mu)K^{\omega}(\mathbf{x}_i,\mathbf{x}_j)+\mu K^{s}(\mathbf{x}_i,\mathbf{x}_j)\Big).\end{aligned}$$

The above equation tells us that the empirical error in SVM just depends on the coefficients $\{\alpha_i\}_{i=1}^{\ell}$.

Moreover, observed that

$$\begin{aligned}\|f\|_K&=\Big\|\sum_{i=1}^{\ell}\alpha_i\Big((1-\mu)K^{\omega}(\mathbf{x}_i,\cdot)+\mu K^{s}(\mathbf{x}_i,\cdot)\Big)\Big\|_K+\|f_2\|_K\\&\geq\Big\|\sum_{i=1}^{\ell}\alpha_i\Big((1-\mu)K^{\omega}(\mathbf{x}_i,\cdot)+\mu K^{s}(\mathbf{x}_i,\cdot)\Big)\Big\|_K.\end{aligned}$$

Hence, the minimizer of CK SVM must have $f_2 = 0$. This completes the proof. ■

ACKNOWLEDGMENT

The authors would like to thank Prof. D. Landgrebe for providing the Indian Pines dataset, Prof. P. Gamba for providing the University of Pavia dataset, Prof. M. Crawford for providing the Botswana dataset, Dr. C. Chen for sharing MH source code, Dr. X. Kang for sharing EPF source code, Prof. J. Li for sharing MPM-LBP source code, Dr. A. Soltani-Farani for sharing SADL source code, and Prof. C. Lin for providing LIBSVM toolbox.

REFERENCES

[1] A. Plaza *et al.*, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, 2009.

[2] 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.

[3] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, Jan. 2014.

[4] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[5] Y. Zhou, J. Peng, and C. L. P. Chen, "Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 1082–1095, Feb. 2015.

[6] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[7] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[8] 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.

[9] G. Mercier and M. Lennon, "Support vector machines for hyperspectral image classification with spectral-based kernels," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. 2003*, 2003, pp. 288–290.

[10] B. Guo, S. R. Gunn, R. I. Damper, and J. Nelson, "Customizing kernel functions for SVM-based hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 622–629, Apr. 2008.

[11] Y. Gu, C. Wang, D. You, Y. Zhang, S. Wang, and Y. Zhang, "Representative multiple kernel learning for classification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 7, pp. 2852–2865, Jul. 2012.

[12] Q. Wang, Y. Gu, and D. Tuia, "Discriminative multiple kernel learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3912–3927, Jul. 2016.

[13] D. Tuia and G. Camps-Valls, "Semisupervised remote sensing image classification with cluster kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 6, no. 2, pp. 224–228, Apr. 2009.

[14] G. Camps-Valls, T. Bandos, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.

[15] G. Camps-Valls, L. Gomez-Chova, J. Muñoz Maré, 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.

[16] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A spatial-spectral kernel-based approach for the classification of remote-sensing images," *Pattern Recognit.*, vol. 45, pp. 381–392, 2012.

[17] J. Li, P. R. Marpu, A. Plaza, J. M. 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.

[18] L. Gómez-Chova, G. Camps-Valls, L. Bruzzone, and G. Calpe-Maravilla, "Mean map kernel methods for semisupervised cloud classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 207–220, Jan. 2010.

[19] J. Liu, Z. Wu, Z. Wei, L. Xiao, and L. Sun, "Spatial-spectral kernel sparse representation for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Ses.*, vol. 6, no. 6, pp. 2462–2471, Dec. 2013.

[20] J. Peng, Y. Zhou, and C. L. P. Chen, "Region-kernel-based support vector machines for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 4810–4824, Sep. 2015.

[21] G. Camps-Valls, N. Shervashidze, and K. M. Borgwardt, "Spatio-spectral remote sensing image classification with graph kernels," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 741–745, Oct. 2010.

[22] Y. Zhou, J. Peng, and C. L. P. Chen, "Extreme learning machine with composite kernels for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Ses.*, vol. 8, no. 6, pp. 2351–2360, Jun. 2015.

[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] 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.

[25] J. T. Kwok and I. W. Tsang, "Learning with idealized kernels," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 400–407.

[26] 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.

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

[28] B. Pan, W. Chen, C. Xu, and B. Chen, "A novel framework for learning geometry-aware kernels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 939–951, May 2016.

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

[30] P. Jain, B. Kulis, J. V. Davis, and I. S. Dhillon, "Metric and kernel learning using a linear transformation," *J. Mach. Learn. Res.*, vol. 13, pp. 519–547, 2012.

[31] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[32] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[33] C. Chen, W. Li, H. Su, and K. Liu, "Spectral-spatial classification of hyperspectral image based on kernel extreme learning machine," *Remote Sens.*, vol. 6, no. 6, pp. 5795–5814, May 2014.

[34] X. Kang, S. Li, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification with edge-preserving filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2666–2677, May 2014.

[35] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 844–856, Feb. 2013.

[36] Y. Chen, N. Nasrabadi, and T. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, 2011.

[37] H. Zhang, J. Li, Y. Huang, and L. Zhang, "A nonlocal weighted joint sparse representation classification method for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Ses.*, vol. 7, no. 6, pp. 2057–2066, Jun. 2014.

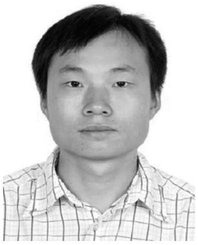
[38] A. Soltani-Farani and S. A. Hosseini, "Spatial-aware dictionary learning for hyperspectral image classificationg," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 527–541, Jan. 2015.

[39] Y. Chen, X. Zhao, and X. Jia, "Spectral-spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Ses.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[40] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Techn. Univ. Denmark, vol. 25, 2012.

[41] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[42] F. Cucker and D. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.



Jiangtao Peng received the B.S. and M.S. degrees from Hubei University, Wuhan, China, in 2005 and 2008, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, China, in 2011.

He is an Associate Professor in the Faculty of Mathematics and Statistics, Hubei University, China. His research interests include machine learning and hyperspectral image processing.

Hong Chen received the B.Sc. and Ph.D. degrees from Hubei University, Wuhan, China, in 2003 and 2009, respectively.

He is an Associate Professor with the Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Huazhong, China. His current research interests include learning theory, approximation theory, and machine learning.

Yicong Zhou (M'07–SM'14) received the B.S. degree from Hunan University, Changsha, China, in 1992, and the M.S. and Ph.D. degrees from Tufts University, Medford, MA, USA, in 2008 and 2010, respectively, all in electrical engineering.

He is currently an Associate Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include chaotic systems, multimedia security, image processing and understanding, and machine learning.

Dr. Zhou is an Associate Editor of the *Journal of Visual Communication and Image Representation* and a Leading Chair of IEEE SMC Technical Committee on Cognitive Computing. He received the third prize of Macau Natural Science Award in 2014.

Luoqing Li received the B.Sc. degree from Hubei University, Wuhan, China, the M.Sc. degree from Wuhan University, Wuhan, China, and the Ph.D. degree from Beijing Normal University, Beijing, China.

He is a Professor with the Faculty of Mathematics and Statistics, Hubei University, Wuhan, China.

Dr. Li is the Managing Editor of the *International Journal on Wavelets, Multiresolution, and Information Processing*.