A Nonlinear Multiple Feature Learning Classifier for Hyperspectral Images With Limited Training Samples

Jiayi Li, *Student Member, IEEE*, Hongyan Zhang, *Member, IEEE*, and Liangpei Zhang, *Senior Member, IEEE*

***Abstract*—A nonlinear joint collaborative representation (CR) model with adaptive weighted multiple feature learning to deal with the small sample set problem in hyperspectral image (HSI) classification is proposed. The proposed algorithm first maps every meaningful feature of the image scene into a kernel space by a column-generation (CG)-based technique. A unified multitask learning-based joint CR framework, with adaptive weighting for each feature, is then undertaken by the use of an alternating optimization algorithm, to obtain accurate kernel representation coefficients, which leads to desirable classification results. The experimental results indicate that the proposed algorithm obtains a competitive performance and outperforms the other state-of-the-art regression-based classifiers and the classical support vector machine classifier.**

***Index Terms*—Classification, collaborative representation (CR), hyperspectral image (HSI), Kernel method, small sample set.**

I. Introduction

HYPERSPECTRAL sensors, spanning the visible to the infrared spectrum, measure the reflected solar signal at hundreds (100 to 200+) of contiguous and narrow wavelength bands (bandwidth between 5 and 10 nm). Hyperspectral imaging can provide ample surface spectral information to identify and distinguish spectrally similar (but unique) materials [1]. Compared with multispectral imagery, the hundreds of bands with rich spectral information in hyperspectral images (HSIs) allow better discrimination of similar ground-cover classes, which can lead to a superior classification performance [2].

However, there are still some obstacles for the classification of pixels in HSIs. One of the difficult and complex problems is the lack of labeled training sets, as the low ratio between the limited available training samples and the large number of spectral bands [3] leads to a decrease in the discriminability and generalization ability. The classification methods originally developed for the labeling of low-dimensional datasets, i.e., multispectral images, generally perform poorly when applied to HSIs, particularly in the case of limited training samples. Furthermore, the unstable spectral signature of the classes in the spatial domain of the scene can lead to an incomplete description of the different ground-object classes. Considering the practical cost of a land survey, some training samples are closed in the spatial domain, and this kind of spatial autocorrelation intrinsically violates the independent identically distributed assumption of the samples. Due to the limited valuable training information, many statistical learning-based classifiers do not work well.

In order to deal with the above problems, various techniques have been proposed to improve the classification result in the small sample set case: 1) *dimensionality reduction*, which reduces the dimensionality of the hyperspectral data to the appropriate subspace without losing the original information that allows for the separation of classes [4]; 2) *active learning*, which builds an efficient training set by iteratively improving the model performance, has also been proposed for HSI classification [5]–[7]; 3) *semisupervised learning*, by jointly leveraging the labeled and unlabeled pixels in the hyperspectral scene, it enriches the available information to improve the classification accuracy [8], [9]; 4) *transfer learning*, which makes use of some existing meaningful and related labeled data from other scenes to enhance the model discrimination, has shown its potential in HSI classification [10]; 5) *kernel methods,* which implicitly mine the high-order discriminative information of the HSIs, have been widely used, and have performed well, due to their insensitivity to the curse of dimensionality [11], [12]; and 6) *the spatial prior knowledge of the HSI*, such as the texture structure [13], the neighborhood similarity [14], and so on, can complement the spectral information of the training samples for the classification.

Recently, Zhang *et al.* [15] proposed a novel linear collaborative representation (CR) approach to deal with the "lack of samples" problem for high-dimensional object recognition. In the CR-based classification procedure [16], the training samples which are located close to the unlabeled pixel contribute most to the representation of the unlabeled pixel, while the rest of the training samples act as collaborative assistants. It has been shown that the CR technique, which utilizes the entire set of training samples, often leads to high computational efficiency and state-of-the-art performance [17]–[19]. In our previous work, the CR technique, which linearly combines the multiple features via a multitask learning (MTL) approach to complement the class discriminability of each feature description, was

Manuscript received September 26, 2014; revised January 01, 2015; accepted February 03, 2015. Date of publication March 15, 2015; date of current version July 30, 2015. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB707105, and in part by the National Natural Science Foundation of China under Grant 61201342 and Grant 41431175. *(Corresponding author: Hongyan Zhang.)*

The authors are with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Collaborative Innovation Center for Geospatial Technology, Wuhan University, Wuhan, China (e-mail: zhanghongyan@whu.edu.cn).

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.2015.2400634

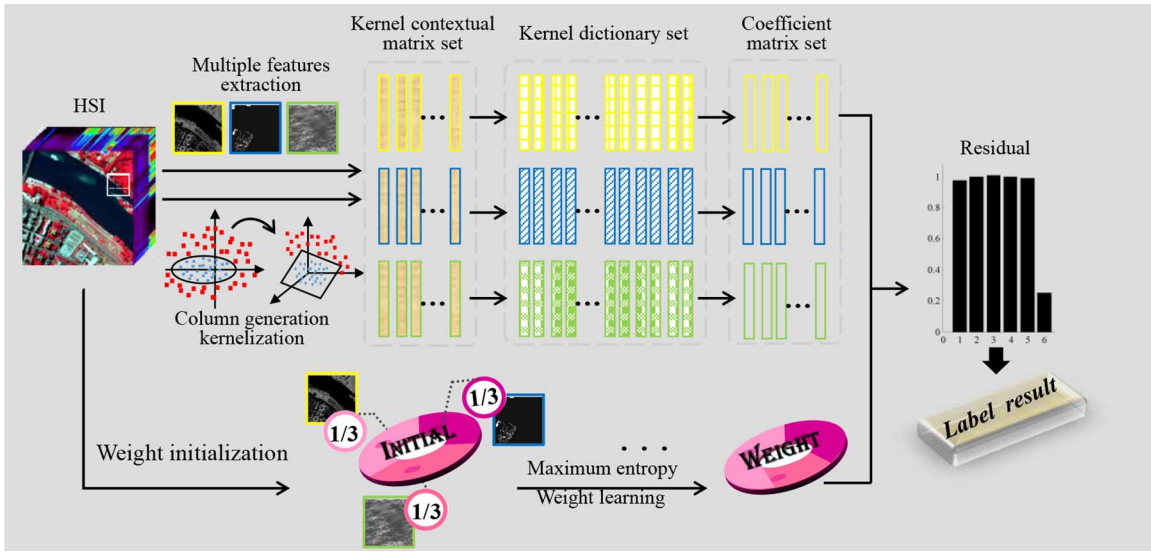


Fig. 1. Schematic illustration of the proposed joint CR classification algorithm with adaptive weighted multitask learning. Given a HSI, multiple modalities of the features are extracted. The multiple features are first constructed and then mapped into the CG kernel one by one. We then constructed the kernel dictionary set with regard to the training labels. In the representation step, each kernel contextual matrix is represented as a linear combination of the corresponding training kernel feature dictionary. To preserve the diversity, the different weights of the various features are also simultaneously estimated in the linear representation procedure. Finally, the classification decision is made according to the weighted overall reconstruction error of the individual class.

applied to HSI classification [20], [21]. Nevertheless, it has been noted that since hyperspectral datasets are not linearly separable [2], [11], a linear regression representation-based model may not cope well with such a nonlinear classification problem. In addition, it is natural that different features contribute different roles in the decision boundary construction, and even pixels in different locations should have specific multiple feature arrangements. That is, treating each feature equally, as in Li *et al.* [20], is less rational.

In this paper, we proposed a novel nonlinear kernel joint CR classification method with adaptive weighted MTL (KJCRC–AWMTL) for HSI via a column-generation (CG) kernel mapping technique [22], [23]. As in the visual illustration shown in Fig. 1, the procedure of the proposed algorithm can be summarized as follows. First, several meaningful features of the HSI are constructed, and each original feature is mapped to a new kernel space. Second, a multiple kernel feature dictionary set is constructed from the training sample set. Third, for each unlabeled pixel that is recognized, a kernel joint signal set containing multiple kernel features is constructed with the contextual information of the hyperspectral scene. Finally, after the kernel dictionary set and the kernel signal set are obtained, the multiple featured CR linear regression model is extended in an adaptive weighted manner to the novel kernel version, to obtain the coding coefficient set for the subsequent recognition. Unlike, the well-used kernel trick that calls for the explicit inner product structure of the solution [24], [25], the CG-strategy is easy to implement and does not require the explicit structure in the regression analytical solution. The proposed method deals with the nonlinear phenomenon of multiple features in HSI classification, and achieves an improved performance. Experiments with several HSI datasets that have been widely used as public evaluation data, confirmed the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows. Section II introduces the CR-based HSI classification method with the fixed weighted multiple feature learning approach. Section III maps the adaptive multiple feature learning framework into a nonlinear space, and proposes the corresponding kernel CR model via CG for HSI. The experimental results of the proposed algorithm with three hyperspectral datasets are given in Section IV. Finally, the conclusion is presented in Section V.

II. CR With Multiple Feature Learning

In this section, we first review the CR technique using only the spectral feature for HSI, and we then shows the linear joint CR-based classification method that integrates the multiple features with a MTL approach.

A. CR

Suppose there are M distinct classes for a HSI, and N_i $(i = 1, \ldots, M)$ training samples for each class. Every spectral signal in this scene can be denoted as a B-dimensional vector, where B refers to the number of bands of the HSI. In this way, training samples from the ith class act as columns of a subdictionary $\mathbf{A}_i = [\boldsymbol{a}_{i,1}, \ldots, \boldsymbol{a}_{i,N_i}] \in \mathbb{R}^{B \times N_i}$, which may not span all of the ith class feature space in the small sample set case, and an unlabeled pixel $\boldsymbol{s}_c$, denoted as feature $\boldsymbol{s} \in \mathbb{R}^B$, belonging to the ith class, can be written as a compact linear combination of the training samples from the ith class

$$\boldsymbol{s} = \boldsymbol{a}_{i,1}\boldsymbol{\alpha}_{i,1} + \cdots + \boldsymbol{a}_{i,N_i}\boldsymbol{\alpha}_{i,N} + \boldsymbol{\varepsilon} = \mathbf{A}_i\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{\varepsilon}$ is the random noise, and $\boldsymbol{\alpha}_i$ which is an unknown N_i-dimensional coefficient vector whose entries are the weights of the corresponding items of $\mathbf{A}_i$, can be solved by

$\boldsymbol{\alpha}_i = \mathbf{argmin}_{\boldsymbol{\alpha}_i} \|\boldsymbol{s} - \mathbf{A}\boldsymbol{\alpha}_i\|_F^2$ and utilized for the subsequent classification. For HSI classification with limited training samples, the quantity of the training samples belonging to the ith class is usually less than the spectral dimension, which makes the coefficient vector unstable and affects the recognition performance.

Instead of $\mathbf{A}_i$, Zhang *et al.* [15] utilized all the training samples to deal with the "lack of samples" problem, and they constructed an overcomplete collaborative dictionary $\mathbf{A} \in \mathbb{R}^{B\times N}$ by stacking all the subdictionaries $\{\mathbf{A}_m\}_{m=1,\ldots,M}$, where $N = \sum N_i$. In this way, the unlabeled ith class pixel $\boldsymbol{s}$ can be represented as a collaborative linear combination of all the training samples as

$$\begin{aligned}\boldsymbol{s} &= \mathbf{A}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} = \mathbf{A}_1\boldsymbol{\alpha}_1 + \cdots + \mathbf{A}_M\boldsymbol{\alpha}_M + \boldsymbol{\varepsilon} \\ &= \mathbf{A}_i\boldsymbol{\alpha}_i + \sum\nolimits_{j=1,j\neq i}^{M} \mathbf{A}_j\boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}\end{aligned} \tag{2}$$

where the whole space constitutes a dominant subspace spanned by $\mathbf{A}_i$, and the complementary subspace is spanned by the rest of the training samples, which can be considered as an external collaborative partner to the dominant subspace. $\boldsymbol{\alpha} \in \mathbb{R}^N$ is an unknown coefficient vector whose entries are the weights of the corresponding items of $\mathbf{A}$, and $\boldsymbol{\varepsilon}$ is the low-level random noise.

B. Joint CR Classification With Multiple Feature Learning

In this section, the joint CR model with multiple feature learning is introduced as an extended CR technique. The proposed model contains two improvements: simultaneous multiple feature learning, and joint multiple neighborhood pixel representation. We first present the multiple feature learning approach, and we then extend the method into a multiple signal framework.

It is widely acknowledged that to extract one optimal feature for all the classes is not realistic [26]. Rather than using a single feature to describe each class, Zhang *et al.* [15] proposed a method that combines multiple complementary features to describe the classes. In this way, the small sample set problem in HSI classification benefits from a more comprehensive description of each pixel, which is induced by the fusion of multiple features. As for the multiple feature case, suppose each pixel has K features, then $\boldsymbol{s}^k$ and $\mathbf{A}^k$ are the signal and dictionary of the kth feature, and $\boldsymbol{\alpha}^k$ is the coding vector of $\boldsymbol{s}^k$ over $\mathbf{A}^k$. An unlabeled pixel described by all the K features can be represented as

$$\begin{aligned}\boldsymbol{s}^1 &= \mathbf{A}_1^1\boldsymbol{\alpha}_1^1 + \cdots + \mathbf{A}_M^1\boldsymbol{\alpha}_M^1 + \boldsymbol{\varepsilon}^1 \\ &\vdots \qquad\qquad\qquad \vdots \\ \boldsymbol{s}^K &= \mathbf{A}_1^K\boldsymbol{\alpha}_1^K + \cdots + \mathbf{A}_M^K\boldsymbol{\alpha}_M^K + \boldsymbol{\varepsilon}^K.\end{aligned} \tag{3}$$

For the multiple signal case, it is assumed that HSI pixels in a small spatial neighborhood, which are highly correlated, can be simultaneously approximated by the common training pixels, while the training pixels are assigned in a different set of coefficients. Here, we utilize the neighboring pixels around the unlabeled pixel to make the representation more robust. We simultaneously stack all the pixels in the neighborhood patch centered at the hyperspectral pixel $\boldsymbol{s}_c$, and of size N_o, to construct the joint signal feature matrix set $\{\mathbf{S}^k\}_{k=1,\ldots,K} = \{\boldsymbol{s}_1^k\boldsymbol{s}_2^k\cdots\boldsymbol{s}_{N_o}^k\}_{k=1,\ldots,K}$, which contains K matrices sized $l^k \times N_o$ for each neighborhood patch, and can be denoted as

$$\begin{aligned}\mathbf{S}^1 &= \left[\boldsymbol{s}_1^1\boldsymbol{s}_2^1\cdots\boldsymbol{s}_{N_o}^1\right] = \sum\nolimits_{i=1}^{M} \mathbf{A}_i^1\boldsymbol{\Psi}_i^1 + \boldsymbol{\Sigma}^1 \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ \boldsymbol{S}^K &= \left[\boldsymbol{s}_1^K\boldsymbol{s}_2^K\cdots\boldsymbol{s}_{N_o}^K\right] = \sum\nolimits_{i=1}^{M} \mathbf{A}_i^K\boldsymbol{\Psi}_i^K + \boldsymbol{\Sigma}^K\end{aligned} \tag{4}$$

where $\boldsymbol{\Psi}^k$, $k = 1,\ldots,K$ is the set of the coding coefficient matrix associated with the corresponding dictionary $\mathbf{A}^k$, $\boldsymbol{\Psi}_i^k$ is the subset of $\boldsymbol{\Psi}^k$ over the subdictionary $\mathbf{A}_i^k$ $(i = 1,\ldots,M)$, and $\boldsymbol{\Sigma}^k$ is the random noise matrix set for the neighborhood patch for the kth feature $(k = 1,\ldots,K)$. Considering that different features can share some similarities as well as some differences, the prior for this issue can be modeled as

$$\sum\nolimits_{k=1}^{K} \omega^k \left\|\boldsymbol{\Psi}^k - \bar{\boldsymbol{\Psi}}^k\right\|_F^2 \tag{5}$$

where $\bar{\boldsymbol{\Psi}}$ is the mean of all $\boldsymbol{\Psi}^k$, and ω^k is the fixed weight for the kth feature. Given the training pixels, $\boldsymbol{\Psi}^{k(k=1,\ldots,K)}$ can be calculated by the joint collaborative model with multiple feature learning [20]

$$\hat{\boldsymbol{\Psi}}^k = \mathop{\mathbf{arg\,min}}_{\boldsymbol{\Psi}^k} \sum\nolimits_{k=1}^{K} \begin{pmatrix} \left\|\mathbf{S}^k - \mathbf{A}^k\boldsymbol{\Psi}^k\right\|_F^2 + \lambda\left\|\boldsymbol{\Psi}_\kappa^k\right\|_F^2 \\ +\tau\omega^k\left\|\boldsymbol{\Psi}_\kappa^k - \bar{\boldsymbol{\Psi}}_\kappa^k\right\|_F^2 \end{pmatrix}. \tag{6}$$

For $i = 1,\ldots,M$, the overall coding error for class i is shown as r_i

$$r_i = \sum\nolimits_{k=1}^{K} \omega^k \left\|\mathbf{S}^k - \mathbf{A}^k\hat{\boldsymbol{\Psi}}^k\right\|_F^2 \tag{7}$$

where $\hat{\boldsymbol{\Psi}}_i^k$ is the estimated regression matrix associated with feature k and class i. The label of the unlabeled pixel is then determined by the minimal total residual

$$\mathbf{class}\left(\boldsymbol{s}_c\right) = \mathop{\mathbf{arg\,min}}_{i=1,\ldots,M} \left\{r_i\right\}. \tag{8}$$

III. Kernel Joint CR Classification Via Adaptive Weighted Multiple Feature Learning

In this paper, two issues are taken into consideration. First, HSI classification is usually a linearly inseparable problem [2], and a higher feature space for mining the nonlinear separability is often beneficial. Second, since different features will have different levels of importance for building the overall coding error in (7), an equal weight for each feature is unreasonable, and a fixed weight with some predefined prior calls for expensive expert knowledge. In view of this, we first map the hyperspectral data into a kernel feature space via a CG technique [22] to make the problem linearly separable, and we then extend the multiple feature learning into an adaptive weighted framework.

A. Proposed CG Kernel Method

The kernel method, which is an approach to deal with nonlinear problems, is to assume that we have some way of measuring the similarity between pixels which does not require preprocessing them into a feature vector format [27]. In the proposed algorithm, we utilize the radial basis function (RBF) $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with the chi-squared distance $\chi^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$, which reflects the relative difference between each subregion of the original feature

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\chi^2(\boldsymbol{x}_i, \boldsymbol{x}_j)/\mu\right), \quad \text{where } \chi^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{2}\sum\nolimits_{b=1}^{B} \frac{(\boldsymbol{x}_i(b) - \boldsymbol{x}_j(b))^2}{\boldsymbol{x}_i(b) + \boldsymbol{x}_j(b)} \tag{9}$$

where $\boldsymbol{x}_i$ is the feature of a pixel at location i in a hyperspectral scene, and μ is set to the mean value of the pairwise chi-squared distance, and is adaptive to the training set.

The utilized CG in this paper directly takes the signal in the kernel space as a feature [23], which is similar to the simplified CG strategy for CG-boost in multiple kernel learning [22]. We denote $\boldsymbol{s}^k \in \mathbb{R}^{B_k}$ as the kth feature of the pixel, and $\boldsymbol{s}_\kappa^k \in \mathbb{R}^N$ as its representation in the kernel feature space. The CG kernel CR of $\boldsymbol{s}^k$ in terms of all the atoms of $\mathbf{A}^k$ can be formulated as

$$\begin{aligned}\boldsymbol{s}_k^\kappa &= \left[\kappa\left(\boldsymbol{a}_1^k, \boldsymbol{s}\right) \cdots \kappa\left(\boldsymbol{a}_N^k, \boldsymbol{s}\right)\right]^T = \mathbf{A}_\kappa^k \boldsymbol{\alpha}_\kappa^k \\ &= \underbrace{\begin{pmatrix} \kappa\left(\boldsymbol{a}_1^k, \boldsymbol{a}_1^k\right) & \cdots & \kappa\left(\boldsymbol{a}_1^k, \boldsymbol{a}_1^k\right) \\ & \ddots & \\ \kappa\left(\boldsymbol{a}_N^k, \boldsymbol{a}_1^k\right) & \cdots & \kappa\left(\boldsymbol{a}_N^k, \boldsymbol{a}_N^k\right) \end{pmatrix}}_{\mathbf{A}_\kappa^k} \underbrace{\left[\boldsymbol{\alpha}_{1,\kappa}^k \cdots \boldsymbol{\alpha}_{N,\kappa}^k\right]^T}_{\boldsymbol{\alpha}_\kappa^k}\end{aligned} \tag{10}$$

where the columns of $\mathbf{A}_\kappa^k$ are the representation of the training samples in the kernel feature space, and $\boldsymbol{\alpha}_\kappa^k$ is assumed to be an $N \times 1$ kernel representation vector.

B. Adaptive Weighted Multiple Feature Learning in Kernel Space

For a hyperspectral dataset, assuming that there is no prior for the different features, we regularize the different weights based on the maximum entropy principle [18]

$$-\sum\nolimits_{k=1}^{K} \omega^k \mathrm{In}\omega^k > \sigma. \tag{11}$$

With the kernel method, the adaptive weighted version of (6) can be extended to deal with such a kernel optimization problem as

$$\left\{\boldsymbol{\Psi}_\kappa^k, \omega^k\right\} = \underset{\boldsymbol{\Psi}_\kappa^k, \omega^k}{\mathbf{arg\,min}} \sum\nolimits_{k=1}^{K} \times \begin{pmatrix} \left\|\mathbf{S}_\kappa^k - \mathbf{A}_\kappa^k \boldsymbol{\Psi}_\kappa^k\right\|_F^2 + \lambda \left\|\boldsymbol{\Psi}_\kappa^k\right\|_F^2 \\ +\tau\omega^k \left\|\boldsymbol{\Psi}_\kappa^k - \bar{\boldsymbol{\Psi}}_\kappa^k\right\|_F^2 + \gamma\omega^k \mathbf{In}\omega^k \end{pmatrix} \tag{12}$$

where $\bar{\boldsymbol{\Psi}}_\kappa^k$ is the mean matrix of the kernel coefficient set $\boldsymbol{\Psi}_\kappa^k$, and λ and τ are two positive regularization parameters. ω^k and $\boldsymbol{\Psi}_\kappa^k$ can be solved by alternating the optimization of the objective function shown in (12) with the two corresponding subproblems, until the solutions converge to a local minimum.

For the first subproblem, we optimize $\boldsymbol{\Psi}_\kappa^k$ by fixing the weight ω^k, and the optimization of (12) becomes

$$\boldsymbol{\Psi}_\kappa^k = \underset{\boldsymbol{\Psi}_\kappa^k}{\mathbf{arg\,min}} \sum\nolimits_{k=1}^{K} \begin{pmatrix} \left\|\mathbf{S}_\kappa^k - \mathbf{A}_\kappa^k \boldsymbol{\Psi}_\kappa^k\right\|_F^2 + \lambda \left\|\boldsymbol{\Psi}_\kappa^k\right\|_F^2 \\ +\tau\omega^k \left\|\boldsymbol{\Psi}_\kappa^k - \bar{\boldsymbol{\Psi}}_\kappa^k\right\|_F^2 \end{pmatrix}. \tag{13}$$

We can obtain a closed-form solution for $k = 1, \ldots, K$

$$\boldsymbol{\Psi}_\kappa^k = \boldsymbol{\Psi}_\kappa^{0,k} + \tau \frac{\omega^k}{\sum_{\eta=1}^{K} \omega^\eta} \mathbf{P}^k \mathbf{Q} \sum\nolimits_{\eta=1}^{K} \omega^\eta \boldsymbol{\Psi}_\kappa^{0,k} \tag{14}$$

where $\mathbf{P}^k = \left(\left(\mathbf{A}_\kappa^k\right)^T \mathbf{A}_\kappa^k + \mathbf{I}\left(\lambda + \tau\omega^k\right)\right)^{-1}$, $\boldsymbol{\Psi}_\kappa^{0,k} = \mathbf{P}^k \left(\mathbf{A}_\kappa^k\right)^T \mathbf{S}_\kappa^k$, $\mathbf{Q} = \left(\mathbf{I} - \sum_{\eta=1}^{K} \varpi^\eta \mathbf{P}^\eta\right)^{-1}$, and $\varpi^\eta = \tau(\omega^\eta)^2 \Big/ \sum_{k=1}^{K} \omega^k$.

For the second subproblem, we optimize the weights ω^k by fixing the coefficient set $\boldsymbol{\Psi}_\kappa^k$, and the optimization of (12) becomes

$$\omega^k = \underset{\omega^k}{\mathbf{arg\,min}} \sum\nolimits_{k=1}^{K} \left(\tau\omega^k \left\|\boldsymbol{\Psi}_\kappa^k - \bar{\boldsymbol{\Psi}}_\kappa^k\right\|_F^2 + \gamma\omega^k \mathbf{In}\omega^k\right) \tag{15}$$

and the weight for each feature can be updated by

$$\omega^k = \mathbf{exp}\left\{-1 - \tau \left\|\boldsymbol{\Psi}_\kappa^k - \bar{\boldsymbol{\Psi}}_\kappa^k\right\|_F^2 \Big/ \gamma\right\}, \quad k = 1, \ldots, K \tag{16}$$

The formulation (12) is a nonconvex optimization problem, which cannot find the globally optimal solution, to the best of our knowledge. As both the subproblems are convex, the alternating optimization is considered to ensure the solutions converge to a local minimum [28].

C. Final Classification Scheme for HSI

For an unlabeled pixel, once $\hat{\boldsymbol{\Psi}}_\kappa^k$ and $\hat{\omega}^k$, $k = 1, \ldots, K$ are estimated by (12), then the label is determined by the minimal total residual

$$\mathbf{class}(\boldsymbol{s}_c) = \underset{i=1,\ldots,M}{\mathbf{argmin}} \sum\nolimits_{k=1}^{K} \hat{\omega}^k \left\|\mathbf{S}_\kappa^k - \mathbf{A}_\kappa^k \hat{\boldsymbol{\Psi}}_{i,\kappa}^k\right\|_F^2 \tag{17}$$

where $\hat{\boldsymbol{\Psi}}_{i,\kappa}^k$ is the kernel coefficient matrix associated with feature k and class i.

To sum up, the implementation details of the proposed algorithm for HSI classification are summarized in Algorithm 1.

D. Computational Complexity Analysis

The computational burden for the proposed algorithm is as follows. The size of $\mathbf{S}_\kappa^k$ is $N \times N_o$, and the size of the kernel dictionary $\mathbf{A}_\kappa^k$ is $N \times N$, where N is the size of the training sample set, and N_o is the number of pixels in the spatial window. Luckily, the kernel mapping does not affect the performance very much, as it can be computed offline. Supposing that q is the iteration number, then the whole computational

Fig. 2. False color image (R: 55, G: 35, and B: 25) of the Houston University image dataset.

complexity is $O\left(q\sum_{k=1}^{K}2\left(1+N_o\right)N^2\right)$, as $\mathbf{P}^k$ and $\mathbf{Q}$ are predefined in each iteration.

Algorithm 1. Kernel Joint CR Classification with Adaptive Weighted MultiTask Learning (KJCRC–AWMTL) for HSI.

Input: 1) An HSI containing training samples

2) Parameters: regulation parameters λ, τ, γ, spatial size N_o

Initialization:

1) Multiple feature extraction from the HSI

2) Construct the multiple feature dictionaries $\mathbf{A}^k$, $k=1,\ldots,K$ and normalize the columns of $\mathbf{A}_\kappa^k$ to have a unit ℓ_2-norm

3) Map the dictionary into the kernel feature space $\mathbf{A}_\kappa^k$

4) Treat each feature with an equal weight

Main iteration:

For each unlabeled pixel $\boldsymbol{s}_c$ in the HSI:

1) Map the features of $\boldsymbol{s}_c$ to the kernel feature space and construct the joint collaborative matrix $\mathbf{S}_\kappa^k$ in the feature space

2) Code $\mathbf{S}_\kappa^k$ over $\mathbf{A}_\kappa^k$ and obtain the coefficient matrices $\hat{\mathbf{\Psi}}_\kappa^k$, and adaptively weight $\hat{\omega}^k$ via (12), and $k=1,\ldots,K$

3) Label the test pixel $\boldsymbol{s}_c$ with (17)

End for

Output: A 2-D matrix which records the labels of the HSI

IV. Experiments

A. Dataset Description

1) Compact Airborne Spectrographic Imager (CASI) Dataset: Houston University Image: The first dataset used in this study, acquired over the University of Houston campus and its neighboring urban area, was distributed by the 2013 GRSS Data Fusion Contest. It contains 144 spectral bands in the 380–1050 nm region, and 349×1905 pixels with a spatial resolution of 2.5 m. Heavy shadows contained in the observed data were removed, and a subregion sized 349×1300 was retained for classification, as shown in Fig. 2. The reference information of the 15 classes was provided by the 2013 IEEE GRSS Data Fusion Contest for this subregion, and is shown in Table I. This is an urban dataset, with most of the land cover consisting of man-made objects. This dataset is challenging, since some of the classes, such as the three kinds of grasses, have quite similar spectral signatures.

2) Hyperion Dataset: Botswana Image: The second dataset was acquired by the NASA Earth Observing-1 satellite over the Okavango Delta, Botswana, on May 31, 2001. This dataset

TABLE I
15 Reference Classes of the CASI Houston University Image Dataset

Class no.	Class name	Class size	Class no.	Class name	Class size
1	Healthy grass	1073	9	Road	1031
2	Stressed grass	810	10	Highway	382
3	Synthetic grass	697	11	Railway	114
4	Trees	1053	12	Parking Lot 1	1233
5	Soil	1242	13	Parking Lot 2	449
6	Water	325	14	Tennis court	428
7	Residential	978	15	Running track	660
8	Commercial	624		Total	11 099

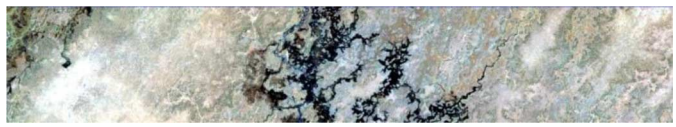


Fig. 3. False color image (R: 80, G: 60, and B: 30) of the Botswana dataset.

TABLE II
14 Reference Classes of the Botswana Hyperion Dataset

Class no.	Class name	Class size	Class no.	Class name	Class size
1	Water	270	8	Island interior	203
2	Hippo grass	101	9	Acacia woodlands	314
3	Floodplain grasses1	251	10	Acacia shrublands	248
4	Floodplain grasses2	215	11	Acacia grasslands	305
5	Reeds1	269	12	Short mopane	181
6	Riparian	269	13	Mixed mopane	268
7	Firescar2	259	14	Exposed soils	95
	Total			3248	

contains 242 spectral bands covering the 400–2500 nm portion of the spectrum in 10 nm windows, and it covers a 7.7-km strip at a 30-m spatial resolution. Uncalibrated and noisy bands that cover the water absorption features were removed, with 145 bands remaining. The size of the dataset is 256×1476, as shown in Fig. 3. We used 14 identified classes to reflect the impact of flooding on vegetation in the study area, and the class names and sizes are listed in Table II.

3) Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Dataset: Kennedy Space Center (KSC): This dataset, acquired over the KSC, Florida, on March 23, 1996, is in 224 spectral bands of 10 nm width, with a wavelength from 400 to 2500 nm, and covers 512×614 pixels, with a spatial resolution of 18 m, as shown in Fig. 4. Low-SNR and water absorption bands were removed, and 176 bands were used in this study. It has been noted that discrimination of land cover for this dataset is difficult, due to the similarity of the spectral signatures of certain vegetation types, and the complicated land cover distribution. The reference information for classification is shown in Table III.

B. Experimental Settings

For all the experiments with the three HSI datasets, limited training samples were randomly selected from the reference data, and the rest of the samples were set as test samples to evaluate the classification result. For classification with the Houston University image dataset, 10 and 15 training samples per class were, respectively, taken, while each case utilized 10 independent trials to alleviate any possible bias, which was also

Fig. 4. False color image (R: 57, G: 26, and B: 17) of the KSC dataset.

TABLE III
13 REFERENCE CLASSES OF THE AVIRIS KSC DATASET

Class no.	Class name	Class size
1	Scrub	761
2	Willow swamp	243
3	Cabbage palm hammock	256
4	Cabbage palm/oak hammock	252
5	Slash pine	161
6	Oak/broadleaf hammock	229
7	Hardwood swamp	105
8	Graminoid marsh	431
9	Spartina marsh	520
10	Cattail marsh	404
11	Salt marsh	419
12	Mud flats	503
13	Water	927
Total		5211

undertaken in the following experiments with the other two datasets. For the experiments with the KSC dataset, the two cases were five training pixels per class and 3% of the pixels of the whole reference label map. For the Botswana dataset, 3% and 5% of the reference data were, respectively, selected as training data. The training sample sizes of all the experiments were quite limited, which is a challenge for the following classification task.

In the quality evaluation tables, the overall accuracy (OA) is the ratio between the correctly classified test pixels and the total number of test samples. The kappa coefficient is a robust measure of the degree of agreement, and the classification accuracies using the different classifiers with the test set for each class can be found in the corresponding columns. The average running times of the repeated trials for each multiple feature related classifier were also recorded. The optimal parameter settings for every trial of each algorithm were acquired by tenfold cross-validation. The classification results were also averaged over 10 runs for each classifier to alleviate any possible bias induced by the random sampling. All of the experiments were carried out using MATLAB on a PC with one 3.50-GHz processer and 16.0 GB of RAM.

The benchmark algorithms are shown in Table IV. The abbreviations in Table IV can be explained as follows. For the SVM-based classifiers: SVMlinear denotes that the classifier is constructed as a linear version with a single feature; SVMrbf refers to the well-used kernel method calling for the explicit inner product with the RBF kernel; and SVM-CK

TABLE IV
COMPARISON OF THE CLASSIFICATION ALGORITHMS

Classifier	Linear/non linear	Multi/single feature	Contextual information
SVMlinear [2]	Linear	Single	No
SVMrbf [2]	Nonlinear	Single	No
SVMlinear-VS	Linear	Multiple	No
SVMrbf–VS	Nonlinear	Multiple	No
SVM-CK [30]	Nonlinear	Multiple	Yes
CRC [15]	Linear	Single	No
KCRC [11]	Nonlinear	Single	No
CRC-VS [20]	Linear	Multiple	No
KCRC-VS	Nonlinear	Multiple	No
JCRC-MTL [20]	Linear	Multiple	Yes
MNFL [21]	Nonlinear	Multiple	No
GCK-MLR [31]	Nonlinear	Multiple	No
KJCRC-AWMTL	Nonlinear	Multiple	Yes

means that a stack of the two spatial features is used to construct a composite kernel with the original spectral feature. For the CR-based classifiers, CRC denotes the classical linear method, and KCRC represents the method with a CG RBF kernel. For the single feature-based algorithms, each feature is applied to show the uneven discriminability. Moreover, vector stacking [29], which is shortened to "VS" in the postfix of the term in Table V, means that we directly stack all three features as an augmented one to simultaneously carry the multiple feature information. JCRC-MTL is a linear version of the proposed method, where the weight of each feature is fixed in advance [20].

Two recently published nonlinear multiple feature learning algorithms for HSI classification, generalized composite kernel-based multivariate logistic regression (GCK-MLR) [31] and multiple nonlinear feature learning with multivariate logistic regression (MNFL) [21], were also considered as benchmarks.

For the parameter settings, the weights for each feature were initially equal, and the regularization parameters λ, γ, and τ for the MTL-based classification algorithms were varied from 10^{-6} to 10^{-1}. The neighborhood sizes for KJCRC–AWMTL and JCRC-MTL were varied from 1 to 169. The optimal parameter settings for every trial of each algorithm were acquired by cross-validation.

C. Experimental Results

In Tables V–VII, the best results for each quality index are labeled in bold, and the suboptimal results for each quality index are underlined. It can be first seen that, compared with the linear classifiers, the corresponding nonlinear versions always obtain a better performance. For the single feature-based classifiers, it can be seen that different features lead to different classification results, and it is difficult to determine an "optimal" feature for the different datasets, which have different kinds of land cover. In view of this, combining the features in a uniform way is of interest. In these three tables, it is also notable that the multiple feature-based algorithms can indeed offer additional complementary information for the classification, as most of the classification results are significantly better than the single feature-based results. Since it is considered that CRC is comparable to SVM [15], KCRC is also superior to SVMrbf in most cases, as with the

TABLE V
CLASSIFICATION ACCURACY FOR THE HOUSTON UNIVERSITY IMAGE DATASET WITH THE TEST SET

	Linear		Training samples per class		Nonlinear		
	10	15			10	15	
SVMlinear	0.8196±0.0226 0.8041±0.0244	0.8470±0.0139 0.8336±0.0150	Spec	SVMrbf	0.8536±0.0182 0.8409±0.0197	0.8609±0.0094 0.8721±0.0101	Spec
	0.6612±0.0295 0.6324±0.0320	0.7076±0.0222 0.6825±0.0240	Gabor		0.6836±0.0235 0.6562±0.0252	0.7514±0.0219 0.7297±0.0238	Gabor
	0.7059±0.0303 0.6818±0.0325	0.7536±0.0268 0.7332±0.0287	DMP		0.7097±0.0228 0.6855±0.0244	0.7670±0.0223 0.7474±0.0237	DMP
CRC	0.7972±0.0200 0.7802±0.0216	0.8131±0.0154 0.7973±0.0166	Spec	KCRC	0.8566±0.0116 0.8444±0.0124	0.8769±0.0089 0.8664±0.0096	Spec
	0.6503±0.0174 0.6209±0.0184	0.7159±0.0257 0.6919±0.0273	Gabor		0.7195±0.0280 0.6955±0.0297	0.7836±0.0154 0.7649±0.0166	Gabor
	0.7276±0.0232 0.7052±0.0249	0.7887±0.0298 0.7708±0.0318	DMP		0.7741±0.0339 0.7554±0.0362	0.8394±0.0099 0.8257±0.0106	DMP
SVMlinear -VS	0.8061±0.0193 0.7894±0.0211 0.5777	0.8564±0.0172 0.8440±0.0185 0.7569	Multi	SVMrbf-VS	0.7970±0.0222 0.7794±0.0242 0.6062	0.8562±0.0175 0.8436±0.0189 0.8404	Multi
CRC-VS	0.7819±0.0191 0.7634±0.0205 2.9296	0.8441±0.0170 0.8307±0.0182 3.2758		KCRC-VS	0.8608±0.0229 0.8487±0.0248 10.6752	0.9065±0.0137 0.8984±0.0148 15.5541	
The third line of the terms for the multiple feature related classifiers records the average running time for the classification.				SVM-CK	**0.9252±0.0140** **0.9186±0.0152** 9.5649	**0.9424±0.0097** **0.9373±0.0105** 12.5381	
				MNFL	0.7759±0.0356 0.7568±0.0385 0.3286	0.8587±0.0254 0.8463±0.0275 0.3582	
				GCK-MLR	0.8885±0.0267 0.8788±0.0289 1.8074	0.9202±0.0155 0.9132±0.0169 2.5474	
JCRC-MTL	0.8187±0.0126 0.8032±0.0137 79.5000	0.8649±0.0212 0.8530±0.0229 195.6667		KJCRC-AWMTL	0.9112±0.0113 0.9035±0.0123 1277.9196	0.9332±0.0108 0.9273±0.0116 1504.8502	

multiple feature stacking classifiers, which also verifies the superiority of the CG kernel. In the multiple feature cases, the proposed KJCRC–AWMTL method is much better than the linear JCRC-MTL, and is only slightly inferior to SVM-CK for the Houston University image dataset. For the other two datasets, the CR-based classifiers are more suitable, and the proposed method gives the best results, which are slightly superior to the results of JCRC-MTL and greatly superior to the results of SVM-CK. In view of this, it can be concluded that the MTL method is the best multiple feature combination approach when compared with the VS and CK/GCK strategies. Furthermore, the CR method is comparable to SVM, and is better than MLR. For the number of training samples issue, the classification results of all the methods improve with the growth in the training sample set, and the multiple feature-based classifiers, especially the proposed KJCRC–AWMTL, can obtain a more desirable performance. It can also be observed that the variations in the results of each classifier in Tables V–VII are large, as the limited training samples seriously affect the performance. Under the small sample set condition, it can be seen that the proposed method can obtain a more stable result. All in all, the proposed kernel multiple feature method can provide a more stable and competitive classification result.

For the running time comparisons, the detailed average running times for every multiple feature related classifier are shown in the third line of the terms in the classification accuracy table. Here, it can be seen that the linear version is faster than the associated kernel method, but, at the same time, the discrimination is inferior. Although the running time of MNFL seems to be optimal, its classification performance is inferior to the other kernel-based classifiers. Comparing the MLR, SVM, and CR methods, it can be observed that MLR is the fastest method, but has the worst classification result, and CR shows better discrimination but requires more running time than SVM. Meanwhile, it should be noted that SVM was implemented by the LibSVM [45] package, which utilizes C++ software to speed it up. Comparing the multiple feature combination approaches (i.e., VS, CK/GCK, and MTL), the VS approach has the fastest speed and the worst discrimination, and MTL obtains the best classification result, but has a burdensome computation cost. Comparing the running time ratios $\frac{time_{\text{KCRC-VS}}/time_{\text{CRC-VS}}}{time_{\text{JCRC-MTL}}/time_{\text{KJCRC-AWMTL}}}$, it can be concluded that the major computational burden of the proposed approach comes from the adaptive weight estimation by several iterations, which causes several matrix inverse calculations. It is, however, reasonable to believe that with the rapid development in computer hardware, the time cost of the proposed method will soon no longer be an issue.

D. Parameter Analysis

In this part, we examine the effect of the parameters on the classification performance of the proposed algorithm in the aforementioned experiments. The experiments for λ, τ, γ, and N_o were repeated 10 times, using different randomly chosen training sets, to alleviate any possible bias induced by the random sampling. When analyzing one specific parameter, we fixed the other parameters as the corresponding optimal. The horizontal axis shown in Fig. 5 is the value range of the

TABLE VI
CLASSIFICATION ACCURACY FOR THE BOTSWANA IMAGE WITH THE TEST SET

	Linear				Nonlinear		
	3%	5%	Training samples percentage		3%	5%	
SVMlinear	0.8579±0.0178 0.8460±0.0194	0.8880±0.0137 0.8786±0.0148	Spec	SVMrbf	0.8605±0.0206 0.8489±0.0223	0.8899±0.0122 0.8807±0.0132	Spec
	0.7191±0.0260 0.6953±0.0283	0.8183±0.0123 0.8029±0.0134	Gabor		0.7560±0.0183 0.7351±0.0199	0.8526±0.0202 0.8401±0.0220	Gabor
	0.8784±0.0257 0.8683±0.0279	0.9178±0.0192 0.9110±0.0208	DMP		0.8677±0.0165 0.8566±0.0179	0.9134±0.0159 0.9062±0.0172	DMP
CRC	0.7861±0.0189 0.7678±0.0207	0.8144±0.0172 0.7985±0.0188	Spec	KCRC	0.8861±0.0093 0.8765±0.0100	0.9149±0.0101 0.9078±0.0110	Spec
	0.7600±0.0221 0.7394±0.0240	0.8587±0.0163 0.8466±0.0177	Gabor		0.8222±0.0176 0.8071±0.0192	0.9015±0.0158 0.8931±0.0172	Gabor
	0.8849±0.0199 0.8753±0.0232	0.9137±0.0197 0.9065±0.0213	DMP		0.9179±0.0126 0.9111±0.0136	0.9411±0.0162 0.9361±0.0175	DMP
SVMlinear -VS	0.9101±0.0147 0.9026±0.0159 0.1247	0.9369±0.0165 0.9316±0.0178 0.1778	Multi	SVMrbf-VS	0.9049±0.0118 0.8970±0.0127 0.1286	0.9320±0.0130 0.9263±0.0141 0.2026	Multi
CRC-VS	0.9245±0.0155 0.9182±0.0168 0.8899	0.9576±0.0124 0.9540±0.0134 0.9897		KCRC-VS	0.9546±0.0144 0.9508±0.0133 3.4327	0.9713±0.0108 0.9689±0.0117 5.3793	
The third line of the terms for the multiple feature related classifiers records the average running time for the classification.				SVM-CK	0.9448±0.0095 0.9402±0.0103 5.0525	0.9657±0.0059 0.9629±0.0063 7.0652	
				MNFL	0.9171±0.0136 0.9102±0.0148 0.2351	0.9521±0.0092 09481±0.0100 0.2769	
				GCK-MLR	0.9004±0.0187 0.8922±0.0200 0.1978	0.9033±0.0293 0.8954±0.0315 0.4190	
JCRC-MTL	0.9523±0.0186 0.9483±0.0202 11.8636	0.9801±0.0059 0.9784±0.0064 18.2716		KJCRC-AWMTL	**0.9747±0.0107** **0.9725±0.0116** 147.4000	**0.9850±0.0081** **0.9837±0.0087** 300.9000	

parameter being analyzed, and the vertical axis shows the OA of the different classifiers.

In Fig. 5(a), it can be seen that except for "KSC 5," all the cases first improve as the regularization parameter λ increases, and then begin to decrease slightly after the maximum value. With the growth in the number of training samples, the variations decrease, in all cases, which suggest the importance of the training samples. The regularization parameter λ makes a tradeoff between the data fidelity term and the prior term of the coefficient matrix for every feature, and it contributes to the objective function with regard to the value of λ. It can be seen that this regularization term can indeed improve the classification result when λ is in a reasonable range.

For parameter τ, which was varied from 10^{-8} to 1, once τ exceeds a certain threshold, the dominant part of the multitask representation optimization can be denoted as $\omega^k \left\| \mathbf{\Psi}^k - \bar{\mathbf{\Psi}} \right\|_F^2$, which has a poor discriminative power for the subsequent classification, as Fig. 5(b) shows.

The performances associated with γ are shown in Fig. 5(c). Here, it can be observed that the proposed kernel method shows a weaker capability with a small γ at first, then increases rapidly to reach the optimal result, and finally decreases a little. It is believed that this parameter affects the weight update procedure, and the effect of the adaptive weighting is shown in Fig. 6.

Finally, we varied N_o from 9 to 169 to investigate the effect of the neighborhood size. From Fig. 5(d), it can be seen that the optimal values for all the datasets are not large, as this approach can be considered as a straightforward spatial smoothing procedure. However, we believe that a more effective way to adaptively utilize the contextual prior should be considered in future work.

E. Weight Analysis

The effect of the adaptive weighting for each hyperspectral dataset in each training sample set case is shown in Fig. 6. Here, ten independent repeated trials were undertaken for each case, associated with the parameter set cross-validations and the classification results in Tables V–VII. In Fig. 6, the horizontal axis is the class name, and the vertical axis shows the statistical weight value over the different classes. All the experiments were initialized with equal weights, and the most intuitive issue from Fig. 6 is that the different features should indeed be weighted differently, as the weights in most cases change a lot in the adaptive weight update mechanism. It can also be seen that the variations in all the classification results are large, as each feature in a specific area is influenced by the complicated surrounding scene. For the Houston University image dataset, it can be observed that the spectral feature is the most discriminative for all the classes, while the performance (except for pixels in the running track class) suggests that the DMP feature is inferior. For the other two datasets, it is demonstrated that the latter two spatial features play more important roles, while the spectral feature is inferior. It is notable that although the Gabor feature shows the worst discriminability in the single feature-based classifiers, its role in the proposed

TABLE VII
CLASSIFICATION ACCURACY FOR THE KSC IMAGE WITH THE TEST SET

	Linear				Nonlinear		
	5	3%	Training samples scale		5	3%	
SVMlinear	0.7893±0.0254 0.7654±0.0280	0.8414±0.0166 0.8229±0.0186	Spec	SVMrbf	0.7884±0.0298 0.7646±0.0331	0.8510±0.0113 0.8337±0.0125	Spec
	0.6365±0.0371 0.5986±0.0408	0.8254±0.0231 0.8052±0.0263	Gabor		0.6706±0.0351 0.6356±0.0387	0.8639±0.0218 0.8479±0.0245	Gabor
	0.8183±0.0395 0.7984±0.0434	0.9082±0.0148 0.8977±0.0165	DMP		0.7947±0.0486 0.7673±0.0532	0.9031±0.0159 0.8921±0.0177	DMP
CRC	0.7949±0.0162 0.7714±0.0178	0.8009±0.0084 0.7766±0.0094	Spec	KCRC	0.8108±0.0209 0.7892±0.0229	0.8423±0.0092 0.8234±0.0104	Spec
	0.7006±0.0231 0.6688±0.0256	0.8684±0.0188 0.8527±0.0213	Gabor		0.7333±0.0370 0.7050±0.0406	0.9051±0.0169 0.8942±0.0189	Gabor
	0.8306±0.0271 0.8121±0.0295	0.9100±0.0149 0.8997±0.0166	DMP		0.8524±0.0309 0.8366±0.0338	0.9336±0.0113 0.9260±0.0126	DMP
SVMlinear -VS	0.8404±0.0294 0.8226±0.0323 0.1492	0.9375±0.0146 0.9304±0.0162 0.3003	Multi	SVMrbf-VS	0.8042±0.0311 0.7827±0.0344 0.1558	0.9265±0.0157 0.9181±0.0175 0.3367	Multi
CRC-VS	0.8502±0.0182 0.8335±0.0234 1.3126	0.9460±0.0147 0.9398±0.0165 1.5175		KCRC-VS	0.8857±0.0153 0.8730±0.0170 3.4487	0.9557±0.0105 0.9507±0.0117 8.0902	
The third line of the terms for the multiple feature related classifiers records the average running time for the classification.				SVM-CK	0.8797±0.0112 0.8731±0.0124 3.9813	0.9526±0.0110 0.9472±0.0123 5.7456	
				MNFL	0.8007±0.0310 0.7788±0.0342 0.2655	0.9409±0.0170 0.9431±0.0189 0.2950	
				GCK-MLR	0.8315±0.0304 0.8126±0.0336 0.1544	0.9133±0.0220 0.9032±0.0240 0.3347	
JCRC-MTL	0.9231±0.0373 0.9144±0.0416 19.0298	0.9705±0.0166 0.9671±0.0186 33.3043		KJCRC-AWMTL	**0.9422±0.0184** **0.9357±0.0295** 175.1112	**0.9782±0.0087** **0.9757±0.0097** 846.9000	

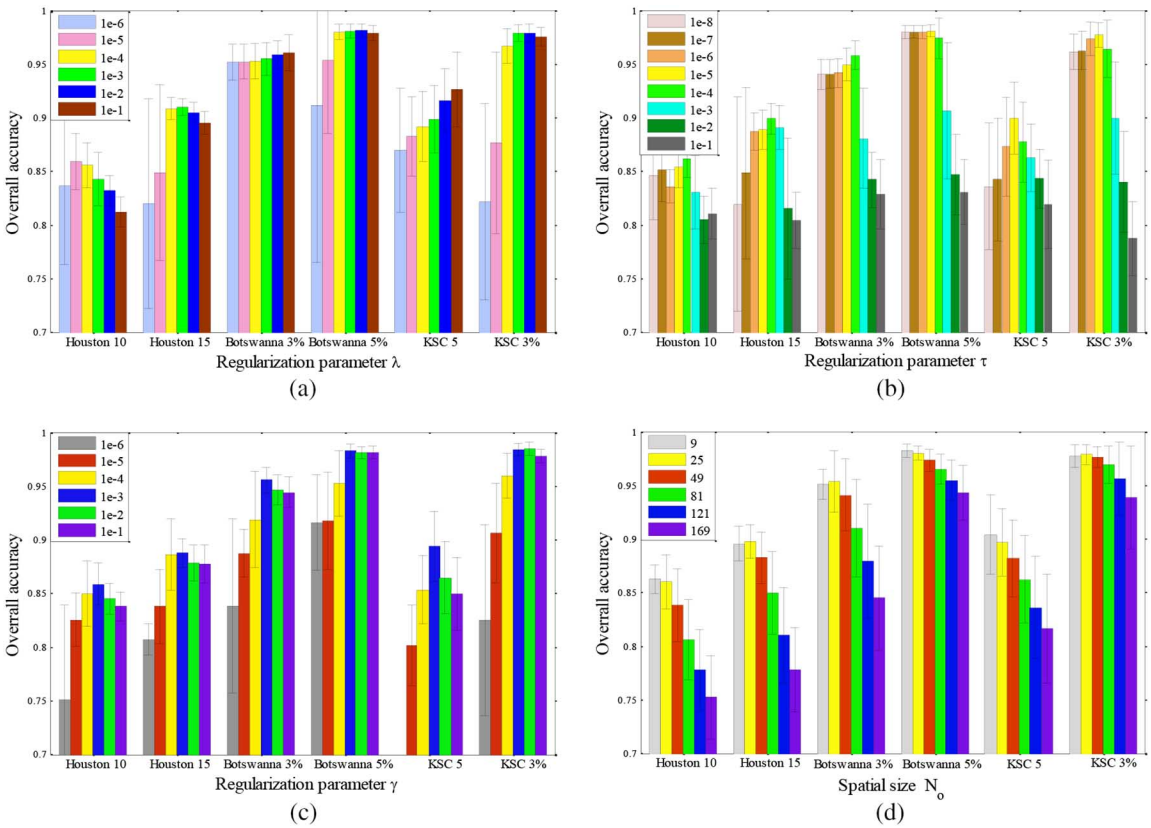


Fig. 5. Classification accuracy versus parameters for KJCRC–AWMTL: (a) regularization parameter λ; (b) regularization parameter τ; (c) regularization parameter γ; and (d) size of the spatial neighborhood window N_o. In this figure, the abbreviations "Houston 10" and "Houston 15" refer to the experiments with 10/15 training samples per class with the Houston University image dataset; "Botswana 3%" and "Botswana 5%" denote the experiments with 3% and 5% data of the whole reference data, respectively, with the Botswana dataset; and the last two terms "KSC 5" and "KSC 3%" are associated with the experiments with the KSC dataset.

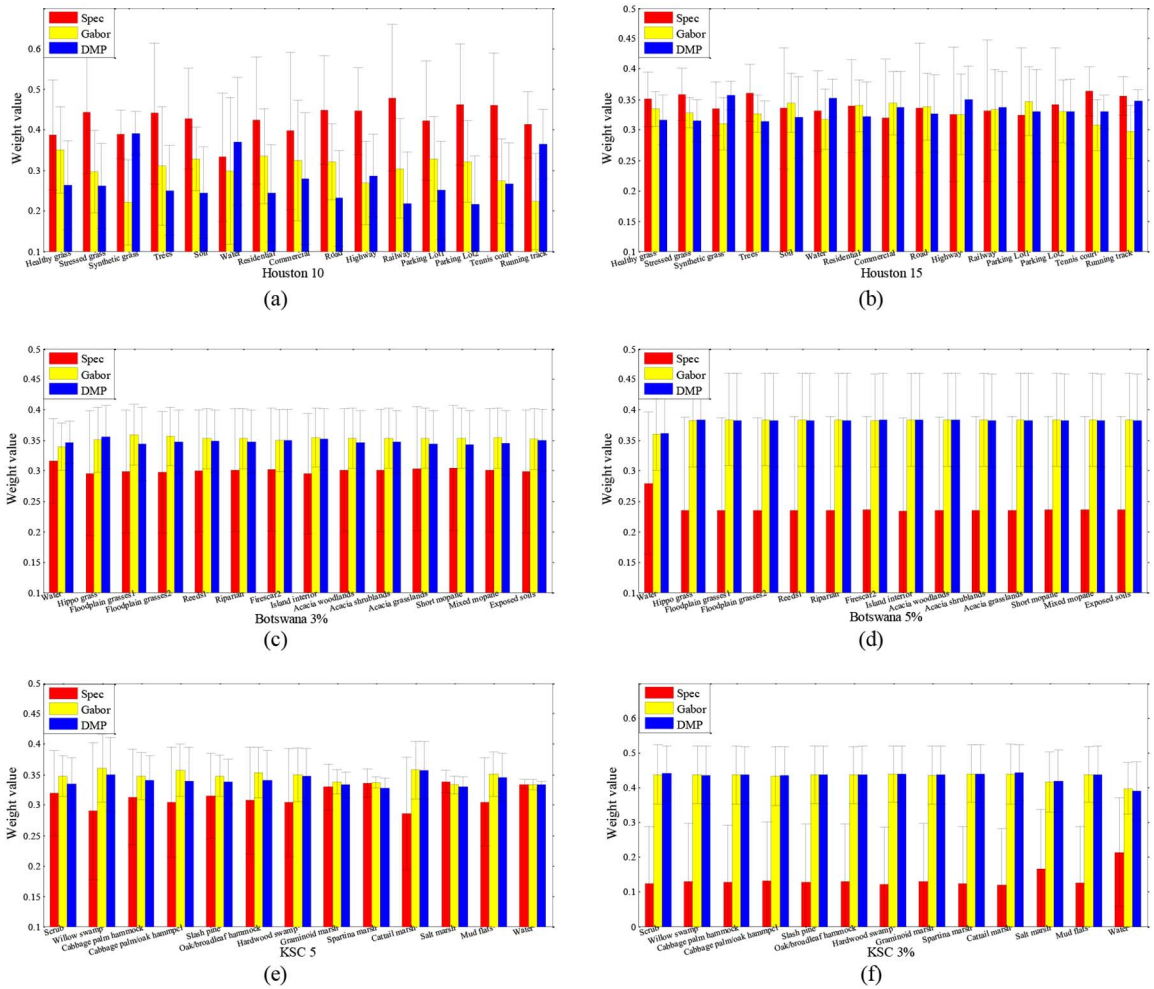


Fig. 6. Effect of ω in all the experiments: (a) Houston University image dataset experiment with 10 training samples per class; (b) Houston University image dataset experiment with 15 training samples per class; (c) Botswana dataset experiment with 3% reference data as the training samples; (d) Botswana dataset experiment with 3% reference data as the training samples; (e) KSC dataset experiment with five training samples per class; and (f) KSC dataset experiment with 3% reference data as the training samples.

multiple feature framework is enhanced. In addition, it is also demonstrated that, under the proposed weight update framework, the weights for the spectral feature and DMP feature are consistent with their performances in the single feature-based classifiers, which validates the effectiveness of this approach.

V. Conclusion

In this paper, we have focused on the linearly inseparable problem of HSI classification, and the different contributions of multiple features in HSI classification, and we have applied a CG kernel method to the adaptive weighted multiple feature learning framework to deal with these issues. The contributions of the proposed algorithm are as follows: 1) it not only maintains the complementary information of multiple meaningful features, but also arranges them in a rational way; 2) it keeps the smoothness of the spatial constraint; and 3) it maps each feature of the original signal into a high-dimensional space. The CG kernel technique, which directly treats the similarity measures between spectral pixels as a feature, shows an efficient performance in the multiple feature learning procedure. The proposed algorithm was tested on CASI, Hyperion, and AVIRIS HSIs, and the extensive experimental results confirmed the effectiveness of this nonlinear technique.

Acknowledgment

The authors would like to thank the committee of IEEE GRSS Data Fusion Contest for providing the CASI image of Houston University dataset. Thanks also to the handling editor and anonymous reviewers for their careful reading and helpful remarks.

References

[1] H. Zhang, L. Zhang, and H. Shen, "A super-resolution reconstruction algorithm for hyperspectral images," *Signal Process.*, vol. 92, no. 9, pp. 2082–2096, 2012.

[2] 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.

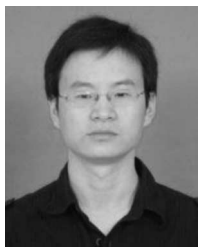
[3] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

[4] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina, "Exploiting manifold geometry in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 441–454, Mar. 2005.
[5] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1014–1031, Mar. 2011.
[6] D. Tuia *et al.*, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 606–617, Jun. 2011.
[7] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3947–3960, Oct. 2011.
[8] A. Plaza *et al.*, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, 2009.
[9] Q. Shi, L. Zhang, and B. Du, "Semi-supervised discriminative locally enhanced alignment for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4800–4815, Sep. 2013.
[10] S. Rajan, J. Ghosh, and M. M. Crawford, "Exploiting class hierarchies for knowledge transfer in hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3408–3417, Nov. 2006.
[11] J. Li, H. Zhang, and L. Zhang, "Column-generation kernel nonlocal joint collaborative representation for hyperspectral image classification," *ISPRS J. Photogramm.*, vol. 94, no. 4, pp. 25–36, 2014.
[12] B. Schölkopf, K. Tsuda, and J. Vert, "A primer on kernel methods," *Kernel Methods in Computational Biology*, 2004, pp. 35–70.
[13] Y. Qian, M. Ye, and J. Zhou, "Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2276–2291, Sep. 2012.
[14] H. Zhang *et al.*, "A nonlocal weighted joint sparse representation classification method for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2056–2065, Jun. 2014.
[15] L. Zhang *et al.*, "Collaborative representation based classification for face recognition," Arxiv preprint arXiv:1204.2358, 2012.
[16] J. Wright *et al.*, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[17] P. Zhu *et al.*, "Multi-scale patch based collaborative representation for face recognition with margin distribution optimization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 822–835.
[18] M. Yang *et al.*, "Relaxed collaborative representation for pattern classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 2224–2231.
[19] J. Li *et al.*, "Hyperspectral image classification by nonlocal joint collaborative representation with a locally adaptive dictionary," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3707–3719, Jun. 2014.
[20] J. Li *et al.*, "Joint collaborative representation with multitask learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5923–5936, Sep. 2014.
[21] J. Li *et al.*, "Multiple feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1592–1606, Aug. 2014.
[22] J. Bi, T. Zhang, and K. P. Bennett, "Column-generation boosting methods for mixture of kernels," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2004, vol. 22, pp. 521–526.
[23] X. Yuan, X. Liu, and S. Yan, "Visual classification with multi-task joint sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4349–4360, Oct. 2012.
[24] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification via kernel sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 217–231, Jan. 2013.
[25] H. Kwon and N. M. Nasrabadi, "Kernel matched subspace detectors for hyperspectral target detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 178–194, Feb. 2006.
[26] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 221–228.
[27] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
[28] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, 2001.
[29] X. Huang and L. Zhang, "An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 257–272, Jan. 2013.
[30] G. Camps-Valls *et al.*, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.
[31] J. Li *et al.*, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.

Jiayi Li (S'13) received the B.S. degree in geomatics engineering from Central South University, Changsha, China, in 2011. She is currently pursuing the Ph.D. degree at the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

Her research interests include hyperspectral imagery, sparse representation, computer vision, and pattern recognition in remote sensing images.

Dr. Li is a Reviewer of more than five international journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the IEEE SIGNAL PROCESSING LETTER, and *International Journal of Remote Sensing*.



Hongyan Zhang (M'13) received the B.S. degree in geographic information system and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2005 and 2010, respectively.

He is currently an Associate Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is a Reviewer of about 10 international academic journals, and has published more than 30 research papers. His research interests include image reconstruction, hyperspectral image processing, sparse representation, and low rank methods for sensing image imagery.

Liangpei Zhang (M'06–SM'08) received the B.S. degree in physics from Hunan Normal University, ChangSha, China, in 1982, the M.S. degree in optics from Xi'an Institute of Optics and Precision Mechanics of Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is currently the Head of the Division of Remote Sensing, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is also a Chang-Jiang Scholar Chair Professor appointed by the Ministry of Education of China. He is currently a Principal Scientist for the China State Key Basic Research Project (2011–2016) appointed by the Ministry of National Science and Technology of China to lead the remote sensing program in China. He has more than 360 research papers. He is the holder of 15 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a Fellow of the Institution of Engineering and Technology, an Executive Member (Board of Governor) of the China National Committee of the International Geosphere-Biosphere Programme, and an Executive Member of the China Society of Image and Graphics. He regularly serves as a Co-Chair of the series SPIE Conferences on Multispectral Image Processing and Pattern Recognition, Conference on Asia Remote Sensing, and many other conferences. He edits several conference proceedings, issues, and geoinformatics symposiums. He also serves as an Associate Editor of the *International Journal of Ambient Computing and Intelligence*, the *International Journal of Image and Graphics*, the *International Journal of Digital Multimedia Broadcasting*, the *Journal of Geo-spatial Information Science*, the *Journal of Remote Sensing*, and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.