Efficient Superpixel-Level Multitask Joint Sparse Representation for Hyperspectral Image Classification

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

***Abstract*—In this paper, we propose a superpixel-level sparse representation classification framework with multitask learning for hyperspectral imagery. The proposed algorithm exploits the class-level sparsity prior for multiple-feature fusion, and the correlation and distinctiveness of pixels in a spatial local region. Compared with some of the state-of-the-art hyperspectral classifiers, the superiority of the multiple-feature combination, the spatial prior utilization, and the computational complexity are maintained at the same time in the proposed method. The proposed classification algorithm was tested on three hyperspectral images. The experimental results suggest that the proposed algorithm performs better than the other sparse (collaborative) representation-based algorithms and some popular hyperspectral multiple-feature classifiers.**

***Index Terms*—Classification, hyperspectral imagery, multitask learning, sparse representation.**

I. Introduction

HYPERSPECTRAL images (HSIs), spanning the visible to infrared spectrum with hundreds of contiguous and narrow spectral bands, are favored by subtle discriminative spectral characteristics, as well as spatial information [1]. Due to the fine spectral differences between various materials of interest, HSIs support improved target detection and classification capabilities, and have important applications in various fields, such as the military [2], precision agriculture [3], [4], and mineralogy [5]. Supervised hyperspectral classification that labels each pixel in the image, which is determined by a given training sample set from each class, is an important task for the subsequent processing and analysis. The framework of supervised classification consists of two procedures: 1) discriminative feature acquisition and 2) classifier design.

1) To obtain the meaningful features for classification, techniques in both the spectral and spatial domain have been studied in recent years. Techniques, such as spectral band selection based on trivariate mutual information and clonal selection [6], particle swarm optimization [7], dimensional reduction based on sparse graph discriminant analysis [8], independent component analysis [9], discriminative metric learning [10], linear spectral unmixing, and manifold learning [11], have been widely utilized to fully exploit the spectral discriminability. On the other hand, the spatial feature construction, which was first studied in high spatial resolution processing and analysis [12], [13], has also been utilized in HSI classification, in techniques such as the 3-D wavelet feature [14], the gray-level cooccurrence matrix [13], and morphological analysis [15]. Furthermore, some multiple-feature fusion approaches with state-of-the-art performance have also been proposed to combine the complementary features and enhance the discriminability [16]–[18].
2) The lack of samples [19], the Hughes phenomenon [20], and the high computational burden caused by high-dimensional signal processing [21] are inevitable obstacles for designing HSI classifiers. One of the most widely used discriminative classifiers is the support vector machine (SVM) [20], [22], which aims to find an optimal separating hyperplane between two classes to solve the binary classification problem. Considering the hyperspectral nonlinear separability and the spectral–spatial prior, some variations of SVM, such as relevance vector machines [23] and semisupervised transductive SVM [24], have also been introduced to solve the hyperspectral classification task. Another discriminative hyperspectral classifier is multinomial logistic regression (MLR) [25], [26], which predicts the probabilities of the different possible outcomes of categorically distributed dependent samples, given a set of independent training samples. Some useful priors, such as composite kernels [27]–[29], sparsity regularization [30], [31], and spatial constraint by Markov random fields [25], [32], have also been utilized in these kinds of classifiers.

A classifier named sparse representation classification (SRC) [33] has attracted great attention for supervised HSI classification [34] in recent years. Based on a collaborative representation (CR) [35] mechanism, where the training samples that 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, SRC can deal with the "lack of samples" problem [21] for high-dimensional

Manuscript received November 4, 2014; revised February 20, 2015; accepted April 6, 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.

The authors are with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing and the Collaborative Innovation Center for Geospatial Technology, Wuhan University, Wuhan 430079, 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/TGRS.2015.2421638

signal recognition. Further studies in sparse representation [14], [19], [21], [36]–[39] have also been implemented and have shown excellent performances. Zhang *et al.* [36] extended the joint sparsity model in [40] with a nonlocal spatial prior. A CR mechanism and a local constraint were emphasized in [19], [41], which are believed to be more useful than a sparse constraint with a general dictionary. Zhang *et al.* [38] also utilized CR for image-set-based face recognition, by modeling the test query images and the whole training dictionary as a convex hull to take the set-to-set CR, and labeling the coefficient vector according to the similar decision rule in the original SRC. By not only exploiting the correlation and distinctiveness of the sample images in the query set but also reducing the computational burden caused by the redundancies in the query set, this novel CR-based framework has shown a desirable recognition performance.

For HSIs, it is natural that spatial similarity and redundancy exist in the scene, which can link the HSI classification with the aforementioned novel set-to-set CR-based work. In this paper, we propose a superpixel-level multitask joint SRC (SMTJSRC) algorithm, which exploits the joint sparsity prior for the multiple-feature fusion, and the correlation and distinctiveness of pixels in a spatial local region for the HSI classification. The algorithm is implemented in the following steps. First, several complementary features of the HSI are extracted. Second, superpixels, which can be regarded as small local regions with an adaptive shape and size [42], [43], are partitioned from the whole scene. Third, the multiple-feature joint sparse linear regression model is extended in a set-to-set CR manner to obtain the coding coefficient vector for the subsequent classification. The proposed method aims to enhance the discrimination of pixels by combining the complementary information of different features and highlighting the major patterns of pixels within a spatial region in a multitask learning fusion to achieve an improved classification result. Experiments with several HSIs confirm the effectiveness of the proposed SMTJSRC algorithm.

We next discuss the relationship between the proposed SMTJSRC classification method and the other relevant CR-based methods. For the multiple-feature-based approaches, Li *et al.* [21] investigated the distinctiveness and similarity of the different features in a model with two ℓ_2-norm-induced constraints, whereas Yuan *et al.* [39] exploited the invariant similarities between various features with an $\ell_{1,2}$-norm induced by a sparse constraint. Although the ℓ_2-norm leads to less computational complexity than the sparsity-induced norm, the iteration process in [21] calls for multiple matrix inverse calculations, which is time consuming when dealing with a large-scale training sample set case. Meanwhile, the multitask $\ell_{1,2}$-norm in [39] utilizes the accelerated proximal gradient (APG) method [44] to accelerate the convergence speed, which should be faster than the method in [21], in practice. For the proposed approach optimized by an alternating minimization method [45], iterating two convex subproblems can ensure the convergence of the whole objective function. One subproblem can be solved by the efficient multitask $\ell_{1,2}$-norm, and the other one has a closed-form solution. For the approaches that utilize a spatial prior, most of the current studies have defined a fixed-size local region for each test pixel and then simultaneously represented the neighboring pixels (weighted or equal) in the region to classify the test pixel [19], [21], [36], [40]. Although some improvements have been achieved by these techniques, the following issues should still be addressed. First, by involving the neighboring pixels, the computational burden is increased by several times over that of the single-pixel-based approach, as each pixel is repeatedly regarded as a collaborative assistant for its surrounding pixels. Second, the motivation behind treating the neighboring pixels as collaborative assistants refers to a smoothing filter by the use of the average pattern of the features in the fixed region, which will inevitably result in some oversmoothing [21]. For the proposed approach based on superpixel labeling, there is no redundant representation calculation for classifying each pixel, as we regard each superpixel as a nonoverlapping pure parcel. With consideration of the homogeneity and compactness, the superpixel covers the perceptual uniform region with an adaptive shape and structure. Compared with the aforementioned methods that utilize a spatial prior, the proposed superpixel-level sparse (collaborative) representation can adaptively exploit the correlation and distinctiveness of representative pixels in the superpixel, which maintains the discriminative representation for the pixel set. In addition, the proposed approach can be reduced as in [39], for the case of there being only one pixel for each superpixel, and can be reduced as an instance in [38], for the case of only a single feature being involved.

The rest of this paper is organized as follows: Section II presents the proposed framework for HSI classification. The experimental results of the proposed algorithm are given in Section III. Finally, Section IV concludes the paper with some concluding remarks.

II. Proposed Framework of SMTJSRC

Here, we first utilize the superpixel segmentation to partition the whole hyperspectral scene into several superpixels, and we then label each superpixel with the uniform multitask learning approach.

A. Segmentation and Superpixel Distance Definition

The remote sensing image analysis literature can be divided into the pixelwise processing approaches and the object-oriented approaches, and the classical pixelwise approaches often suffer from the "salt-and-pepper" phenomenon. In recent years, researchers of the first category of approaches [19], [21], [25] have smoothed the thematic map with the help of neighboring pixels, but this comes with the risk of oversmoothing, particularly when dealing with a complicated spatial structure. Although the later object-oriented approaches can alleviate the "salt-and-pepper" phenomenon by the use of spatial segmentation [46], the optimal segmentation scale is difficult to obtain, as the scale difference of the land cover is often complicated. In this paper, the superpixel containing several pixels in a local region is an intermediate concept between these two approaches. Superpixels in a hyperspectral scene are defined as pure perceptual uniform parcels, and a land-cover object is composed of

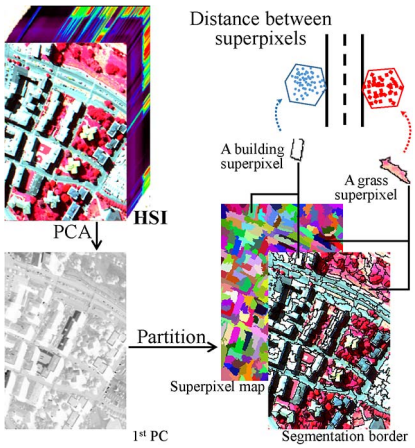


Fig. 1. Procedure of the superpixel segmentation and the hull-based superpixel-to-superpixel distance.

several adjacent superpixels. In this way, an oversegmentation procedure is utilized in the superpixel acquisition step.

The utilized superpixel segmentation method in this paper is based on graph partitioning and the entropy rate [42], which favors compact and homogenous nonoverlapping clusters, and has an efficient computational complexity approximated as $O(|V|\log|V|)$, where V refers to the number of pixels in the scene. The only free parameter T (the number of superpixels) controls the segmentation scale of the scene. More details of the segmentation mechanism can be found in [42]. For the proposed approach, the first principal component (PC) of the HSI is extracted, which maintains the most important information of the whole scene, and is utilized as the base image for the superpixel segmentation. As shown in Fig. 1, with the oversegmentation, a desirable superpixel with comparable scale and adaptive shape can avoid the oversmoothing problem described in [21], which is further discussed in the following experimental section.

As shown in Fig. 1, a parcel belonging to the grass class can be clustered as a superpixel, and denoted by $\mathbf{Y} = \{y_1, \ldots, y_i, \ldots, y_g\}$, where $y_i \in \mathbb{R}^d$. d is the dimension of the feature, and g is the number of pixels in the parcel. The hull of set $\mathbf{Y}$ is defined as $H(\mathbf{Y}) = \{\sum a_i y_i\}$. Usually, $\sum a_i = 1$ is required to be bounded as follows:

$$H(\mathbf{Y}) = \left\{ \sum a_i y_i | \sum a_i = 1, 0 \le a_i \le \tau \right\}. \quad (1)$$

If $\tau = 1$, $H(\mathbf{Y})$ is a convex hull, and if $\tau < 1$, $H(\mathbf{Y})$ is a reduced convex hull [47]. Both cases are referred to as a "convex hull" in the following. In a full HSI scene, another superpixel belonging to the building class can also be denoted as a sample set $\mathbf{Z} = \{z_1, \ldots, z_i, \ldots, z_l\}$, where $z_i \in \mathbb{R}^d$, and l is the number of pixels in the building superpixel. The distance between two superpixels is defined by modeling each set as a convex set, as follows:

$$\begin{aligned} &\min_{a,b} \left\| \sum a_i y_i - \sum b_i z_i \right\|_2^2 \\ &\text{s.t.} \sum a_i = 1, 0 \le a_i \le \tau; \sum b_i = 1, 0 \le b_i \le \tau. \end{aligned} \quad (2)$$

If there is no intersection between the two sets (as shown in Fig. 1), it can be observed that (2) denotes the distance between two support hyperplanes, which is equivalent to the distance measured by SVM [20]. That is to say, minimizing the distance can be transformed into maximizing the margin between the two convex hulls. If there are intersections, the associated classifier, which is related to soft-margin SVM and $\upsilon-$SVM [48], should be sensitive to the location of the samples and outliers. A more detailed discussion of this issue can be found in [48] and [49]. For a hyperspectral superpixel, it is believed that the oversegmentation procedure usually ensures that there is no intersection between superpixels of different classes.

B. Superpixel-Level CR Classification

The superiority of the CR-based hyperspectral classifiers [34], [41] is due to the utilization of the similar training samples from different classes to represent the test pixel. In view of this, it is natural to inherit such a mechanism in the superpixel CR classification framework.

Suppose that we have M distinct classes, then we set $\mathbf{D}_i \in \mathbb{R}^{d\times N_i}, i = 1, \ldots, M$ as the ith subdictionary whose columns are the N_i training samples from the ith class, and $N = \sum_i N_i$. It is believed that each $\mathbf{D}_i$ can model a convex set for a specific class. The collaborative dictionary $\mathbf{D}$, which is made up of all the subdictionaries $\{\mathbf{D}_i\}_{i=1,\ldots,M}$, and is concatenated as a uniform convex set, maps each hyperspectral pixel into a high-dimensional space corresponding to the dictionary. To classify the unlabeled superpixel $\mathbf{Y}$, we model each of $\mathbf{Y}$ and $\mathbf{D}$ as a hull, i.e., $\mathbf{Y}\boldsymbol{a} \in \mathbb{R}^d$ and $\mathbf{D}\boldsymbol{b} \in \mathbb{R}^d$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are coefficient vectors. The CR model of the convex hull of the superpixel $\mathbf{Y}$ can then be expressed as follows:

$$\mathbf{Y}\boldsymbol{a} = [\mathbf{D}_1\boldsymbol{b}_1, \ldots, \mathbf{D}_i\boldsymbol{b}_i, \ldots, \mathbf{D}_M\boldsymbol{b}_M] = \mathbf{D}\boldsymbol{b} + \boldsymbol{\varepsilon} \quad (3)$$

where $\boldsymbol{b}_i$ represents the coefficient subvector over the ith subdictionary $\mathbf{D}_i$, and $\boldsymbol{\varepsilon}$ is the random noise.

As described in [34], the CR vector contains discriminative information, to some degree, and can be used to perform the subsequent classification task. This is the motivation behind representing the convex hull of the unlabeled superpixel $\mathbf{Y}$ as a weighted composite of dictionary $\mathbf{D}$, whereas the coefficient vector $\boldsymbol{a}$ can be regarded as the weight to construct the convex hull of the superpixel. The collaborative vector $\boldsymbol{b}$ and the pixel weight $\boldsymbol{a}$ can be easily obtained by solving the following optimization problem:

$$\min_{a,b} \|\mathbf{Y}\boldsymbol{a} - \mathbf{D}\boldsymbol{b}\|^2 \quad \text{s.t.} \sum a_i = 1 \quad (4)$$

where we set $\sum a_i = 1$ to avoid the trivial solution $\boldsymbol{a} = \boldsymbol{b} = \mathbf{0}$. By minimizing the distance between $\mathbf{Y}\boldsymbol{a}$ and $\mathbf{D}\boldsymbol{b}$, the unexpected interruption in both the superpixel $\mathbf{Y}$ and the dictionary $\mathbf{D}$ will be assigned with a very small representation coefficient. As with SRC, the simplest method to find its label is by the least squares method, which can be expressed as $\text{class}(\mathbf{Y}) = \arg\min_i \|\mathbf{Y}\boldsymbol{a} - \mathbf{D}_i\boldsymbol{b}_i\|_2^2$. However, in practice, some light disturbance may be misclassified to class $j (j \ne i)$ as $\|\mathbf{Y}\boldsymbol{a} - \mathbf{D}_i\boldsymbol{b}_i\|_2^2 > \|\mathbf{Y}\boldsymbol{a} - \mathbf{D}_j\boldsymbol{b}_j\|_2^2$, which leads to an unstable

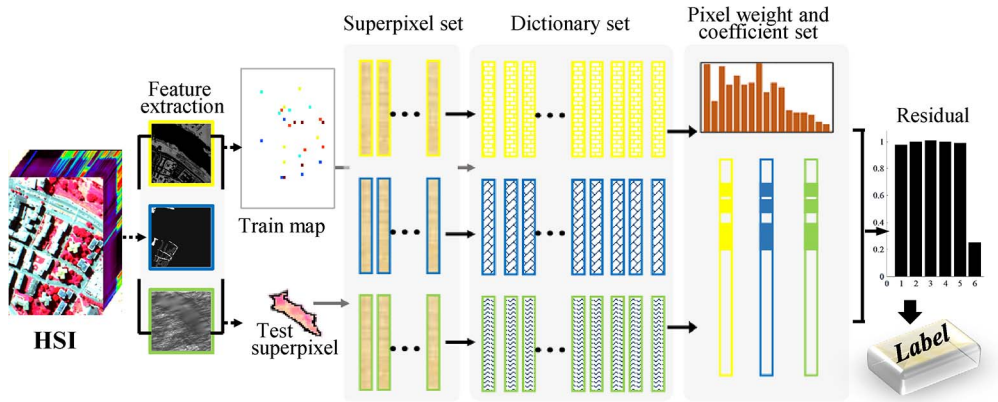


Fig. 2. Schematic illustration of the SMTJSRC algorithm for HSI classification. Multiple modalities of the features are first extracted for the whole hyperspectral scene. A dictionary set and a superpixel set both containing multiple features are then constructed. Each coefficient vector is represented as a linear combination of the corresponding training feature dictionary. To preserve the invariant similarities between various features, a multitask joint sparsity norm, which enforces the joint selection of a few common classes of training samples to represent a test superpixel over each feature and each instance, is introduced. Meanwhile, the practical instance of the convex hull can be simultaneously learned with the bounded coefficient (known as the pixel weight in this framework). Finally, the classification decision is made, according to the reconstruction error of the individual class.

classification result [41]. To alleviate this problem, a regularization term can be considered to generate a more stable model for the superpixel CR as follows:

$$\min_{a,b} \|\mathbf{Y}\boldsymbol{a} - \mathbf{D}\boldsymbol{b}\|_2^2 \text{ s.t.} \sum a_i = 1, \|\boldsymbol{a}\|_{\ell_p} < \delta_1, \|\boldsymbol{b}\|_{\ell_q} < \delta_2 \tag{5}$$

where ℓ_p, ℓ_q can be the ℓ_1- or ℓ_2-norm, with regard to the different instances of the general model in (5). In theory, the ℓ_p-norm represents the distribution of the image noise, and the ℓ_q-norm refers to the prior of the coefficient. δ_1 and δ_2 are two bound constants to constrain the convex hulls.

For classification, the class of the superpixel $\mathbf{Y}$ can be determined by minimizing the residual r_i (i.e., the error between $\mathbf{Y}\boldsymbol{a}$ and the linearly recovered approximated result from the training samples in the ith class) as follows:

$$\text{class}(\mathbf{Y}) = \min_{i=1,\ldots,M} r_i(\mathbf{Y}\boldsymbol{a}) = \min_{i=1,\ldots,M} \|\mathbf{Y}\boldsymbol{a} - \mathbf{D}_i\boldsymbol{b}_i\|_2. \tag{6}$$

Compared with the pixelwise joint CR-based approach incorporating neighboring pixels [19], [40], [41], superpixel-level CR has the following differences. First, the former approach utilizes $\|\mathbf{Y} - \mathbf{D}_i\mathbf{B}_i\|_F^2$ (where each column of $\mathbf{B}$ refers to the coefficient vector associated with the pixels in $\mathbf{Y}$), which means the residual of the average vector of $\mathbf{Y}$, and it acts as a smoothing filter in practice. Meanwhile, for the proposed superpixel-based approach, the coding step can adaptively utilize the convex combination of the pixel set $\mathbf{Y}$, which reflects the centroid of this parcel with a more stable feature representation. Second, it is redundant to repeatedly take one pixel into the coding step of its surrounding pixel labeling process, and the proposed superpixel-level approach is more efficient, as there are no overlapping parcels in the segmentation step. Finally, it is also noted that the utilized superpixel based on oversegmentation extracts more adaptive pure parcels with comparable scales, whereas the pixelwise joint CR-based approaches with a fixed neighborhood may get into problems with neighborhoods containing several different kinds of land cover.

C. Multitask Learning Algorithm for Superpixel-Level CR Classification

Designing a classifier that combines multiple features (e.g., spectral, texture, and shape) to improve the classification accuracy is natural, as one single feature can only depict the HSI from one perspective. Since none of the feature descriptors will have the same discriminative power for all classes, the multitask learning approach can further fuse the complementary discriminative abilities of the different features by simultaneous use of the specific learned convex hull of each feature. A visual illustration of the classification scheme for an HSI with the proposed SMTJSRC algorithm is shown in Fig. 2.

Considering each feature as a modality, (3) can be extended as follows:

$$\begin{aligned}
&\mathbf{Y}^1\boldsymbol{a}^1 = \left[\mathbf{D}_1^1\boldsymbol{b}_1^1, \ldots, \mathbf{D}_i^1\boldsymbol{b}_i^1, \ldots, \mathbf{D}_M^1\boldsymbol{b}_M^1\right] = \mathbf{D}^1\boldsymbol{b}^1 + \boldsymbol{\varepsilon}^1 \\
&\quad\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
&\mathbf{Y}^K\boldsymbol{a}^K = \left[\mathbf{D}_1^K\boldsymbol{b}_1^K, \ldots, \mathbf{D}_i^K\boldsymbol{b}_i^K, \ldots, \mathbf{D}_M^K\boldsymbol{b}_M^K\right] = \mathbf{D}^K\boldsymbol{b}^K + \boldsymbol{\varepsilon}^K
\end{aligned} \tag{7}$$

where the convex sets of the multiple different features $\{\mathbf{Y}^k\}_{k=1,\ldots,K}$ are extracted from the different perspectives of the unlabeled hyperspectral superpixel, and their corresponding subdictionaries $\{\mathbf{D}^k\}_{k=1,\ldots,K}$ are constructed with the features of the same training samples. K is the number of modalities. $\{\boldsymbol{a}^k\}_{k=1,\ldots,K}$, $\{\boldsymbol{b}^k\}_{k=1,\ldots,K}$, and $\{\boldsymbol{\varepsilon}^k\}_{k=1,\ldots,K}$ are the pixel weight set, the collaborative coefficient set, and the random noise set, respectively.

To make the representation step stable, the multitask-learning-based framework also utilizes regularizations to extend (5) as follows:

$$\begin{aligned}
&\min_{\{a^k,b^k\}} \sum_k^K \|\mathbf{Y}^k\boldsymbol{a}^k - \mathbf{D}^k\boldsymbol{b}^k\|_2^2 \\
&\text{s.t.} \sum a_i^k = 1, \|\boldsymbol{a}^k\|_{\ell_p} < \delta_1, \|\boldsymbol{b}^k\|_{\ell_q} < \delta_2,\ k = 1, \ldots, K.
\end{aligned} \tag{8}$$

To combine the discriminative abilities of the multiple features for classification, a joint learning procedure is utilized for (8) by imposing a class-level sparsity-inducing term on $\{\boldsymbol{b}^k\}_{k=1,\ldots,K}$. Similar to the motivation behind SRC, it can be useful to jointly select a few common classes of training samples to represent a test convex hull over each feature. That is to say, the desired representation vectors for the multiple features should share certain class-level sparsity patterns. Given the optimal superpixel convex hull $\mathbf{Y}^k\boldsymbol{a}^k$, the CR coefficient vector $\boldsymbol{b}^k$ can be rewritten as $\boldsymbol{b}^k = [\boldsymbol{b}_1^{k'}, \ldots, \boldsymbol{b}_M^{k'}]$, in which $\boldsymbol{b}_i^k$ consists of the components of $\boldsymbol{b}^k$ restricted on class i. We stack all the CR vectors together, and $\boldsymbol{B}_i = [\boldsymbol{b}_i^1, \ldots, \boldsymbol{b}_i^K]$ denotes the representation coefficients associated with class$i$ across the different features. Inspired by the sparsity constraint utilized in SRC and the role of the multiple features, the class-level joint sparsity-inducing term [the ℓ_q-norm in (8)], which applies the ℓ_0-norm across the ℓ_2-norm of $\mathbf{B}_i$, can be shown as $\|[\|\boldsymbol{B}_1\|_F, \ldots, \|\boldsymbol{B}_M\|_F]\|_0$. Since the ℓ_0-norm problem is NP-hard, convex relaxation $\|[\|\boldsymbol{B}_1\|_F, \ldots, \|\boldsymbol{B}_M\|_F]\|_1$, which still encourages the test superpixel convex hull set to be sparsely reconstructed by the most representative classes in the training set, is also workable. Considering the superpixel constructed before, it is believed that pixels in the parcel will be homogenous and similar. In view of this, the ℓ_2-norm regularization is available to make the problem stable, and with a light computational complexity. We can rewrite the regularized multitask learning model in (8) as its Lagrangian formulation, as follows:

$$\begin{aligned}&\min_{\{\mathbf{A},\mathbf{B}\}} \sum_k^K \|\mathbf{Y}^k\boldsymbol{a}^k - \mathbf{D}^k\boldsymbol{b}^k\|_2^2 + \lambda\|\mathbf{A}\|_F + \eta\|\mathbf{B}\|_{1,2}\\ &\text{s.t. } \mathbf{A} = [\boldsymbol{a}^1, \ldots, \boldsymbol{a}^K], \mathbf{B} = [\boldsymbol{b}^1, \ldots, \boldsymbol{b}^K],\\ &\qquad \sum a_i^k = 1,\ k = 1, \ldots, K \end{aligned} \tag{9}$$

where λ and η are positive constants to balance the data fidelity term and the regularizations.

To solve the proposed problem, we use the alternating minimization method, which is very efficient and effective in solving multiple variable optimization problems [45]. The Lagrange function of (9) can be reorganized as follows:

$$\begin{aligned}L\left(\mathbf{A}, \mathbf{B}, \{\gamma^k\}_{k=1,\ldots,K}\right) = &\sum_k^K \|\mathbf{Y}^k\boldsymbol{a}^k - \mathbf{D}^k\boldsymbol{b}^k\|_2^2\\ &+ \lambda_1\|\mathbf{A}\|_F + \eta\|\mathbf{B}\|_{1,2} + \sum_k^K \langle\gamma^k, \boldsymbol{e}\boldsymbol{a}^k - 1\rangle\end{aligned} \tag{10}$$

where $\{\gamma^k\}_{k=1,\ldots,K}$ is the Lagrange multiplier set, $\langle\bullet,\bullet\rangle$ is the inner product, and $\boldsymbol{e}$ is a row vector whose elements are 1. As a convex optimization problem with two variables, $\mathbf{A}$ and $\mathbf{B}$ can be solved by alternating the optimization with the two corresponding subproblems, until the solutions converge to a minimum.

For the first subproblem, we optimize $\mathbf{A}$ by fixing $\mathbf{B}$, and the optimization of (10) becomes

$$\begin{aligned}L(\boldsymbol{a}^k, \gamma^k) = \|\mathbf{Y}^k\boldsymbol{a}^k - \mathbf{D}^k\boldsymbol{b}^k\|_2^2 + \lambda\|\boldsymbol{a}^k\|_2^2 + \langle\gamma^k, \boldsymbol{e}\boldsymbol{a}^k - 1\rangle\\ \text{s.t. } k = 1, \ldots, K\end{aligned} \tag{11}$$

where

$$\frac{\partial L}{\partial \gamma^k} = \boldsymbol{e}\boldsymbol{a}^k - 1 \tag{12}$$

$$\frac{\partial L}{\partial \boldsymbol{a}^k} = 2\mathbf{Y}^{k'}(\mathbf{Y}^k\boldsymbol{a}^k - \boldsymbol{x}^k) + 2\lambda\boldsymbol{a}^k + \gamma_k\boldsymbol{e}^T, \quad \text{s.t. } \boldsymbol{x}^k = \mathbf{D}^k\boldsymbol{b}^k. \tag{13}$$

According to (12) and (13), there is a closed-form solution to (11) as follows:

$$\gamma^k = \frac{2(\boldsymbol{e}\boldsymbol{p}^k - 1)}{\boldsymbol{e}\mathbf{Q}^k\boldsymbol{e}^T}, \quad \boldsymbol{w}^k = \boldsymbol{p}^k - 0.5\gamma_k\mathbf{Q}^k\boldsymbol{e}^T \tag{14}$$

where $\mathbf{Q}_k = (\mathbf{Y}^{k'}\mathbf{Y}^k + \lambda)^{-1}$ and $\boldsymbol{p}_k = \mathbf{Q}_k\mathbf{Y}^k\boldsymbol{x}^k$.

For the second subproblem, we optimize $\mathbf{B}$ by fixing $\mathbf{A}$, and the optimization of (10) becomes

$$\min_{\mathbf{B}} \sum_k^K \left\|\boldsymbol{s}^k - \mathbf{D}^k\boldsymbol{b}^k\right\|_2^2 + \eta\|\mathbf{B}\|_{1,2} \quad \text{s.t. } \boldsymbol{s}^k = \mathbf{Y}^k\boldsymbol{a}^k. \tag{15}$$

The problem of (15) is known as the multitask joint covariate selection model [50] in sparse learning, and can be efficiently solved with several iterations by the APG method [39]. As discussed in [45], the alternating minimization approach for such a general convex problem will converge to the correct solution, as both subproblems are convex. It is notable that running (15) until convergence is not necessary for the best classification performance. Indeed, a satisfactory classification accuracy can be obtained within a few hundred iterations. This can be partially explained by the fact that the objective of SMTJSRC is to minimize the reconstruction error of a test convex hull set. In practice, it is desirable to set a bound of the iteration times for the proposed alternating optimization problem as a "convergence condition." As in the multitask learning approaches [21], the label of the unlabeled superpixel is then determined by the minimal total residual as follows:

$$\text{class}(\mathbf{Y}) = \arg\min_{i=1,\ldots,M} \sum_{k=1}^K \left\|\mathbf{Y}^k\boldsymbol{a}^k - \mathbf{A}_i^k\boldsymbol{b}_i^k\right\|_2^2. \tag{16}$$

The computational burden for the proposed SMTJSRC algorithm is as follows. The running time for the first subproblem is $O(Kd_{\max}^2)$, where $d_{\max}$ is the maximum of the dimension of the multiple features, as $\mathbf{Q}$ and its inverse can be precomputed. For the second subproblem, L_1 is the average iteration times utilized to solve (15), and it takes $O((2L_1 + 1)Kd_{\max})$ [39]. It is also notable that the computational load of (16) is negligible in the superpixel multitask learning procedure. Finally, L_2 is the average iteration times for the alternating optimization, and the computational cost for labeling the whole hyperspectral scene is $O(TL_2((2L_1 + 1)Kd_{\max} + Kd_{\max}^2) + |V|\log|V|)$.

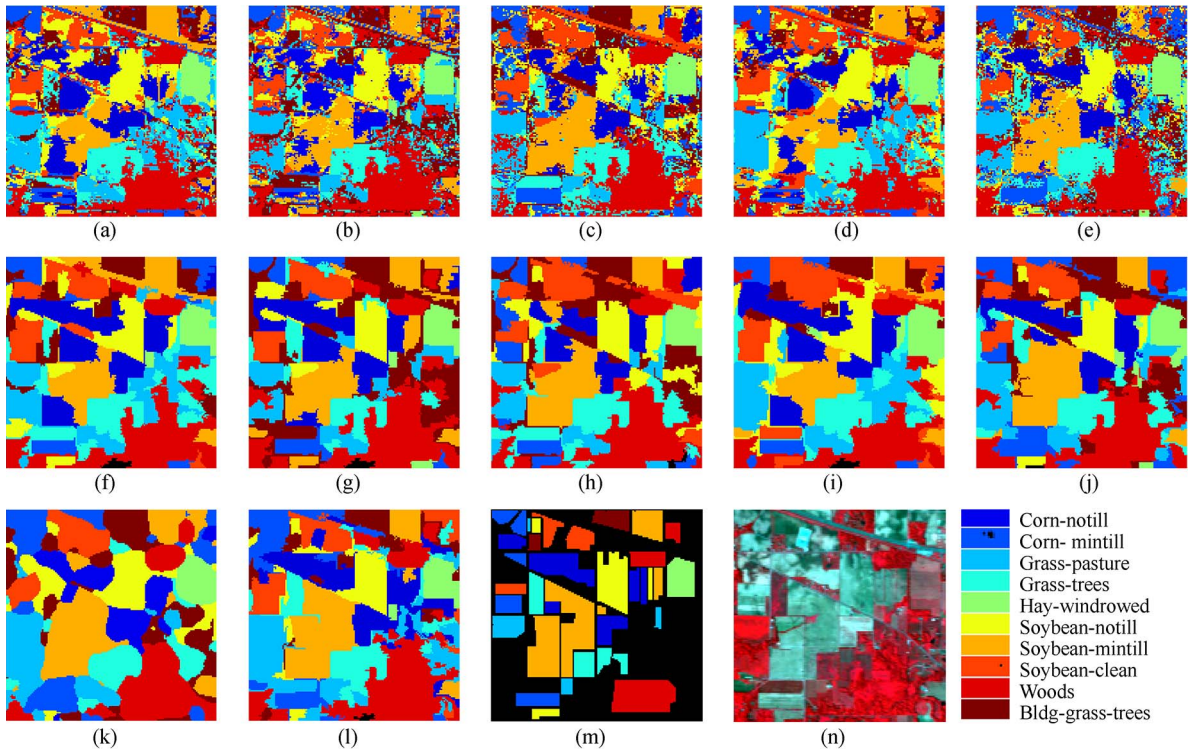


Fig. 3. Classification results for the Indian Pines image: (a) SVM-VS, (b) SVM-CK, (c) GCK-MLR, (d) MNFL, (e) MTJSRC, (f) SVM-VS-P, (g) SVM-CK-P, (h) GCK-MLR-P, (i) MNFL-P, (j) MTJSRC-P, (k) JCRC-MTL, (l) SMTLJSRC, (m) reference set, and (n) false-color image (R: 57, G: 27, B: 17).

The implementation details of the proposed SMTJSRC algorithm for HSI classification are shown in Algorithm 1.

Algorithm 1: The Framework of the Proposed Algorithm for HSI Classification

Input: 1) A HSI containing training samples
2) Regularization parameters λ and η, and number of superpixels T

Step 1: Segment the whole scene into T superpixels

Step 2: Extract multiple features and construct the multitask dictionary from the image

Step 3: For each test superpixel in the hyperspectral scene:
Initialize $\mathbf{B}^{(0)}$, and $t \leftarrow 0$
While not meeting the convergence condition:
Step 3.1: Fix $\mathbf{B}^{(t)}$ and update $\mathbf{A}^{(t+1)}$ by (14).
Step 3.2: Fix $\mathbf{A}^{(t+1)}$ and update $\mathbf{B}^{(t+1)}$ by (15).
Step 3.3: $t \leftarrow t+1$
End while
End For

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

III. Experiments

A. Data Set Description

Three hyperspectral data sets collected by different instruments were used in the experiments. The first scene was gathered by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor on June 12, 1992, over the Indian Pines test site in Northwestern Indiana, and consists of 145 $\times$ 145 pixels and 224 spectral reflectance bands in the wavelength range of 0.4–2.0 μm. The false-color composite of the Indian Pines image is shown in Fig. 3(a). We also reduced the number of bands to 200, by removing bands covering the regions of water absorption: [104–108], [150–163], and 220. The spatial resolution for this image is about 20 m. This image contains ten ground-truth classes, and the numbers of the reference data and the corresponding visual map are shown in Table I and Fig. 3(b), respectively. This data set is the most widely used data set for testing the accuracy of hyperspectral classification algorithms. It is a challenging task, since some of the classes are highly mixed with quite similar spectral signatures.

The second scene was acquired by the Reflective Optics Systems Imaging Spectrometer (ROSIS) sensor over Pavia University, in northern Italy. It consists of 610 $\times$ 610 pixels and 115 spectral reflectance bands. We selected 103 of the bands and cut a patch sized 610 $\times$ 340. The false-color composite of the Pavia University image is shown in Fig. 4(a). The geometric resolution of this image is 1.3 m. This image contains nine reference classes, and details of the quantities and the corresponding visual map are shown in Table II and Fig. 4(b), respectively.

The final image was a part of an airborne hyperspectral data flight line over the Washington DC Mall, which was acquired by the Hyperspectral Digital Image Collection Experiment (HYDICE) sensor, and was provided with the permission of the Spectral Information Technology Application Center of Virginia, who was responsible for its collection. The sensor system used in this case measured the pixel response in 210 bands in the 0.4–2.4-μm region of the visible and infrared spectrum. Bands in the 0.9- and 1.4-μm regions, where the atmosphere is

TABLE I
REFERENCE INFORMATION, CLASSIFICATION ACCURACY (%), AND RUNNING TIME (IN SECONDS) FOR THE INDIAN PINES IMAGE WITH THE TEST SET

CLASS No.	Name	Num.	SVM-VS	SVM-VS-P	SVM-CK	SVM-CK-P	GCK-MLR	GCK-MLR-P	MNFL	MNFL-P	MTJSRC	MTJSRC-P	JCRC-MTL	SMTLJSRC
1	Corn-notill	1428	58.01	69.89	54.93	70.63	65.37	71.93	55.63	65.97	63.21	75.56	68.41	**81.15**
2	Corn-mintill	830	61.60	71.84	63.50	80.23	82.55	91.35	64.17	73.67	79.37	**91.44**	87.13	87.15
3	Grass-pasture	483	78.90	83.64	83.45	85.79	85.33	86.66	86.70	87.78	85.90	85.37	89.05	**89.09**
4	Grass-trees	730	91.13	97.99	91.39	97.81	90.74	96.49	92.39	95.17	97.64	99.31	94.32	**99.42**
5	Hay-windrowed	478	98.35	99.81	98.01	99.79	99.55	99.79	99.59	99.79	**99.98**	99.79	99.96	99.85
6	Soybean-notill	972	69.66	77.80	67.20	73.93	79.16	**84.96**	71.87	79.46	77.15	83.42	82.98	84.11
7	Soybean-mintill	2455	63.15	70.74	65.03	78.72	74.17	81.38	73.19	82.07	75.47	84.45	82.28	**87.21**
8	Soybean-clean	593	59.55	77.41	57.50	79.69	86.16	**97.62**	75.61	91.25	83.40	96.84	92.09	96.84
9	Woods	1265	86.58	87.54	86.32	89.45	93.16	**95.82**	88.07	89.45	85.67	88.90	90.24	91.22
10	Bldg-grass-trees	386	81.81	94.23	79.68	93.80	94.65	**97.61**	82.45	93.38	91.06	94.15	97.15	97.34
OA			71.14	79.10	71.07	82.06	81.19	87.04	75.56	82.81	79.99	87.28	85.06	**89.15**
			±3.52	±5.30	±3.05	±4.70	±2.95	±2.44	±4.20	±5.41	±2.90	±3.27	±2.26	±2.30
κ			0.669	0.760	0.667	0.793	0.784	0.851	0.719	0.802	0.770	0.853	0.828	**0.875**
			±0.040	±0.059	±0.035	±0.054	±0.034	±0.028	±0.047	±0.061	±0.033	±0.037	±0.026	±0.026
Time			0.837	1.047	0.710	0.911	4.332	4.570	0.325	0.614	19.958	20.193	91.940	1.551

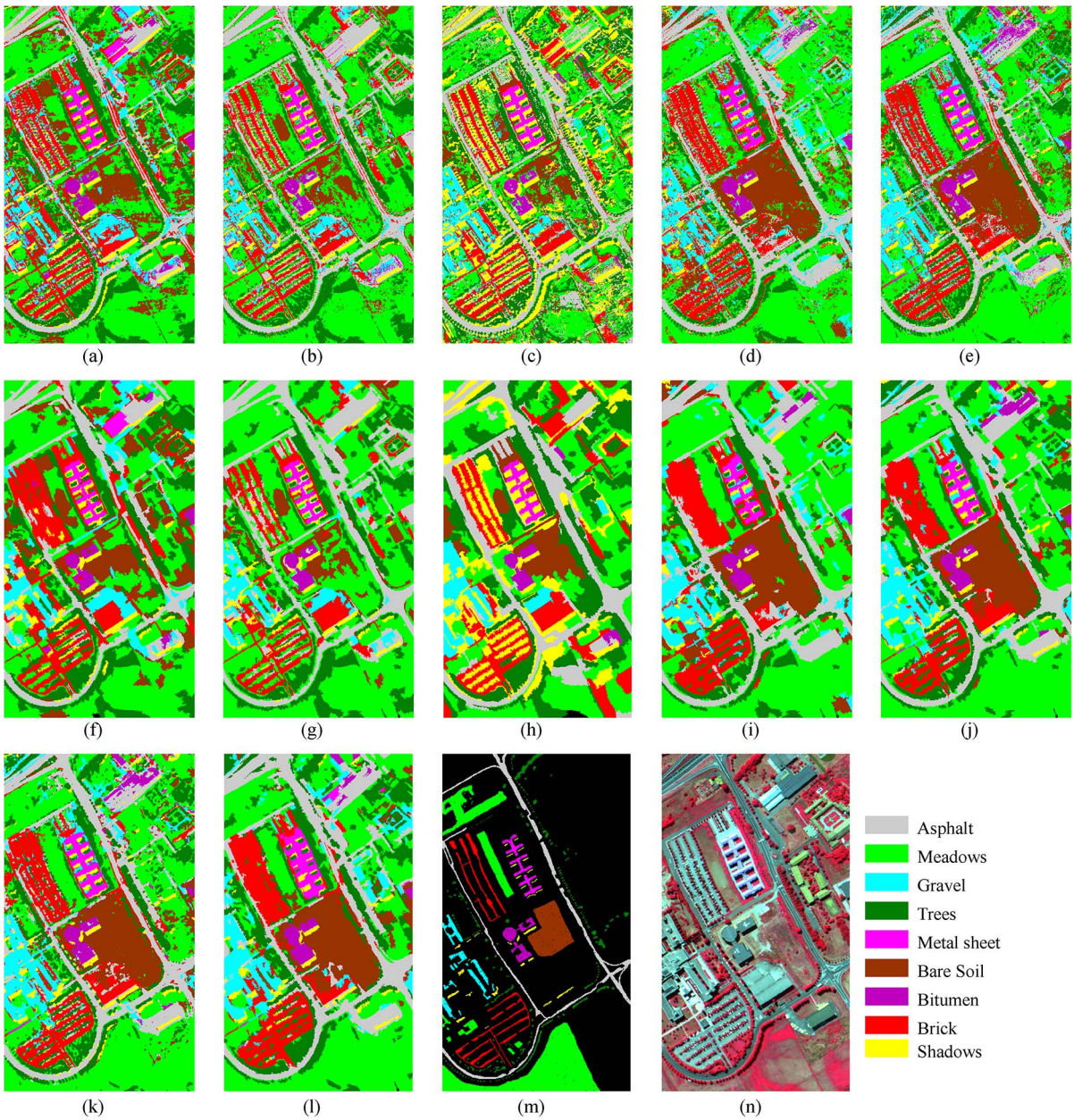


Fig. 4. Classification results for the Pavia University image: (a) SVM-VS, (b) SVM-CK, (c) GCK-MLR, (d) MNFL, (e) MTJSRC, (f) SVM-VS-P, (g) SVM-CK-P, (h) GCK-MLR-P, (i) MNFL-P, (j) MTJSRC-P, (k) JCRC-MTL, (l) SMTLJSRC, (m) reference set, and (n) false-color image (R: 102, G: 56, B: 31).

TABLE II
REFERENCE INFORMATION, CLASSIFICATION ACCURACY (%), AND RUNNING TIME (IN SECONDS) FOR THE PAVIA UNIVERSITY IMAGE WITH THE TEST SET

No.	Class Name	Num.	SVM-VS	SVM-VS-P	SVM-CK	SVM-CK-P	GCK-MLR	GCK-MLR-P	MNFL	MNFL-P	MTJSRC	MTJSRC-P	JCRC-MTL	SMTLJSRC
1	Asphalt	6631	79.95	86.99	84.95	92.90	90.05	93.20	90.69	95.43	88.61	**96.23**	93.76	96.10
2	Meadows	18649	74.81	79.98	76.44	78.64	66.06	77.32	88.86	91.07	93.31	94.54	94.04	**95.74**
3	Gravel	2099	78.20	87.86	91.80	98.56	94.06	**98.58**	92.27	96.18	95.21	96.57	96.45	97.77
4	Trees	3064	94.97	89.48	**95.83**	92.24	93.38	84.91	93.84	86.31	95.42	91.39	92.85	92.16
5	Metal sheet	1345	99.49	98.58	99.35	98.35	99.71	98.80	99.66	98.44	99.90	98.80	**99.99**	99.22
6	Bare soil	5029	75.82	81.79	67.09	69.71	73.41	89.08	96.64	98.96	96.38	**99.97**	98.21	99.86
7	Bitumen	1330	95.86	98.03	95.95	98.71	94.87	98.86	99.86	99.70	99.45	99.59	**99.90**	99.56
8	Brick	3682	80.77	89.30	87.82	92.66	91.41	97.02	91.44	95.19	95.38	**98.12**	97.36	97.43
9	Asphalt	6631	**99.78**	92.37	97.99	94.15	98.66	85.18	80.37	78.84	98.05	94.38	98.53	89.43
	OA		79.82	83.09	81.57	84.53	78.81	85.22	91.29	93.17	93.86	95.91	95.27	**96.36**
			±2.38	±2.80	±4.19	±4.23	±3.16	±4.59	±2.37	±2.44	±1.96	±1.81	±1.96	±1.46
	κ		0.743	0.784	0.764	0.801	0.735	0.814	0.887	0.911	0.920	0.946	0.938	**0.952**
			±0.027	±0.033	±0.052	±0.053	±0.037	±0.054	±0.030	±0.031	±0.025	±0.023	±0.025	±0.019
	Time		6.16	10.15	2.51	5.20	36.25	42.12	0.8903	4.18	193.72	196.86	443.02	20.49

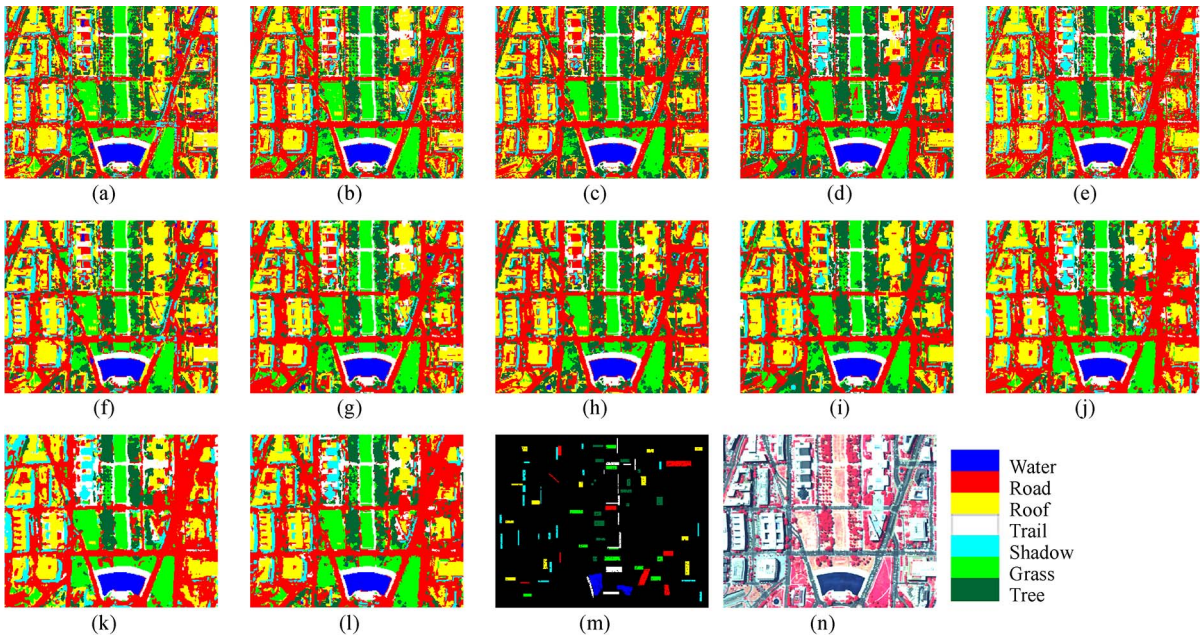


Fig. 5. Classification results for the Washington DC image: (a) SVM-VS, (b) SVM-CK, (c) GCK-MLR, (d) MNFL, (e) MTJSRC, (f) SVM-VS-P, (g) SVM-CK-P, (h) GCK-MLR-P, (i) MNFL-P, (j) MTJSRC-P, (k) JCRC-MTL, (l) SMTLJSRC, (m) reference set, and (n) false-color image (R: 63, G: 52, B: 36).

opaque, have been omitted from the data set, leaving 191 bands. The data set contains 250 scan lines, with 307 pixels in each scan line. The false-color composite of the Washington DC image is shown in Fig. 5(a). This image contains seven reference classes, as shown in Table IV. This data set is challenging due to its complicated spatial distribution.

B. Experimental Setting

The goal of the experiments was to investigate the effectiveness of the proposed multiple-feature fusion algorithm for the classification of hyperspectral data sets. For the multiple-feature extraction, we utilized three meaningful and widely used features for HSI analysis: 1) the spectral value feature; 2) the Gabor texture feature; and 3) the extended morphological attribute profile (EMAP) feature. For each data set, the detailed parameter settings of each spatial feature followed the specific guidelines [15], [21].

Based on these multiple features, a number of state-of-the-art multiple-feature fusion algorithms in the literature were taken as benchmarks, as illustrated in Table III, in detail. In addition, superpixel segmentation, as described in Section II-A, was utilized as a postprocessing step for each classifier that does not utilize a spatial prior, for a fair comparison [51]. It should be noted that all the SVM-related algorithms were implemented with the help of the LibSVM [52] package, which utilizes C++ software to speed up the process. For each data set, we randomly selected ten pixels for each class as the training samples, and the rest as the test samples from the reference data to validate the performances. The range of the regularization parameters for the tenfold cross-validation was from 10^{-10} to 10^{-2}, and the number of superpixels was empirically varied

TABLE III
CLASSIFICATION APPROACHES IN THE COMPARISON FOR SECTION III-C

Acronym	Description in short	Linear/nonlinear	Spatial prior or not	Manner of multiple-feature combination
SVM-VS	Standard SVM with a radial basis function (RBF) kernel, where VS denotes the approach that stacks multiple features into an "incremental" feature	Nonlinear	No	Vector stacking
SVM-CK	SVM with RBF kernel, where CK denotes a composite kernel, as in [27] (MLR)	Nonlinear	No	Composite kernel
GCK-MLR	MLR with a generalized composite kernel (GCK) to combine the multiple features [27]	Nonlinear	No	Generalized composite kernel
MNFL	A regularization-free multiple-feature learning method based on MLR [18]	Nonlinear	No	Vector stacking
MTLJSRC	Class-level sparsity constraint on the representations across multiple features, which is also named SRC-MTL in [21]	Linear	No	Sparse (collaborative) representation-based multi-task learning
JCRC-MTL	Assuming both similarity for labeling and dissimilarity for the pixel description between multiple features, as well as some contextual prior in the spatial domain [21]	Linear	Neighborhood contextual smoothing	Sparse (collaborative) representation-based multi-task learning
SMTLJSRC	The proposed superpixel-based algorithm, which can be seen as MTLJSRC when each superpixel has been degraded into pixels	Linear	Superpixel	Sparse (collaborative) representation-based multi-task learning

Addition: For the first five algorithms that do not utilize a spatial prior, the superpixel segmentation, as in Section II.A, can be utilized as a post-processing step. Taking "SVM-VS" as an example, we label the associated approach as "SVM-VS-P".

TABLE IV
REFERENCE INFORMATION, CLASSIFICATION ACCURACY (%), AND RUNNING TIME (IN SECONDS) FOR THE WASHINGTON DC IMAGE WITH THE TEST SET

CLASS No.	Name	Num.	SVM-VS	SVM-VS-P	SVM-CK	SVM-CK-P	GCK-MLR	GCK-MLR-P	MNFL	MNFL-P	MTJSRC	MTJSRC-P	JCRC-MTL	SMTLJSRC
1	Water	625	95.15	96.54	93.54	93.30	95.40	94.83	97.71	97.45	98.18	97.85	98.10	**98.21**
2	Road	952	87.60	89.46	93.68	94.34	94.80	95.49	95.67	**96.87**	94.23	95.63	95.22	96.58
3	Roof	780	86.09	**87.06**	83.21	84.10	81.48	83.01	82.14	82.75	83.82	84.42	84.69	86.51
4	Trail	754	84.10	83.66	92.03	91.52	90.99	90.90	95.51	95.27	94.97	**95.79**	89.95	93.01
5	Shadow	764	92.60	91.82	97.06	95.56	95.70	94.79	96.67	95.76	97.14	96.29	90.23	**97.59**
6	Grass	725	90.76	90.69	93.55	93.69	97.57	**97.65**	92.06	92.17	96.41	96.71	94.99	97.31
7	Tree	745	97.36	95.90	96.49	95.24	98.93	97.82	**99.36**	98.57	97.36	96.31	96.24	97.89
	OA		90.27 ±2.49	90.52 ±2.45	92.76 ±1.98	92.54 ±1.92	93.47 ±0.82	93.46 ±0.86	94.08 ±1.77	94.09 ±1.59	94.42 ±3.04	94.61 ±2.67	92.67 ±5.02	**95.22** ±1.91
	κ		0.886 ±0.029	0.889 ±0.029	0.915 ±0.023	0.913 ±0.022	0.924 ±0.010	0.923 ±0.010	0.931 ±0.021	0.931 ±0.019	0.935 ±0.36	0.937 ±0.031	0.914 ±0.059	**0.944** ±0.022
	Time		2.22	6.63	1.26	5.54	7.51	11.78	0.39	4.77	58.52	62.78	129.53	25.30

from $W/100$ to $W/10$, where W is the total number of pixels in the scene. The parameter settings for the other benchmark classifiers were also selected by tenfold cross-validation from a reasonable range. The classification accuracy values were averaged over ten runs for each classifier to reduce the possible bias induced by the random sampling. 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 (κ) is a robust measure of the degree of agreement, and the classification accuracy of the different classifiers with the test set for each class can be found in the corresponding columns. All the experiments, except for the SVM-related work, were carried out using MATLAB on a personal computer with a single i7-3770K 3.50-GHz processor and 32.0 GB of RAM.

C. Experimental Results

The thematic maps of the various classifiers are visually shown in Figs. 3–5. For the test pixels, the average individual labeling accuracy for each class, the average OA with the associated standard deviation, and the average kappa coefficient (κ) with the standard deviation for the different classifiers are shown in Tables I, II, and IV, respectively. The average running times for the labeling of all the pixels in the scene are shown in the bottom line of the quantitative evaluation tables.

It can be observed in Table I that the proposed SMTLJSRC method yields the best accuracy and the most stable performance. For these multiple-feature-based classifiers, it can be first observed that the inferior accuracy of SVM-VS and SVM-CK suggests the limitations of the pixel-based classification methods without a spatial prior. With the postprocessing step, the corresponding classification results are significantly improved, which indicates the superiority of the segmentation over the neighborhood smoothing for such a homogenous scene. That is to say, for such a homogenous image with subtle spectral differences between classes, it is believed that the spatial prior, which is utilized to stabilize the signal and alleviate the "salt-and-pepper" phenomenon, is highly effective, as both the quantitative evaluation and the thematic map indicate.

For the three representation-based classifiers using spatial priors, including the neighborhood smoothing, the postprocessing, and the proposed approach, it can be observed that JCRC-MTL suffers from the "oversmoothing" problem, whereas the other two methods can maintain the edge information in most cases. Compared with the proposed approach, the postprocessing-related subfigures suffer from blockwise misclassification, as the pixelwise classifier cannot undertake major correction in a local region. For the running time comparison, it can be seen that MNFL is the fastest and that SVM-VS, SVM-CK, and the proposed SMTJSRC classifiers are comparable but a bit slower than the former, and the other classifiers are the slowest. Overall, it is concluded that the proposed SMTLJSRC is more efficient than the other two representation-based classifiers, with superior classification accuracy at the same time.

As both the Pavia University image and the Washington DC image are hyperspectral and high-spatial-resolution remote sensing images, the classification results with these multiple-feature-based classifiers show some similarities. First, the spatial features are more meaningful for this kind of remote sensing data, and can effectively improve the classification result. Based on the complementary multiple features, different learning approaches for utilizing these features can significantly affect the discriminability. For both hyperspectral data sets, the VS-based classifiers are generally inferior to the others, whereas the proposed superpixel-level sparse representation-based multitask learning mechanism shows the best performance, which is consistent with the classification results of the Indian Pines image. However, there are still some individual issues for these two hyperspectral and high-spatial-resolution images. For the Pavia University image, utilizing the contextual information with a regular pattern (i.e., stacking neighboring pixels to the test pixel together with an equal weight) is still useful, as each parcel in the scene covers tens of pixels at the least. The optimal neighborhood size of JCRC-MTL for classifying the Pavia University image is 3×3 (as fully analyzed in [21]), and the "oversmoothing" problem is not significant. In view of this, JCRC-MTL is superior to MTLJSRC, and is comparable to MTLJSRC-P with this image. However, for the Washington DC image, the spatial structure of each parcel is more complicated, and the contextual prior utilized in JCRC-MTL imports some pixels in different classes as "fake assistance," which reduces the discriminability, as shown in Table IV. Meanwhile, as the desired size of superpixels in the Washington DC image is small, the running time of the proposed approach is somewhat longer than for the MLR-related approach, due to the complicated spatial structure. For both data sets, it is still suggested that the proposed spatial-prior-related algorithm shows a more efficient and superior performance.

D. Analysis of Superpixel-Related Issues

Here, the superiority of the superpixel CR and the relationship between the computational cost and the superpixel segmentation scale is analyzed.

Fig. 6 presents the effect of the superpixel size on the performance of the proposed method. Considering the complicated spatial structure, the Washington DC image was utilized here to demonstrate the relationships. The experimental setting was the same as before, and the parameters were set by cross-validation. In this figure, the horizontal axis indicates the number of superpixels in the scene, the vertical axis shows the corresponding average running time for ten independent trials, and the color of each dot shows the associated average OA. As shown in Fig. 6, with the growth in the number of superpixels in the scene, the running time increases, and more superpixels become pure. For the classification accuracy, the plot rises quickly and reaches a maximum point and then remains relatively stable with only a tiny fluctuation. That is to say, while a desirable classification accuracy calls for oversegmentation of the scene, there is some tradeoff between the number of superpixels in the scene and the computational cost, which is due to the complexity of the land-cover distribution.

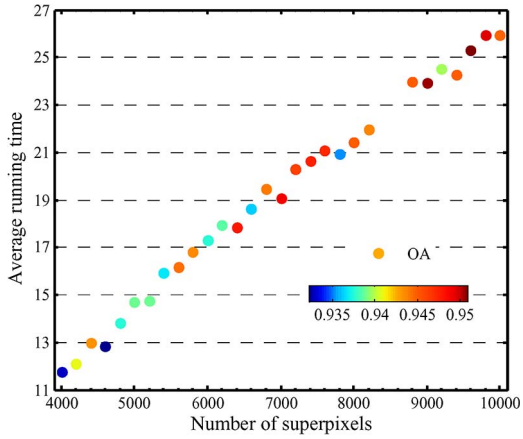


Fig. 6. Classification results versus different superpixel scales for the Washington DC image.

TABLE V
CLASSIFICATION APPROACHES IN THE COMPARISON FOR THE ANALYSIS OF THE ROLE OF THE SUPERPIXEL CR

Acronym	ISCRC-VS	MTLJSRC-C
Similarity	1) Same segmentation step. 2) Same classification rule as SMTLJSRC	1) Same segmentation step. 2) Same classification rule as JCRC-MTL
Dissimilarity	1) It relaxes the class-level sparsity constraint in (9) with a non-sparsity ℓ_2 norm induced regularization, and the details can be found in [38] 2) Vector stacking multiple features of a superpixel 3) Each feature is separable, and there is no interaction between features	1) Stacking pixels in the segmentation parcel as a set, it is a matrix form of MTLJSRC, which can be found in [39], and each pixel contributes equally in the classification step 2) Pixel set joint representation based multi-task learning

Although the oversegmentation is effective, it is believed that both mixed superpixels and pure superpixels can exist in the hyperspectral scene, in practice, where the former are made up of several classes of pixels, and the latter consist of pixels from a single class. To allow a comparison with the proposed algorithm, two benchmark classifiers were utilized, and a brief introduction to them can be found in Table V.

To demonstrate the performance of the proposed classifier, some representative superpixels under different scales

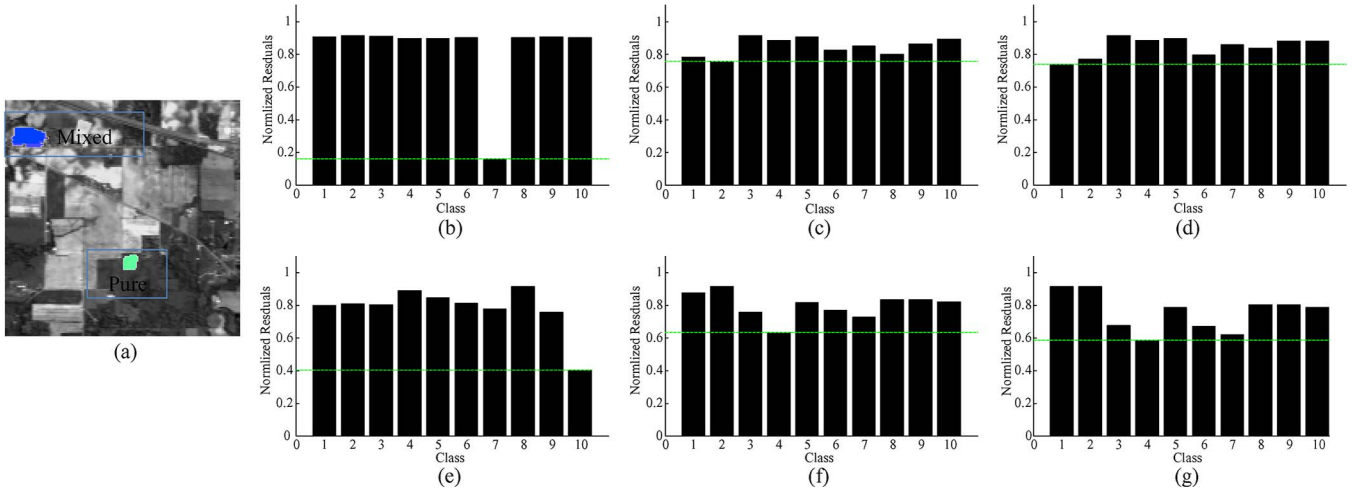


Fig. 7. Normalized residuals for each class for two superpixels in the Indian Pines image. Detailed locations and labels of the involved pixels are shown in (a), where the colors are consistent with the legend in Fig. 3. The upper subfigures represent a mixed superpixel, with the majority of the pixels belonging to class 1, and limited pixels belonging to class 2 and class 10. The lower subfigures represent a pure superpixel, which belongs to class 4. (b) ISCRC-VS for the mixed superpixel, (c) MTLJSRC-C for the mixed superpixel, (d) SMTLJSRC for the mixed superpixel, (e) ISCRC-VS for the pure superpixel, (f) MTLJSRC-C for the pure superpixel, and (g) SMTLJSRC for the pure superpixel.

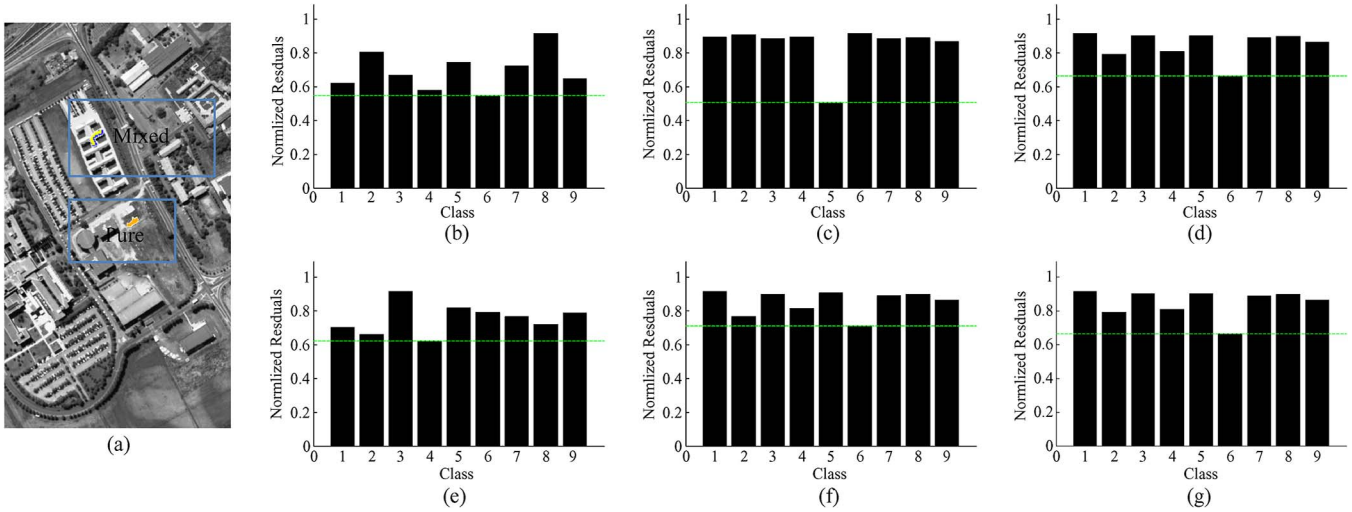


Fig. 8. Normalized residuals for each class for two superpixels in the Pavia University image. Detailed locations and labels of the involved pixels are shown in (a), where the colors are consistent with the legend in Fig. 4. The upper subfigures represent a mixed superpixel, with the majority of the pixels belonging to class 6, and limited pixels belonging to class 5. The lower subfigures represent a pure superpixel, which belongs to class 9. (b) ISCRC-VS for the mixed superpixel, (c) MTLJSRC-C for the mixed superpixel, (d) SMTLJSRC for the mixed superpixel, (e) ISCRC-VS for the pure superpixel, (f) MTLJSRC-C for the pure superpixel, and (g) SMTLJSRC for the pure superpixel.

(i.e., we manually set the undersegmentation and oversegmentation cases) for each HSI were extracted as examples. For the residual-related subfigures in Figs. 7–9, the horizontal axis indicates the class number, the vertical axis shows the corresponding residual of the algorithms, and the green dotted line refers to the smallest normalized residual of the classes. In each figure, the upper subfigures represent a mixed superpixel, and the lower subfigures represent a pure superpixel. All the parameters of these classifiers were optimized by tenfold cross-validation.

For the examples in the homogenous Indian Pines image, the upper superpixel is a mixture of corn-notill pixels (the first class) with rare highly mixed pixels without label in the reference map, and the superpixel belonging to class 4 is a pure grass-trees pixel. As the corn-notill pixels dominate the mixed superpixel, labeling it as neither class 2 nor class 10 is more appropriate. It is shown in Fig. 7(b)–(d) that the proposed algorithm and MTLJSRC-C can successfully fulfill the labeling task, whereas the ISCRC-VS method fails. For the pure instance, similar results can also be found in Fig. 7(b)–(d). The performances of the superpixels in the Pavia University data set are similar to those of the superpixels in the Indian Pines image. In view of the residual distribution, the superiority of the proposed approach is more significant. That is to say, compared with the other two classifiers, it is believed that both the class-level sparsity for multiple features and the superpixel CR can improve the performance of the sparse (collaborative) representation-based classifier.

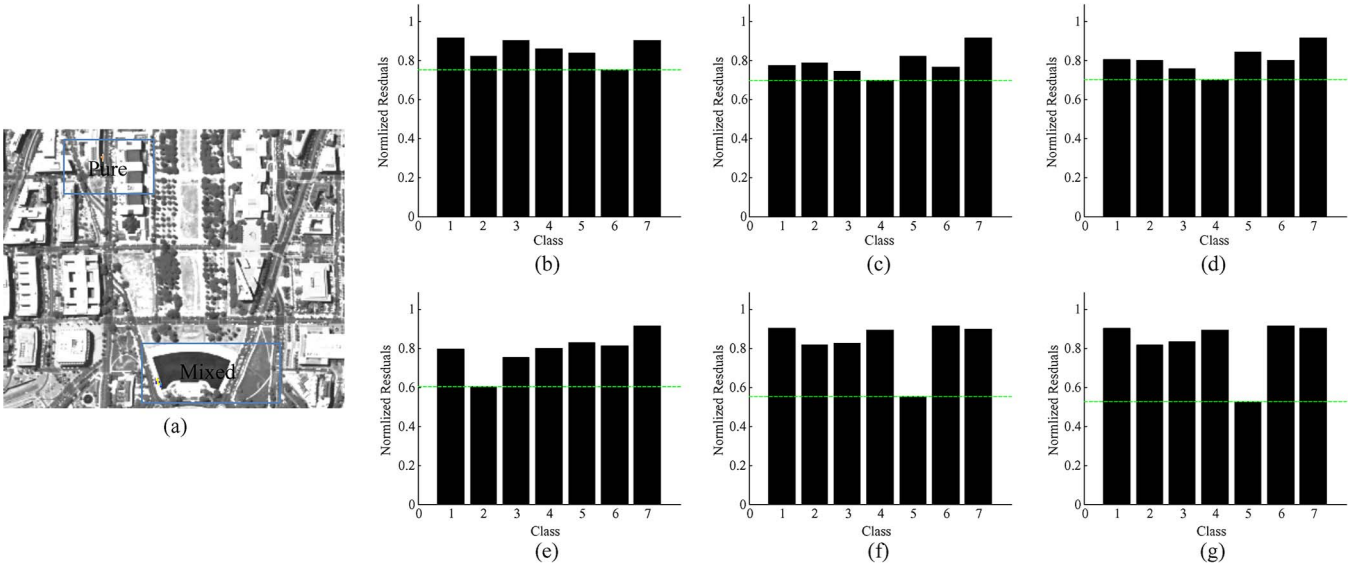


Fig. 9. Normalized residuals for each class for two superpixels in the Washington DC image. Detailed locations and labels of the involved pixels are shown in (a), where the colors are consistent with the legend in Fig. 5. The upper subfigures represent a mixed superpixel, with the majority of the pixels belonging to class 4, and limited pixels belonging to class 1. The lower subfigures represent a pure superpixel, which belongs to class 5. (b) ISCRC-VS for the mixed superpixel, (c) MTLJSRC-C for the mixed superpixel, (d) SMTLJSRC for the mixed superpixel, (e) ISCRC-VS for the pure superpixel, (f) MTLJSRC-C for the pure superpixel, and (g) SMTLJSRC for the pure superpixel.

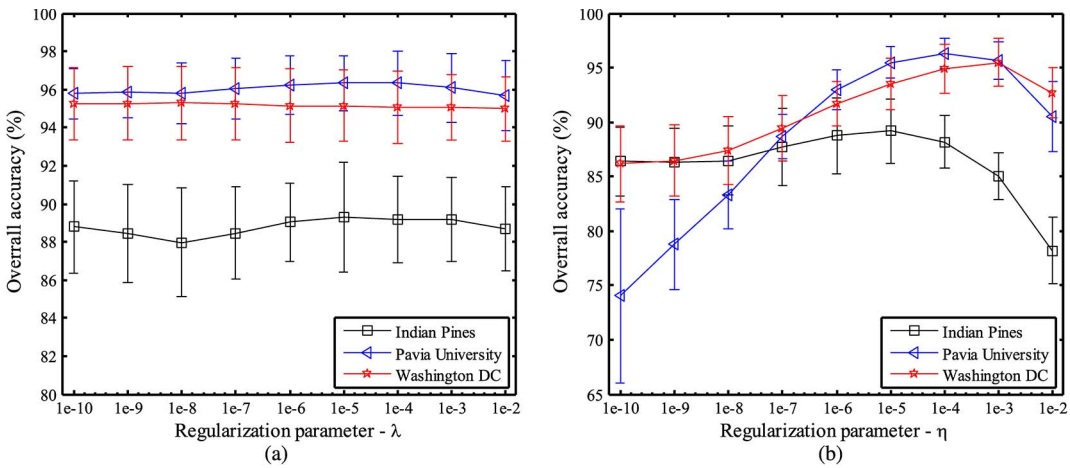


Fig. 10. Classification accuracy versus the regularization parameters of the proposed approach, with the three HSIs: (a) parameter λ and (b) parameter η.

To sum up, several conclusions can be made from the classification results of the three hyperspectral data sets. First, in view of the recognition accuracy, it is believed that the proposed superpixel-level sparse representation-based multitask learning is more effective than the other multiple-feature combination techniques. The spatial prior that is utilized in the superpixel segmentation approach is more stable and superior than the contextual prior used in the current pixel representation-based multitask learning classifiers. Second, for the time complexity issue, it can be seen that the proposed superpixel-oriented approach is approximately comparable to the MLR-related multiple-feature learning approach, and is much more efficient than the pixel-level representation multitask learning approach. Note that, as the SVM-related classifiers were implemented with the help of the LibSVM [52] package, which utilizes C++ software to speed up the process, they should not be included in the comparison, for fairness. Third, there is some tradeoff between the computational cost and the scale of the segmentation, which is closely related to the complexity of the land cover.

E. Parameter Analysis

The final subsection examines the effect of the parameters on the classification performance of the proposed algorithm, with the three HSIs. We fixed the other parameters as the corresponding optimal and focused on one specific parameter at a time. The experiments for λ and η were repeated ten times using different randomly chosen training sets to reduce the possible bias induced by the random sampling. The horizontal axis shown in each subfigure in Fig. 10 is the value range of the

corresponding parameter, while the vertical axis shows the OA (%) of the different data sets.

In Fig. 10(a), it can be observed that the performances of all the hyperspectral data sets are stable with some small variations as the regularization parameter λ increases. Since λ makes a tradeoff between the data fidelity term and the prior term of the superpixel adaptive representation, it is demonstrated that the proposed algorithm is quite robust to this parameter over a wide range of values.

In Fig. 10(b), the experimental results for all the data sets generally improve slowly as the regularization parameter η increases, and then begin to decrease after the maximum value. It is found that once η exceeds a certain threshold, the final class-level sparsity term will dominate the optimization, and the discriminative power of the classification rule that minimizes the representative residual of a single class will be weakened, as the data fidelity term may work in vain.

IV. Conclusion

In this paper, we have proposed an efficient SMTJSRC algorithm for hyperspectral imagery. In the proposed approach, an HSI superpixel is represented by an adaptive combination of the pixels in a parcel, and class-level sparsity is utilized to simultaneously integrate the multiple features into a uniform classification framework. The main advantage of the proposed SMTLJSRC is that the superiorities of the multiple-feature combination approach, the spatial prior utilization, and the computational complexity can be maintained at the same time. The extensive experimental results clearly indicate that the proposed method can achieve competitive classification results.

However, the proposed multitask learning framework could still be further improved in certain aspects. For instance, the segmentation step is independent of the following multitask learning, and the current framework only considers a linear model. Therefore, our future work will focus on how to build a uniform framework to simultaneously cover the nonlinear segmentation, extraction, representation, and recognition tasks.

Acknowledgment

The authors would like to thank Prof. D. Landgrebe from Purdue University, for providing the AVIRIS image of Indian Pines and the HYDICE image of Washington DC Mall; Prof. Gamba from the University of Pavia, for providing the ROSIS data set; and the handling editor and anonymous reviewers, for their careful reading and helpful remarks.

References

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

[2] M. T. Eismann, A. D. Stocker, and N. M. Nasrabadi, "Automated hyperspectral cueing for civilian search and rescue," *Proc. IEEE*, vol. 97, no. 6, pp. 1031–1055, Jun. 2009.

[3] P. K. Goel *et al.*, "Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn," *Comput. Electron. Agric.*, vol. 39, no. 2, pp. 67–93, May 2003.

[4] B. Datt, T. R. McVicar, T. G. Van Niel, D. L. B. Jupp, and J. S. Pearlman, "Preprocessing EO-1 Hyperion hyperspectral data to support the application of agricultural indexes," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1246–1259, Jun. 2003.

[5] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, Jun. 2011.

[6] F. Jie, L. C. Jiao, X. Zhang, and T. Sun, "Hyperspectral band selection based on trivariate mutual information and clonal selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4092–4105, Jul. 2014.

[7] H. Su, Q. Du, G. Chen, and P. Du, "Optimized hyperspectral band selection using particle swarm optimization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2659–2670, Jun. 2014.

[8] L. Nam Hoai, Q. Du, and J. E. Fowler, "Sparse graph-based discriminant analysis for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 3872–3884, Jul. 2014.

[9] J. Wang and C.-I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, Jun. 2006.

[10] Q. Zhang, L. Zhang, Y. Yang, and L. Weng, "Local patch discriminative metric learning for hyperspectral image feature extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 3, pp. 612–616, Mar. 2014.

[11] D. Lunga, S. Prasad, M. M. Crawford, and O. Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning," *IEEE Signal Process Mag.*, vol. 31, no. 1, pp. 55–66, Jan. 2014.

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

[13] 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.

[14] 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, Apr. 2013.

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

[16] L. Zhang, D. Tao, and X. Huang, "On combining multiple features for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 879–893, Mar. 2012.

[17] X. Huang and L. Zhang, "A multilevel decision fusion approach for urban mapping using very high-resolution multi/hyperspectral imagery," *Int. J. Remote Sens.*, vol. 33, no. 11, pp. 3354–3372, 2012.

[18] J. Li *et al.*, "Multiple feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1592–1606, Mar. 2014.

[19] 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.

[20] 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.

[21] J. Li, H. Zhang, L. Zhang, X. Huang, and L. Zhang, "Joint collaborative representation with multitask learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5923–5936, Sep. 2014.

[22] J. A. Gualtieri and R. F. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Proc. 27th AIPR Workshop, Adv. Comput.-Assist. Recog.*, 1999, pp. 221–232.

[23] F. A. Mianji and Y. Zhang, "Robust hyperspectral classification using relevance vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2100–2112, Jun. 2011.

[24] A. Singla, S. Patra, and L. Bruzzone, "A novel classification technique based on progressive transductive SVM learning," *Pattern Recog. Lett.*, vol. 42, pp. 101–106, Jun. 2014.

[25] 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.

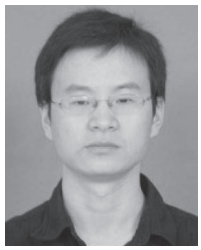
[26] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3947–3960, Oct. 2011.

[27] 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.
[28] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.
[29] J. Li, H. Zhang, and L. Zhang, "A nonlinear multiple feature learning classifier for hyperspectral images with limited training samples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 2015, to be published.
[30] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image classification using soft sparse multinomial logistic regression," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 318–322, Mar. 2013.
[31] P. Gurram and K. Heesung, "Sparse kernel-based ensemble learning with fully optimized kernel parameters for hyperspectral classification problems," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 787–802, Feb. 2013.
[32] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral-spatial classification of hyperspectral images based on hidden Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2565–2574, May 2014.
[33] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[34] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Sparse representation for target detection in hyperspectral imagery," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 3, pp. 629–640, May 2011.
[35] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang, "Collaborative representation based classification for face recognition," unpublished paper, 2012. [Online]. Available: http://arxiv.org/abs/1204.2358
[36] 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 Sens.*, vol. 7, no. 6, pp. 2056–2065, Jun. 2014.
[37] S. Yang, H. Jin, M. Wang, Y. Ren, and L. Jiao, "Data-driven compressive sampling and learning sparse coding for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 479–483, Feb. 2014.
[38] P. Zhu, W. Zuo, L. Zhang, S. C.-K. Shiu, and D. Zhang, "Image set-based collaborative representation for face recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 7, pp. 1120–1132, Jul. 2014.
[39] 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.
[40] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, no. 49 no. 10, pp. 3973–3985, Oct. 2011.
[41] J. Li, H. Zhang, Y. Huang, and L. Zhang, "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.
[42] M. Liu, O. Tuzel, S. Ramalingam, and R. Chellapa, "Entropy rate superpixel segmentation," in *Proc. IEEE CVPR*, 2011, pp. 2097–2104.
[43] R. Achanta *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[44] X. Chen, W. Pan, J. T. Kwok, and G. Carbonell, "Accelerated gradient method for multi-task sparse learning problem," in *Proc. 9th IEEE ICDM*, 2009, pp. 746–751.
[45] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *J. Mach. Learn. Res.*, vol. 6, pp. 2049–2073, 2005.
[46] 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. 2012.
[47] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[48] K. P. Bennett and E. J. Bredensteiner, "Duality and geometry in SVM classifiers," in *Proc. ICML*, 2000, pp. 57–64.
[49] D. Burges and C. Crisp, "A geometric interpretation of v-SVM classifiers," *Adv. Neural Inf. Process. Syst.*, vol. 12, no. 12, pp. 244, 2000.
[50] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statist. Comput.*, vol. 20, no. 2, pp. 231–252, Apr. 2010.
[51] X. Huang, Q. Lu, L. Zhang, and A. Plaza, "New postprocessing methods for remote sensing image classification: A systematic study," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7140–7159, Nov. 2014.
[52] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, Apr. 2011.

Jiayi Li (S'13) received the B.S. degree from Central South University, Changsha, China, in 2011. She is currently working toward the Ph.D. degree with 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, computation vision, and pattern recognition in remote sensing images.

Ms. 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 LETTERS, and the *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 has published more than 30 research papers. His current research interests focus on image reconstruction, hyperspectral image processing, sparse representation, and low-rank methods for sensing image imagery.

Dr. Zhang is a reviewer of about ten international academic journals.

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 the 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 Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is also a Changjiang 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 410 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 Program, and an Executive Member of the China Society of Image and Graphics. He regularly serves as a Cochair of the SPIE Series Conferences on Multispectral Image Processing and Pattern Recognition, the 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. He was a recipient of the 2010 Boeing Best Paper Award and the 2013 Earth Resource Data Analysis System (ERDAS) Best Paper Award from the American Society of Photogrammetry and Remote Sensing (ASPRS).