Hyperspectral Image Classification via Multitask Joint Sparse Representation and Stepwise MRF Optimization

Yuan Yuan, *Senior Member, IEEE*, Jianzhe Lin, *Student Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Hyperspectral image (HSI) classification is a crucial issue in remote sensing. Accurate classification benefits a large number of applications such as land use analysis and marine resource utilization. But high data correlation brings difficulty to reliable classification, especially for HSI with abundant spectral information. Furthermore, the traditional methods often fail to well consider the spatial coherency of HSI that also limits the classification performance. To address these inherent obstacles, a novel spectral–spatial classification scheme is proposed in this paper. The proposed method mainly focuses on multitask joint sparse representation (MJSR) and a stepwise Markov random filed framework, which are claimed to be two main contributions in this procedure. First, the MJSR not only reduces the spectral redundancy, but also retains necessary correlation in spectral field during classification. Second, the stepwise optimization further explores the spatial correlation that significantly enhances the classification accuracy and robustness. As far as several universal quality evaluation indexes are concerned, the experimental results on Indian Pines and Pavia University demonstrate the superiority of our method compared with the state-of-the-art competitors.**

***Index Terms*—Hyperspectral image (HSI) classification, Markov random field (MRF), multitask, sparse representation.**

I. Introduction

HYPERSPECTRAL cameras are designed for collecting hyperspectral images (HSIs) with narrow continuous spectral bands [1]. Because of its wide spectral range and high spectral resolution, the acquired HSIs contain rich discriminative physical clues to pinpoint the ground objects laying on the observed surface. This makes the hyperspectral imagery suitable for land cover classification.

Existing work toward HSI classification [2], [3] has been extensive. However, there are spaces to improve the performance. In this paper, the main classification procedure is the same as the traditional ones. But different from previous methods, the proposed one puts more emphasis on bridging the gap between the benefits of the high-dimensional data and redundancy of the correlated bands. Moreover, for HSI which is a 3-D cube, improving the effective utilization of band spectra is not enough. The spatial coherence is with the same importance. Therefore, the enhancement of spatial correlation of HSI classification result is also attached great significance in this paper.

A. Related Work

In remote sensing, existing classification methods mainly concentrate on two fields: 1) the feature extraction and 2) classifier construction. For a given data set, it is always difficult to find the best combination of the feature and the classifier, for the reason that the different categories of HSI may share quite similar spectral signatures and identical materials may have different signatures. Therefore, for different HSI, finding the most suitable feature and classifier simultaneously is not a simple task.

To extract the most essential feature, many methods have been proposed recently. Most of the state-of-the-arts are based on basic methods such as principle component analysis [4], [5], discrete wavelet transform [6], and independent component analysis [7] methods. The common point of these methods is that they all exploit spectral signatures to transform the initial HSI to reduced data set. Recent advance comes from the joint sparse representation (JSR)-based feature extraction method [8]. The main idea of them is to use different kinds of features to represent the initial image jointly. The applications at the very first mainly concentrate on natural image processing including gait recognition [9], image annotation [10], and face recognition [11]. The first work which uses JSR for visual classification is shown in [12] that casts the feature combination to a multitask JSR (MJSR) problem. However, the model complexity is the main problem for this framework. Therefore, the multiple tasks are always projected

Manuscript received April 15, 2015; revised July 1, 2015; accepted September 27, 2015. Date of publication October 15, 2015; date of current version November 15, 2016. This work was supported in part by the National Basic Research Program of China (Youth 973 Program) under Grant 2013CB336500, in part by the State Key Program of the National Natural Science of China under Grant 61232010, in part by the National Natural Science Foundation of China under Grant 61172143, Grant 61379094, and Grant 61105012, in part by the Natural Science Foundation Research Project of Shaanxi Province under Grant 2015JM6264, in part by the Fundamental Research Funds for the Central Universities under Grant 3102014JC02020G07, and in part by the Open Research Fund of the Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences. This paper was recommended by Associate Editor X. He. *(Corresponding author: Qi Wang.)*

Y. Yuan and J. Lin are with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: yuany@opt.ac.cn; linjianzhe@opt.cn).

Q. Wang is with the School of Computer Science and the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@nwpu.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/TCYB.2015.2484324

to a discriminative subspace manifold regularization to reduce the time cost [13]. The theory of information bottleneck is also introduced to formulate the multitask problem as encoding a communication system with multiple senders that finds a tradeoff between the accuracy and complexity of the multiview model [14]. Then, a kernel sparse multitask learning method specializing in HSI processing is proposed in [15], in which various extracted features of HSI are viewed as different modalities, and all the features work jointly by sparse representation to achieve the final classification. A similar work which combines the spectral, gradient, shape, and texture features together to jointly represent the initial HSI is also proposed in [16]. However, we do not think the integration of these features [17] can achieve the best performance. In these methods, too much redundant information is included for the reason that the different features are all extracted on the same data though sparse framework is introduced. Moreover, classification with the integration of so many traditional features is not efficient. On this point, compared with the feature extraction methods above, band selection method which only extracts a subset of bands to represent the initial HSI may have advantage. And we would introduce the JSR framework to band selection in this paper.

The other task of HSI classification is classifier construction based on the extracted features. The classifier can be roughly divided into three categories: 1) unsupervised; 2) supervised; and 3) semisupervised. *A priori* knowledge brought in the learning phase is the most distinctive difference among them. In most time, support vector machine (SVM) [18] as a robust supervised classifier demonstrates its superiority among a majority of classifiers. This method finds the optimal hyperplane between two categories to address a binary classification problem. But it is found that introducing additional spatial constraint will bring more excellent classification performance [19], [20]. Ji *et al.* [21] proposed a hypergraph-based spectral–spatial classification method that constructs both spectral- and spatial-based hypergraph, on which the probability of pixels belonging to different categories is learned. Another spectral–spatial method is proposed in [22] that selects the labels of pixels by edge preserving method. Both the two methods conduct the classification process by exploiting the spatial features of neighboring pixels. However, the most popular one of these spatial features is exploited in the neighborhood system by using Markov random fields (MRFs) [23]. MRF as a statistic modeling tool effectively incorporates the spatial characteristic into the classification process under Bayes inferring framework. An SVM- and MRF-based classification method can be found in [24] that combines these two as an integrated framework to conduct a contextual HSI classification. But the edge information used for computing the spatial energy in MRF framework is not effective enough that an adaptive MRF approach is proposed [25], in which the weighting coefficient of MRF classification is dependent on the a relative homogeneity index of each pixel. Problems remain in MRF for the reason that the wealth of spectral information in hyperspectral data cannot often be complemented by extremely fine spatial resolution. This phenomenon may lead to the problem of mixed pixels. To overcome this drawback, Li *et al.* [26] combined subspace projection and MLR to separate these classes. Another limitation of MRF is that these MRF-based methods only concentrate on the spatial prior with the discrete-valued labels, which leads to a hard discrete optimization problem. In [27], this problem is generally solved by utilizing a set of hidden real-valued fields. However, weight imbalance problem still exists that provides potential of improvement for MRF to exploit contextual information of HSI.

B. *Limitation of Existing Methods*

Although many hyperspectral classification methods which take both spatial and spectral information into consideration are presented in this paper, the limitations of these methods are not concluded. We want first to summarize these limitations.

1) Poor Band Usage in HSI: The sufficient spectral bands of HSI yield the potential to complete the classification process gracefully. However, the classification result is barely satisfactory, owing to the fact that the high interband correlation will pull down the discrimination among pixels. Traditional HSI classification always introduces dimensionality reduction method to conquer this drawback, but the quality of the remaining bands is undetermined. In other word, we cannot always summarize the most valuable spectral information we need from the original image. It has been demonstrated alternatively [28], [29] that classification with selected bands which are highly uncorrelated may bring better result. But the inevitable information loss will still affect the result severely.

2) Weight Imbalance in MRF Framework: As shown in recent work for HSI classification [30], MRF is an effective framework to introduce the spatial connection of HSI pixels. However, we have to emphasize that MRF framework cannot always distribute the most proper weight to spectral and spatial energy terms. In different area of an HSI, the importance of these two kinds of clues are not the same, which means the weight between them should change from one area to another. But unfortunately, existing work does not consider this aspect.

To overcome the aforementioned drawbacks, two main works are proposed in this paper. For the first problem, we strick a balance between limited band use and acquisition of useful information. To achieve this, we construct several tasks, each of which includes limited number of bands with low correlation, and make them work jointly. The second work is stepwise strategy for MRF optimization. It effectively avoids the traditional setting of the balance parameter between the two energy terms.

C. *Contribution*

Two main contributions are claimed in this paper. They are listed as follows.

1) Multitask Joint Sparse Representation on HSI: In order to use a limited number of uncorrelated bands and maintain their interactive relationship, the MJSR is explored [31]–[33] in this paper. The method utilizes several sets of features simultaneously and translates them into tasks to complete the classification process. In the preparation phase, we make use of cluster method to divide bands into several sets, and select

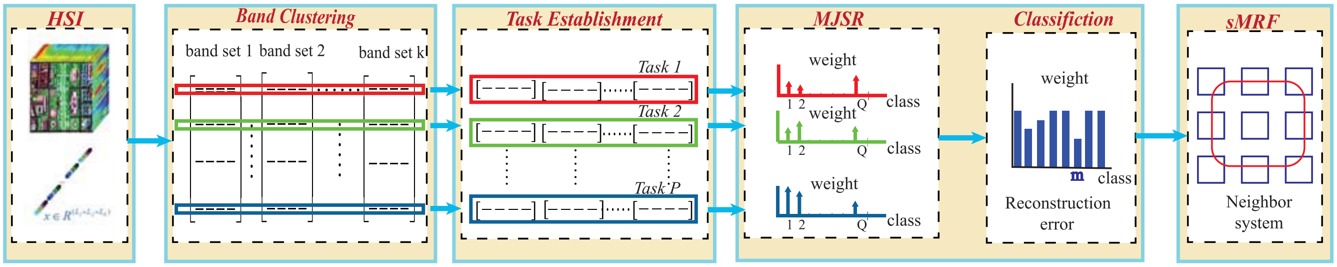


Fig. 1. HSI classification pipeline. For an input HSI, the first step is to divide the spectral bands into different band sets, and the bands in one set have similar attributes. Then, based on the obtained band sets, several tasks are established, which are used for further multitask representation. After that, we reconstruct a test sample with a JSR of tasks and its class label is inferred with the lowest reconstruction error. Finally, the obtained results are refined by sMRF optimization.

bands from clusters to construct tasks. Then all the tasks are integrated into a whole union to work together and find the best solution. We choose the accelerated proximal gradient (APG) method [34], [35] to optimize the classification model and find the most appropriate parameters to synthesize the tasks. The presented strategy can maximize the usage of bands and is the first attempt in HSI classification, as far as we know.

2) Stepwise MRF Framework: To keep the spatial coherence of the classification results in HSI, MRF framework [30] is a traditional solution. It defines an energy function containing a data term and smoothness term and through its minimization, the best labeling (classification) result is obtained. Generally, the two functional terms are always optimized together, which cannot always maximize the usage of spectral and spatial information simultaneously. In this paper, we find that if we optimize these two parts one after another, a better result can be obtained. This stepwise MRF (sMRF) not only improves the classification result of our method, but also demonstrates its applicability in other classification paradigms.

The remainder of this paper is organized as follows. In Section II, the proposed work is described in detail, including band clustering, task establishment, joint classification with MJSR, and sMRF optimization. In Section III, experimental results, as well as a comprehensively qualitative and quantitative comparison and analysis, are presented. Finally, we conclude this paper in Section IV.

II. HSI Classification

In this paper, we focus on conducting the classification by specific bands. There are mainly four steps for this procedure: 1) band clustering; 2) task establishment; 3) joint classification with MJSR; and 4) sMRF optimization. The general flowchart is shown in Fig. 1.

First, since the neighboring bands of HSI tend to be similar, we tactfully divide them into several highly uncorrelated band sets. This is the prerequisite for further task establishment. For this purpose, a modified *k*-means method is designed to get a cluster result that can be repeated. After the process of band clustering, we define multiple tasks to fulfill the classification, each of which contains a few selected bands representing the whole band sets and can perform the classification task individually. But treating them independently is not appropriate because the correlation and redundancy among the bands are not well modeled. Therefore, an MJSR strategy is explored to simultaneously solve the classification of different tasks. In the end, a stepwise optimization is applied to refine the obtained results.

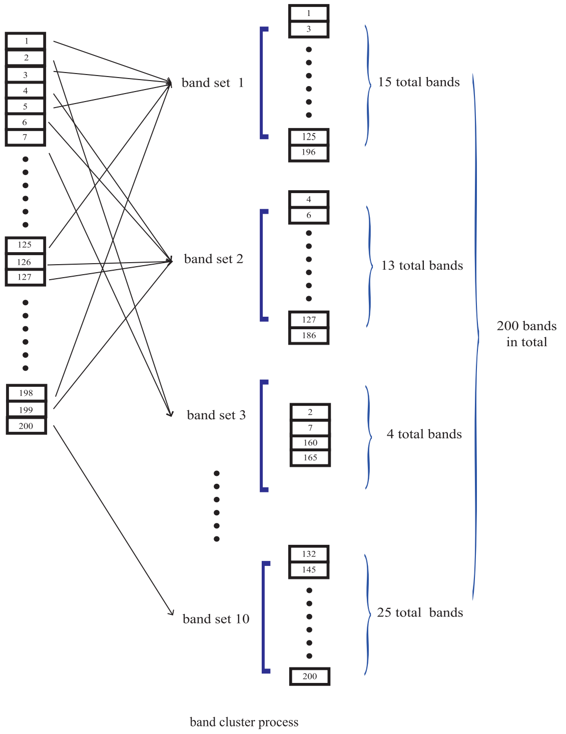


Fig. 2. Illustration of band clustering for the Indian Pines image. Suppose the $n = 200$ original bands are clustered into $k = 10$ band sets. Since the clustering procedure is kind of random that the volume of each set differs from each other. Heuristically, the band set with smaller volume has stronger ability of discrimination.

A. Band Clustering

For HSI, we always find that the high interband correlations affect the classification result a lot, especially for the neighboring bands. Therefore, we need to select the bands with high discrimination to complete the classification process. To achieve this purpose, the first step can be viewed as a preparation step, in which we divide the bands into different sets according to their similarities. Bands with high similarities will be arranged to the same cluster, and only representative of every cluster will be selected in the next step. To be

Algorithm 1 Modified k-Means

```
Input: $X = \{x_1, x_2, \dots, x_n\}$, $R$, $k$
$\mu_x \leftarrow$ CENTERINITIALIZATION($X$,$R$,$k$)
Removed isolated points returns
Initialize $C^x$ through clustering centering on $\mu_x$
Calculate the new cluster centers and re-cluster to get $C^x_{new}$
while $C^x_{new} \neq C^x$ do
    $C^x \leftarrow C^x_{new}$
    Calculate the new cluster centers and re-cluster to get $C^x_{new}$
end while
Output: $C^x$

function CENTERINITIALIZATION($X$,$R$,$k$)
    Remove isolated points
    while The total number of clusters $N \neq k$ do
        if $k - N > N$ then
            For each current cluster, find $\mu_x^1$ and $\mu_x^2$ with maximum Euclidean distance shown in $R$
            Divide each current cluster to two centering on $\mu_x^1$ and $\mu_x^2$
            $N \leftarrow 2N$
        else
            For the first $k-N$ clusters, find $\mu_x^1$ and $\mu_x^2$ with maximum Euclidean distance shown in $R$
            Divide the $k - N$ cluster to two centering on $\mu_x^1$ and $\mu_x^2$
            N = k
        end if
    end while
    return $\mu_x$
end function
```

more specific, we cast every band of the image to a vector of data, and then cluster them to band sets. Moreover, different sets have different volume, which reflects its significance. Heuristically, the band set with smaller volume has stronger ability of discrimination. A more detailed illustration is shown in Fig. 2.

As for the clustering technique, a modified k-means is designed. The motivation is mainly because of the unrepeated clustering results for distinct initializations of traditional k-means. To solve this problem, we select the initial cluster centers as the points with maximum Euclidean distances. To be more specific, suppose the initial bands is $X = \{x_1, x_2, \dots, x_n\}$, we first eliminate the outliers in the data points [36], [37], which might be some isolated noises [38] disorderly distributed at the edge of the picture, and will cause inaccuracy. Then we calculate the affinity matrix R, in which the elements represent the Euclidean distances between each two points. The two bands μ_x^1 and μ_x^2 with the maximum Euclidean distances shown in R are assigned as the two initial clustering centers, and the other bands are divided into two parts according to their distances to the two centers. Then for each cluster, the division continues with the same principle until the number of clusters N reaches the required volume k. This center initialization not only ensures the initial dissimilarity of the chosen centers, but also makes the clustering procedure more stable. The other important setting lies in the choice of band set number k. Smaller size will make the task separation less desirable; larger size will induce a decreased discrimination and high computational cost. Suppose the final clustering result is represented by C^x, a more detailed pseudocode is shown in Algorithm 1. Detailed discussion of this parameter will be introduced in Section III-B2.

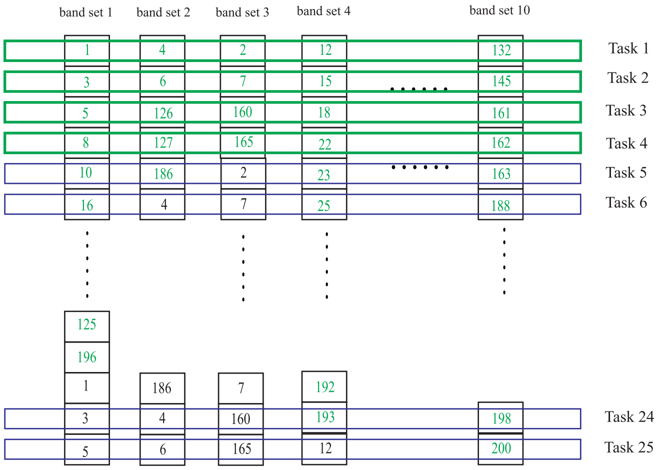


Fig. 3. Illustration of task establishment. Followed from Fig. 2, there are $t = 25$ possible tasks but only $P = 4$ tasks are selected as the final ones, indicated by the green rectangles.

B. Task Establishment

It has been demonstrated that for HSI classification, if a suitable number of bands are selected, the result of classification would be more accurate than we just make use of all the n bands [28], [29]. The reason is that the data after band selection is with lower correlation and has more discriminative ability. Inspired by this fact, we just select a limited number of bands to conduct the classification.

After the data are divided into k clusters in the previous step, we will define t tasks to fulfill the classification objective. Each task contains a batch of elements chosen from the band sets, one from each. That means every task contains all the representative information from the HSI and it is adequate to perform the classification. To be specific, suppose the volume of each cluster set is denoted as $c_1, c_2, \dots, c_k$ and consequently $t = \max\{c_1, c_2, \dots, c_k\}$. In real applications, too many tasks may lead to a high computational cost in the next JSR step. Therefore, we only select $P = \min\{c_1, c_2, \dots, c_k\}$ tasks from t for efficiency. This is illustrated in Fig. 3.

C. Joint Classification With MJSR

The above two steps can be considered as a grouping strategy for the original bands. In this step, we focus on utilizing the obtained P tasks to accomplish the classification process jointly. This can be divided into four substeps as follows.

1) Dictionary Construction: In this phase, pixels of different ground classes are selected to construct tasks. For each task, since there are Q classes of pixels, we denote it by $X^p = [X_1^p, \dots, X_Q^p], p = 1, 2, 3, \dots, P$, of which $X_q^p \in R^{m_p \times n_q}$.

Here, m_p is number of bands contained in the pth task and n_q is the number of training samples in the qth class. And $n_t = \sum_{q=1}^{Q} n_q$ represents the number of training samples in total. In fact, X_q^p is the dictionary for the class q and task p.

2) Testing Sample Representation: Given a test sample y, we first construct its pth task representation y^p by the linear representation of dictionaries. The reconstruction equation is as follows:

$$y^p = \sum_{q=1}^{Q} X_q^p w_q^p + \xi^p, p = 1, \ldots, P \tag{1}$$

of which w_q^p is the reconstruction coefficient vector and ξ^p is the residual term. Suppose that $w_q = [w_q^1, \ldots, w_q^P]$ represents the coefficient of every task in qth class. Our MJSR can be concluded as the following equation:

$$\min_{W} \frac{1}{2} \sum_{p=1}^{P} \left\| y^p - \sum_{q=1}^{Q} X_q^p w_q^p \right\|_2^2 + \lambda \sum_{q=1}^{Q} \left\| w_q \right\|_2 \tag{2}$$

where $W = [w_q^p]_q^p$. The equation above is a multitask least square regressions with $\ell_{1,2}$ mixed-norm regularization.

3) Parameter Optimization: The goal of this step is to find the most proper representative coefficients $w_q = [w_q^1, \ldots, w_q^P]$ of (2). To this end, we apply the APG method [34], [35] to get the solution. Compared with other existing methods such as the projected subgradient method [39] and the blockwise coordinate descent method [46] which are also applicable for the solution of this equation, APG is with higher convergence rate and learning accuracy. This method introduces the variation in Nesterov's method that calls a black-box oracle in the projection step in each iteration. This projection can be simply solved and the time complexity is reduced greatly. Suppose the iterative number is t. The converge rate of APG is $O(1/t^2)$ compared with $O(1/\sqrt{t})$ of other methods. To be more specific, this method iteratively update the weight matrix sequence $\{\hat{W}^t = [w_q^{p,t}]\}_{t \geq 1}$ and a newly introduced aggregation matrix sequence $\{\hat{V}^t = [v_q^{p,t}]\}_{t \geq 1}$. Each iteration includes two steps as follows, which update $\hat{W}^t$ and $\hat{V}^t$ alternately.

Step 1 (Generalized Gradient Mapping): Given the current aggregation matrix $\hat{V}^t$, then we update $\hat{W}^{t+1}$ according to the following:

$$\begin{aligned} \hat{w}^{p,t+1} &= \hat{v}^{p,t} - \eta \nabla^{p,t}, p = 1, \ldots P \\ \hat{w}_q^{t+1} &= \left[1 - \frac{\lambda \eta}{\left\| \hat{w}_q^{t+1} \right\|_2} \right]_+ \hat{w}_q^{t+1}, q = 1, \ldots Q \end{aligned} \tag{3}$$

of which $\nabla^{p,t} = -(X^p)^T y^p + (X^p)^T X^p \hat{v}^{k,t}$, η is a size parameter and $[\cdot]_+ = \max(\cdot, 0)$.

Step 2 (Aggregation): Now we will update $\hat{V}^{t+1}$ by the linear representation of $\hat{W}^t$ and $\hat{W}^{t+1}$

$$\hat{V}^{t+1} = \hat{W}^{t+1} + \frac{\alpha_{t+1}(1 - \alpha_t)}{\alpha_t} \Big(\hat{W}^{t+1} - \hat{W}^t \Big). \tag{4}$$

It has been demonstrated [12] that the best parameter setting of $\{\alpha_t\}_{t \geq 1}$ is $\alpha_t = 2/(t + 2)$.

4) Final Classification: For the testing sample, we can get the optimal representative coefficients $\hat{w}_q^p$ from the previous steps. And now we assign the label of the test sample by finding the lowest reconstruction error accumulated across all the P tasks. The class can be found by the following equation:

$$q^* = \arg\min_{q} \sum_{p=1}^{P} \theta^p \left\| y^p - X_q^p \hat{w}_q^p \right\|_2^2 \tag{5}$$

where θ^p is the weight we assign to task p, whose numerical value will be discussed in the experimental section. The eventual class label of the test sample is the one with the lowest total reconstruction error.

D. Stepwise MRF Framework for Optimization

The obtained results after previous processing still have some outliers. So we just take the former process as the initialization of the MRF framework [40], [41], which is utilized to refine the classification map. Traditionally, the MRF describes the following energy minimization problem:

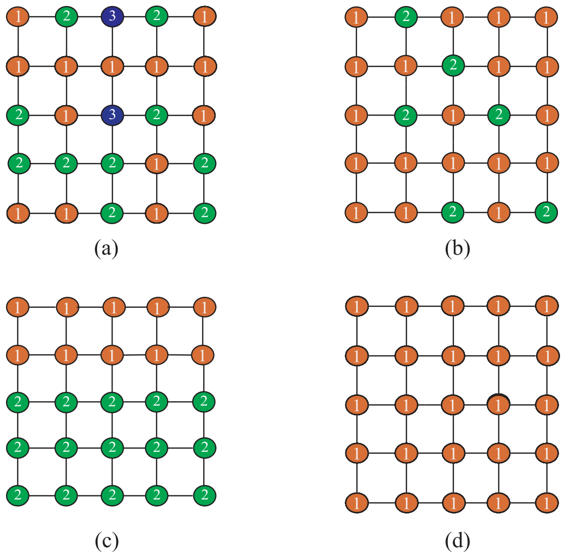
$$E = E_d + \lambda E_s \tag{6}$$

where E_d is the data term representing the likelihood of the objective data, and E_s is the smoothness term reflecting the joint Gibbs distribution [42] of the label field which satisfies the Markov property. By finding the minimum solution of the energy function E, the corresponding label field can be acquired.

However, there is a coefficient λ which is imposed to assign a proper proportion to both the data term and the smoothness term. But in most time, this coefficient is difficult to be determined adaptively. In some area of an image where there exist only one class of pixels, the smoothness term is more important to eliminate the noisy labels. In other areas that different classes exist, the data term counts more. We take the examples as shown in Fig. 4 to illustrate the situation.

The first case in Fig. 4(a) contains three different labels that are mixed together. In this case, the accurate classification is mostly determined by the spectral information. That means the data term should be more decisive than the smoothness term and a small λ is much proper. Otherwise if λ is large, the optimization result may be the case as shown in Fig. 4(c), where both the correct and incorrect labels spread. The second case shown in Fig. 4(b) only consists of one kind of label but is contaminated by sparse noisy labels. The majority work should be smoothing the whole patch to make the results consistent as shown in Fig. 4(d). This corresponds to a larger λ on the contrary. Altogether, we cannot define a most proper coefficient λ in advance to balance the functional two terms.

To solve this problem, we abandon the coefficient λ by a stepwise strategy to separate these two terms. The principle is that the smoothness term and data term should, respectively, take their own effect in accord with the actual configuration of the objective data. To be specific, the smoothness term is first used to make a local optimization, and then we use the data term to get rid of the wrong labels, which can be viewed as a global restriction. With this strategy, an improved framework for MRF is proposed. A revised ISING model is employed



(a) (b) (c) (d)

Fig. 4. Illustration of different λ choices. (a) Initial correct labeling and its (c) corresponding results applying large λ. (b) Initial corrupted labeling and its (d) corresponding results applying large λ.

for the smoothness term, and for the data term, we introduce GMM model [43].

1) Energy Function: In the energy function [44], the smoothness term is typically defined as $E_s = \sum_{a,b \in C} V_c(l_a, l_b)$, where C is the set of cliques in a specific neighborhood system, a and b are the pixels in the image and l_a and l_b are the corresponding labels. Moreover, we treat the pixel with different orders separately by assigning various weights to them. These weights reflect the significance of the neighboring pixels in the system. The ISING model of smoothing term is novelly revised as follows:

$$V_{\{a,b\}}(l_a, l_b) = \begin{cases} \dfrac{-\rho}{n} & \text{if } l_a = l_b \\ \dfrac{+\rho}{n} & \text{if } l_a \neq l_b \end{cases} \tag{7}$$

where n is the order of pixel. Suppose we have pixel x_i and the central pixel x_o, the n-order means the distance between them is n. In fact, this formulation means that the larger the distance between the pixel and the center is, the less effect it owns.

To penalize solutions which are not consistent with the prior knowledge, the data term is introduced. The general form of which is $E_d = \sum_{a \in \mathcal{C}} w_a(l_a)$, where $w_a(l_a)$ shows the cost of giving a label l_a to pixel a. We introduce GMM model in this step, for the reason that the hyperspectral scenery is always formed by several categories. Suppose that there are L classes of pixels and the Gaussian distributions in our GMM model indicate each of these categories. Each distribution is constructed by a set of parameters $\theta_i = \{\mu_i, \sigma_i, z_i\}, i \in \{1, \ldots, L\}$, in which μ_i and σ_i represent the expectation and the variance respectively, and z_i represents the mixing coefficient of the ith category. Expectation-maximization algorithm is introduced iteratively to estimate and update these three parameter. Then, to measure the probability of an observed pixel belonging to each specific class, the data term is formulated as follows:

$$w_a(l_a) = f\left(a|\theta_{l_a}\right) \tag{8}$$

where $f(.)$ is the probability density function of the GMM distribution.

2) Stepwise Optimization: After the definitions above, we give out the stepwise optimization as below.

Step 1: Conduct the local optimization by smoothness term. The initial result obtained by multitask representation is not very robust. Through the minimization $\mathcal{L}_s = \min_{\mathcal{L}} E_s$, we, on one hand, eliminate the isolated class labels first, and on the other hand, get two kind of labels such as the changed and the unchanged.

Step 2: Justify the correctness of the changed labels by the data term. After the first step, we get an updated label field. But the induced label changes are not necessarily appropriate. Besides, labels which remain unchanged might be wrong in the next iteration, either. Therefore, we further check the changed labels through the data term $\mathcal{L}_d = \min_{\mathcal{L}}\{E_{dl_1}, E_{dl_2} + \alpha\}$, where l_1 and l_2 are the initial label and the changed label, respectively, and α is the coefficient that punishes the change process.

Step 3: Check if the energy function becomes a smaller one. If so, restart the iteration from step 1. Otherwise, we reach the final optimization result.

To be noted further, the steps 1 and 2 cannot be exchanged, for the reason that the data term here mainly acts as the judgement of the correctness of the changes brought by smoothness term.

III. Experiments and Analyzes

In this section, experiments are conducted to evaluate the effectiveness of the proposed method. We first introduce the experimental setting and the parameter selection in detail. And then, evaluative analysis is presented.

A. Data Set

We verify the proposed method on two publicly available HSIs: 1) Indian Pines and 2) Pavia University. The description of these images is as follows.

The Indian Pines image was gathered over a vegetation area in northwestern Indiana by AVIRIS sensor, which consists of 145 × 145 pixels and 220 spectral reflectance bands with spatial resolution of 20 m/pixel. These 220 bands include 20 water absorption bands that are not discriminative enough and we conventionally remove them in our experiment. Sixteen classes of interest are contained in this image, of which we select the major nine categories to accomplish our experiment. We choose our training samples in the image randomly, and the count is 10%.

The Pavia University is captured by ROSIS sensor over Pavia, Northern Italy. This image is characterized by spatial resolution of 1.3 m/pixel with 103 spectral bands and comprises 610 × 340 samples including nine classes of interest. For this image, we complete the trial with 10% training samples.

B. Experimental Details

Before detailed analyzing the performance of the proposed method in this paper, the competitors and parameter selection will be introduced in the following part.

1) Competitors: To verify the effectiveness of the proposed MJSR with stepwise MRF optimization (MSMRF), we first compare it with five most typical classification algorithms: 1) SVM [45]; 2) orthogonal matching pursuit (OMP) [46], [47]; 3) kernel logistic regression (KLR) [48], [49]; 4) subspace pursuit (SP) [50]; and 5) k-nearest neighbor (kNN) algorithms [51], [52]. These six classical methods are widely accepted in remote sensing applications. The comparison reveals the superiority of our method.

Moreover, we also compare our method with some other spatial–spectral HSI classification methods, including the state-of-the-arts. The first spatial–spectral method named SVM+ISO data is proposed in [53], which combines the pixel-wise classification map with the segmentation map. The classification map is got by SVM classifier and the spatial relation is achieved by clustering-based method. After the post regularization, the final classification result is obtained by a majority vote based on these two maps. The second spatial–spectral method [54] integrates SVM to the morphological profiles framework. In this method, several morphological profiles are built and used all together in one extended morphological profiles, which illustrates the spatial correlation of HSI gracefully. This method is denoted as Spec-EMP. The third spatial–spectral method [55] is based on stochastic minimum spanning forest approach. The graph-based minimum spanning forest is used to correct the pixel-wise classification map. This method is named as random marker-minimum spanning forest (RD-MSF). Note that since the original papers [54], [55] lack a thorough report on the three HSIs, the comparison of Spec-EMP only exists in Pavia University data and RD-MSF only exists in the Indian Pines data.

2) Parameter Selection: There are a number of critical parameters to be elaborated in our experiment.

The very first parameter lies in the band clustering process, which is the cluster number k. If k is large, the correlation among different sets is high and there will exist few tasks. On the contrary, if k is a small number, the bands within each set will have more correlation and the effect of clustering process is not obvious. This will lead to a high computational complexity. Therefore, in our experiment, we heuristically set $k = 20$.

The second parameter is the task number denoted by P. Too many tasks may boost the computational complexity. But with a limited number of tasks, the effect brought by the interaction among every task may be cut down. In this paper, we set $P = \min\{c_1, c_2, \ldots, c_k\}$ according to the minimum volume of clusters (owing to fact that the result of clustering is not stable, P in most time ranges from $2 \sim 5$. Since there are generally 20 cluster sets, the total representative bands used for classification are $40 \sim 100$. On the both data sets in our experiments, P is heuristically set as 2 for efficiency).

The third parameter θ^p exists in the final classification step of joint classification with MJSR. θ^p represents the weight we put to every task in the process of classification. In most cases, we simply pick $\theta^p = 1/P$ to give every task the same weight, for the reason that all these tasks constructed by bands of the HSI are with equal importance. But they can be assigned with different weights.

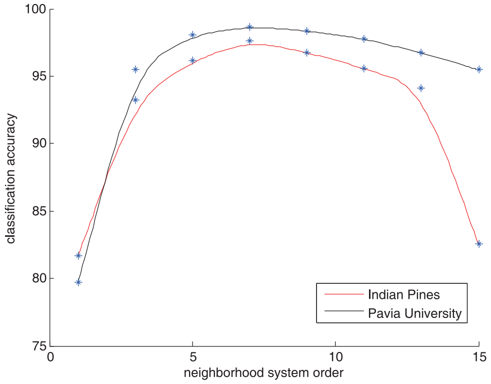


Fig. 5. Experimental comparison of different neighborhood orders.

The fourth and the fifth parameters appear in the optimization process. In the first step of stepwise optimization, we conduct the local optimization in a neighborhood system. However, the size of this neighborhood system influences the final result severely. Suppose the maximum order of neighborhood system we choose is denoted by r. Then we examine the performance of the method with different choices of r. The results on Indian Pines image and Pavia University with 10% training samples are shown in Fig. 5. It is obvious that with the increase in r, the performance is enhanced. But when r exceeds 7, the performance on the contrary decreases. The theoretical analysis is that a reasonable enlargement of neighborhood system may boost the accuracy because it can eliminate the influence of noise. However, too large neighborhood system will introduce too much uncorrelated information that deteriorates the discriminative ability. Therefore, we select $r = 7$ in our experiment. In the second step of stepwise optimization, we punish the change process by coefficient α which is the fifth parameter. This coefficient should avoid two extreme conditions, too large or too small. If α is a large one, the change will be hard to occur. On the contrary, if the change is easy to happen, the final result will almost totally differs from the initial labeling, which means that after the smoothing process, the result of MJSR will be covered. So, in most time, we set α as 1 or 2.

C. Performance Analysis

In this section, a number of experimental results are gained to evaluate our method. Results of both typical methods and the proposed one (MSMRF) are shown in detail, as described below. The main evaluation criteria are the overall accuracy (OA) and the average accuracy (AA).

1) Experiment on Indian Pines: In the very first step, we compare our method with the traditional classification methods on Indian Pines, with 10% training samples [49], [54]. The comparison is illustrated in Fig. 6, which consists of four subfigures, including the results of the proposed MSMRF

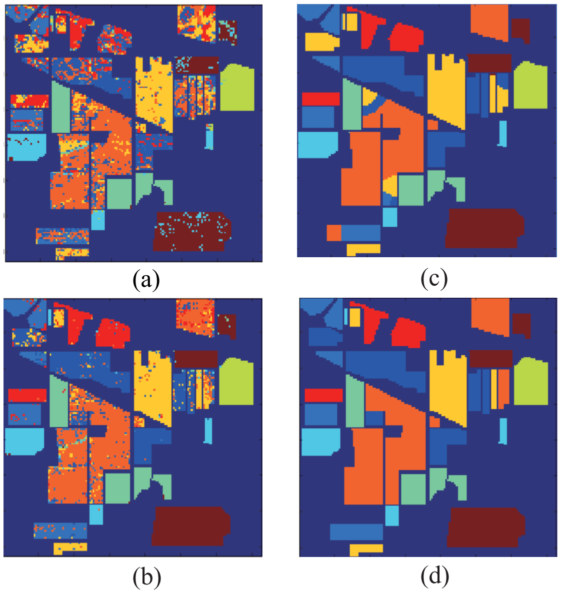


Fig. 6. Results on Indian pine image. (a) kNN classification map, OA = 69.46%. (b) SVM classification map, OA = 79.69%. (c) MSMRF classification map, OA = 92.11%. (d) Ground truth.

method, two most competitive classical methods kNN and SVM, and the ground truth. Fig. 6(a) is got by the traditional kNN classifier and the OA is 69.46%. Fig. 6(b) is achieved by SVM method, for which the OA is 79.69%. Fig. 6(c) is obtained by MSMRF, whose OA achieves 92.11% and is the best among these three methods. At last, Fig. 6(d) is the ground truth.

An objective comparison is shown in Table I, which includes the results of another five comparison methods, SP, OMP, SLK, SVM+ISOdata, and RD-MSF. We may find that almost all the best OA and AA appear in our method. To be more specific, except for the areas of corn-notill, Grass-pasture, and soybeans-mintill on which the best results exist in SVM+ISOdata or RD-MSF method, the other highest scores are all achieved by our method. For several classes, our method has the perfect results which reach 100% accuracy. Moreover, as seen in Table I the main errors concentrate on the classification results of corn-mintill, soybeans-notill, and woods for most methods. Nevertheless, our results are 90.84%, 99.69%, and 100%, which are still satisfying. In the following, we will give a detailed analysis of the illustrated classification map.

As shown in Fig. 6, for traditional kNN and SVM methods, these three types are mainly disrupted by chaotic outliers while the proposed one is with few noises. To account for this phenomenon, it has to be taken into consideration that the minor differences of neighboring pixels cause the wrong labeling, so that the outliers randomly occur with no rules. But in our method, we just select limited bands contained in each task that cut down the interaction of similar bands, which means the confusing bands might have already been got ride of through the phase of MJSR. Moreover, apart from these scattered errors, there are also coherent errors for the results of traditional methods. For example, a high percentage of the area of soybean-mintill is wrongly classified as soybean-notill in the result of kNN method. This type of error is mainly caused by the similarity of two different object bands. But for our method, this wrong labeling does not exist. We handle this problem by sMRF assistance which considers the tendency that the neighboring pixels may have the same labels and the different ground categories have different physical bands.

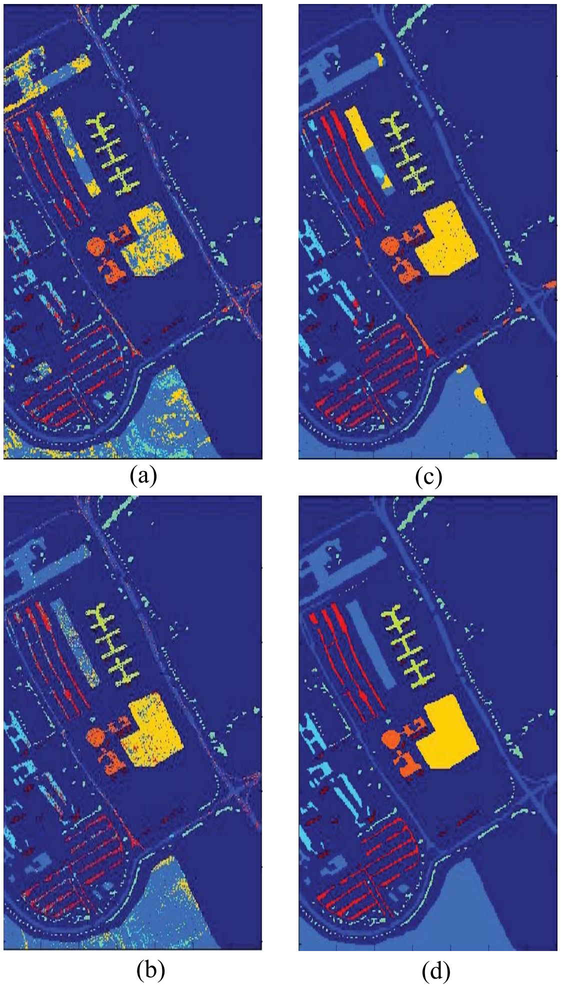


Fig. 7. Results on Pavia University image. (a) kNN classification map, OA = 74.86%. (b) SVM classification map, OA = 79.15%. (c) MSMRF classification map, OA = 92.52%. (d) Ground truth.

2) Experiment on Pavia University: To further demonstrate the proposed method is appropriate for different HSIs, another experiment is conducted on Pavia University image with the same parameters. The visualized results are shown in Fig. 7. It is obvious that the best performance comes from our method, with 95.41% AA and 92.42% OA, respectively. A more objective result including comparison with SP, OMP, SLK, Spec-EMP, and SVM+ISOdata is shown in Table II.

To be more specific, we can see from Table II that the error-prone areas mainly distribute in the Gravel and Meadows regions. This can be proved from Fig. 7(a) and (b), where the Meadows area in the lower portion of the image is severely affected by gravel and bare soil. But in Fig. 7(c), this error almost do not exist. For these two types, our accuracies are 90.56% and 93.55%, respectively, which are all best performances. The main advantage of our method still lies in the elimination of scattered errors and coherent errors. Thanks to

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY FOR DIFFERENT METHODS ON INDIAN PINES

	Pixel-number	KLR	OMP	SP	kNN	SVM	SVM-ISOdata	MD-MSF	MSMRF
Feature number		(206)	(206)	(206)	(206)	(206)	(206)	(206)	(40)
Overall accurary		79.15%	73.27%	74.45%	69.46%	79.69%	91.47%	91.15%	**92.11%**
Average accurary		87.66%	81.65%	82.77%	74.78%	85.34%	91.59%	93.87%	**94.86%**
Classes									
Corn-notill	1428	89.46%	65.97%	74.65%	54.07%	77.79%	80.48%	**97.50%**	82.56%
Corn-mintill	830	70.67%	60.67%	63.20%	55.34%	80.14%	88.02%	81.40%	**90.84%**
Grass-pasture	483	90.60%	89.49%	89.04%	87.99%	92.69%	93.75%	**98.70%**	96.27%
Grass/Trees	730	98.07%	95.24%	95.98%	99.05%	97.62%	96.88%	97.30%	**100%**
Hay-windrowed	478	98.86%	97.05%	99.09%	99.20%	99.21%	97.51%	99.80%	**100%**
soybeans-notill	972	74.97%	68.20%	70.72%	77.55%	75.69%	84.19%	96.60%	**99.69%**
Soybeans-min	2455	84.87%	75.96%	77.94%	57.60%	61.61%	**95.77%**	78.20%	84.40%
Soybean-clean	593	81.16%	54.53%	61.23%	53.55%	84.58%	89.87%	95.90%	**100%**
Woods	1265	95.02%	92.87%	95.62%	88.67%	98.71%	97.85%	99.40%	**100%**

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY FOR DIFFERENT METHODS ON PAVIA UNIVERSITY

	Pixel-number	KLR	OMP	SP	kNN	SVM	Spec-EMP	SVM-ISOdata	MSMRF
Feature number		(103)	(103)	(103)	(103)	(103)	(103)	(103)	(40)
Overall accurary		83.56%	73.30%	74.86%	73.63%	79.15%	83.53%	91.20%	**92.52%**
Average accurary		84.83%	81.79%	83.19%	82.51%	87.66%	89.39%	92.94%	**95.41%**
Classes									
Asphalt	6304	82.96%	68.23%	69.78%	70.86%	84.30%	**95.33%**	94.40%	87.83%
Meadows	18146	83.34%	67.04%	67.90%	68.48%	67.01%	73.46%	87.45%	**90.56%**
Gravel	1815	64.13%	65.45%	69.20%	73.29%	68.43%	65.89%	61.32%	**93.55%**
Trees	2912	96.33%	97.29%	96.77%	93.52%	97.80	**99.18%**	98.63%	97.37%
Metal sheets	1113	99.19%	99.73%	99.64%	99.36%	99.37%	99.48%	99.91%	**100%**
Bare soil	4572	80.05%	73.27%	78.96%	69.81%	92.45%	84.15%	97.88%	**100%**
Bitumen	981	84.51%	87.26%	88.18%	93.09%	89.91%	97.22%	**100%**	**100%**
Bricks	3364	83.17%	81.87%	83.68%	84.58%	92.42%	96.12%	**99.02%**	89.36%
Shadows	795	89.81%	95.97%	94.59%	99.88%	97.23%	93.66%	97.86%	**100%**

the MJSR and sMRF processes, this image is also classified with high precision.

To conclude, compared with the former classification methods, a brief explanation for the advantage of the proposed method mainly lies on the elimination of redundant information that brings confusion to the classification. This improvement contributes to the construction of a more robust classification model. Moreover, according to the visualized classification maps, two main types of errors caused by this confusion in traditional methods can be concluded. The first is the wrong labeling that caused by confusing bands among different classes, and the second is the bad influence brought by the similarity of the same class that spreads the wrong labels. While as shown in the classification map of our method, the first error is eliminated by the maximized usage of the selected bands through MJSR that gets rid of confusing information, and the second one is solved gracefully via sMRF assistance which can preserve spatial coherence.

D. Discussion

There are still some problems to be discussed. The discussion focuses on two topics: 1) the effect of bare MJSR without spatial constraint and 2) the expansibility of sMRF optimization.

1) Effect of MJSR: One of the main contributions of this paper is MJSR. In order to see the actual effect of this functional step, we separate it from sMRF optimization. Fortunately, we find that although the accuracy declines obviously, the bare MJSR result still outperforms the classical methods mentioned above. On Indian Pines, the overall classification accuracy is 80.54% for MJSR, while the existing pixel-wise classification methods KLR, SVM, OMP, SP, and kNN are all no more than 80%. Moreover, on Pavia University, the accuracy is still the best 84.45%, compared with classical methods whose highest one is 83.56% for KLR.

We have to claim that another most significant contribution of our MJSR method lies in the band utilization. In our method, we make the best of every band to reach the final result. On the experimental images, the numbers of bands we use are no more than 40. The low correlation among bands contributes to the final result of MJSR.

2) Expansibility of sMRF Optimization: The second discussion of the proposed method is the expansibility of sMRF optimization. We apply our optimization process (sMRF) to both the existing kNN and SVM methods to verify this problem. The comparison conducted on Indian Pine is shown in Fig. 8. It is found that the accuracy of the traditional algorithm kNN tends to be around 70%, which is not a nice result. However, once the kNN algorithm is improved by our optimization algorithm, the accuracy reaches 90%, which is a distinct improvement. These two results are shown in Fig. 8(a) and (b), respectively. As shown in Fig. 8(c) and (d), the precision of SVM is around 80% and boosts to 97.59% after optimization. The corresponding comparisons are shown in Fig. 8(e) and (f) is the groundtruth. This improvement verifies the extraordinary expansibility of sMRF optimization.

TABLE III
COMPARISON OF CLASSIFICATION ACCURACY FOR MRF AND sMRF ON INDIAN PINE AND PAVIA UNIVERSITY

Image	Accuracy	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 5$	stepwise
Indian Pine	Overall	90.42%	92.05%	92.31%	92.53%	92.78%	**97.59%**
	Average	94.70%	95.83%	89.81%	83.24%	91.76%	**98.16%**
Pavia University	Overall	85.30%	86.40%	86.89%	87.57%	86.93%	**98.68%**
	Average	91.64%	91.72%	92.12%	91.90%	91.75%	**98.28%**

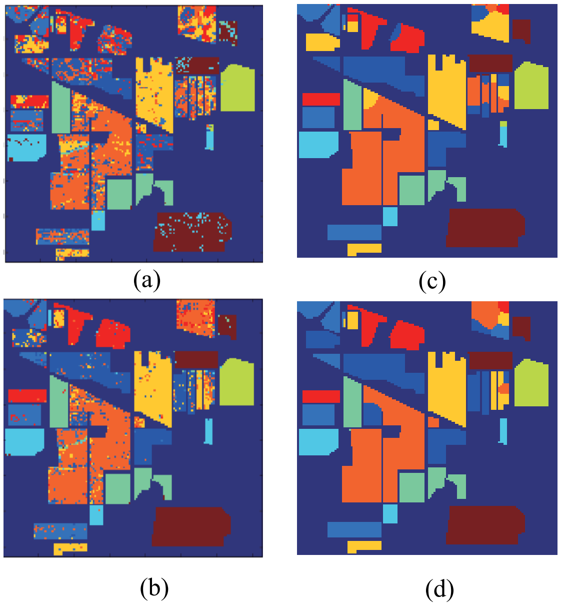


Fig. 8. Indian pine image. (a) kNN classification map, OA = 69.46%. (b) kNN with sMRF classification map, OA = 87.77%. (c) SVM classification map, OA = 79.69%. (d) SVM with sMRF classification map, OA = 97.59%.

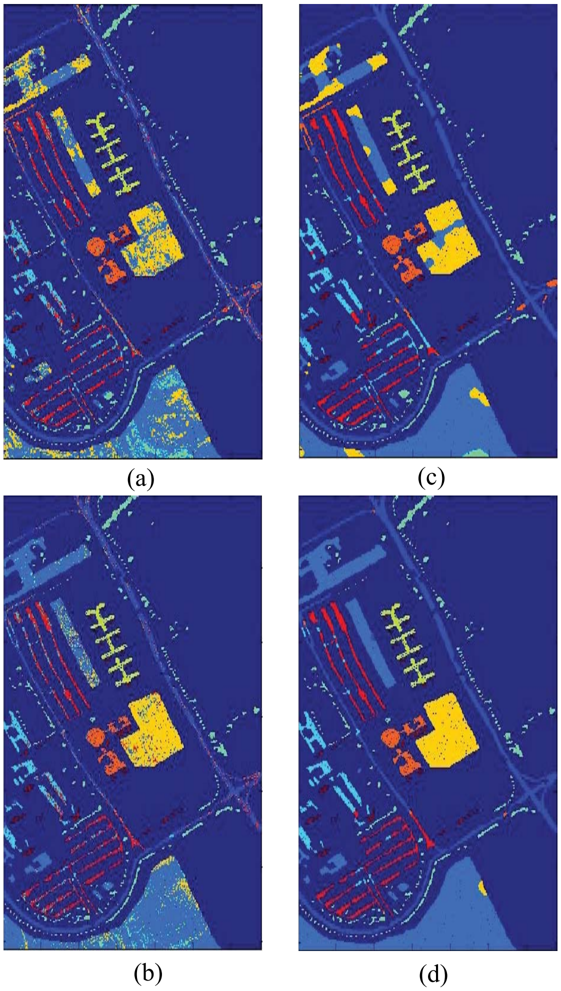


Fig. 9. Pavia University image. (a) kNN classification map, OA = 73.63%. (b) kNN with sMRF classification map, OA = 88.73%. (c) SVM classification map, OA = 79.15%. (d) SVM with sMRF classification map, OA = 98.68%.

The same improvement is also found in Pavia University that validates this high expansibility of sMRF too. The accuracy of kNN and SVM increases by 15.10% and 19.53%, respectively. More detailed result is shown in Fig. 9. To conclude, the AA of every method increases at least 10%, which means this optimization is not barely suitable for MJSR.

3) Comparison of MRF With sMRF: At last, a discussion of the effect of sMRF compared with MRF [30] is conducted to further prove the effectiveness of our stepwise optimization. The comparison is still based on Indian Pine and Pavia University both with 10% training samples. In the traditional MRF framework, as shown in (6) the weights of smoothness term and data term are balanced by coefficient λ. This λ ranges from 0.5 to 5 in most time. However, in sMRF this coefficient no longer exists. We take SVM as the initialization of these two methods and find sMRF outperforms MRF with distinct advantage. The detailed comparison is shown in Table III. We can see that on Indian Pine, the best OA of MRF is 92.78% when $\lambda = 5$ but the OA reaches 97.59% for sMRF. Similarly, for Pavia University, the best result of MRF is obtained when $\lambda = 3$ and OA is 87.57%, while the OA of sMRF is 98.68%. This can be viewed as an apparent improvement.

4) Analysis of Suboptimal Performances: We do not deny that though the overall performance of the proposed method is the best, suboptimal performances on several specific categories still exist. For Indian Pines, the accuracies of both Corn-notill and Soybean-min rank the third and the Grass-pasture ranks the second; for Pavia University, the Asphalt ranks the third and the Trees and Bricks rank the fourth. These accuracies in most time range from 80% to 90% which are barely satisfying. The main reason is that the classifier cannot always get the best performance on all the categories due to their different data properties. But overall, the proposed method performs the best for most ground categories and the averaged accuracy is the highest. A more specific research on these categories may be viewed as our main future work.

IV. CONCLUSION

In this paper, we have presented a new spectral–spatial HSI classification scheme. It aims to solve the problem brought by interband correlation among HSI bands and uses the spatial coherency to adaptively refine the classification results. For this purpose, the bands are first divided into several clusters, whose elements of every cluster are with low correlation. Then tasks are constructed by selecting bands from each band set. After that, a jointly sparse representation is applied to represent a specific input pixel, according to the reconstruction error of which the class label can be identified. At last, the sMRF spatial restriction helps with keeping the label consistency within a small neighborhood. Two main contributions are claimed in this paper: 1) the MJSR of HSI data and 2) the stepwise sMRF framework. Experimental results on two HSIs namely Indian Pines and Pavia University demonstrate that the proposed framework yields better performance, when compared with traditional popular methods.

References

[1] F. F. Sabins, *Remote Sensing: Principles and Applications*. Long Grove, IL, USA: Waveland Press, 2007.

[2] Q. Shi, B. Du, and L. Zhang, "Domain adaptation for remote sensing image classification: A low-rank reconstruction and instance weighting label propagation inspired algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5677–5689, Oct. 2015.

[3] Q. Shi, B. Du, and L. Zhang, "Spatial coherence-based batch-mode active learning for remote sensing image classification," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2037–2050, Jul. 2015.

[4] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[5] J. Zabalza *et al.*, "Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4418–4433, Aug. 2015.

[6] K. Kavitha, P. Nivedha, S. Arivazhagan, and P. Palniladeve, "Wavelet transform based land cover classification of hyperspectral images," in *Proc. Int. Conf. Commun. Netw. Technol.*, Sivakasi, India, 2014, pp. 109–112.

[7] N. Falco, L. Bruzzone, and J. A. Benediktsson, "An ICA based approach to hyperspectral image feature reduction," in *Proc. IEEE Trans. Geosci. Remote Sens. Symp.*, Quebec City, QC, Canada, 2014, pp. 3470–3473.

[8] Y. Mohsenzadeh, H. Sheikhzadeh, A. M. Reza, N. Bathaee, and M. M. Kalayeh, "The relevance sample-feature machine: A sparse Bayesian learning approach to joint feature-sample selection," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2241–2254, Dec. 2013.

[9] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[10] W. Liu and D. Tao, "Multiview Hessian regularization for image annotation," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2676–2687, Jul. 2013.

[11] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 980–993, Mar. 2015.

[12] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4349–4360, Oct. 2012.

[13] Y. Luo, D. Tao, C. Xu, B. Geng, and S. J. Maybank, "Manifold regularized multitask learning for semi-supervised multilabel image classification," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 523–536, Feb. 2013.

[14] C. Xu, D. Tao, and C. Xu, "Large-margin multi-viewinformation bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.

[15] H. Zhi, Q. Wang, Y. Shen, and M. Sun, "Kernel sparse multitask learning for hyperspectral image classification with empirical mode decomposition and morphological wavelet-based features," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5150–5163, Aug. 2014.

[16] 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.

[17] Q. Wang, Y. Yuan, and P. Yan, "Visual saliency by selective contrast," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 7, pp. 1150–1155, Jul. 2013.

[18] A. C. Braun, U. Weidner, and S. Hinz, "Support vector machines, import vector machines and relevance vector machines for hyperspectral classification—A comparison," in *Proc. IEEE 3rd Workshop Hyperspect. Image Signal Process. Evol. Remote Sens.*, Lisbon, Portugal, 2011, pp. 1–4.

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

[20] C.-H. Li, H.-S. Chu, B.-C. Kuo, and C.-T. Lin, "Hyperspectral image classification using spectral and spatial information based linear discriminant analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Vancouver, BC, Canada, 2011, pp. 1716–1719.

[21] R. Ji *et al.*, "Spectral-spatial constraint hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1811–1823, Mar. 2014.

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

[23] J. Lin, Q. Wang, and Y. Yuan, "In defense of iterated conditional mode for hyperspectral image classification," in *Proc. IEEE Int. Conf. Multimedia Expo*, Chengdu, China, 2014, pp. 1–6.

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

[25] B. Zhang, S. Li, X. Jia, L. Gao, and M. Peng, "Adaptive Markov random field approach for classification of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 973–977, Sep. 2011.

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

[27] J. L. Marroquin, E. A. Santana, and S. Botello, "Hidden Markov measure field models for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 11, pp. 1380–1387, Nov. 2003.

[28] P. Latorre-Carmona, J. Martinez Sotoca, F. Pla, J. Bioucas-Dias, and C. J. Ferre, "Effect of denoising in band selection for regression tasks in hyperspectral datasets," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 6, no. 2, pp. 473–481, Apr. 2013.

[29] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297–2307, May 2010.

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

[31] J. Wright *et al.*, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[32] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.

[33] Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-spectral saliency detection," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 34–41, 2013.

[34] Y. Nesterov, *Gradient Methods for Minimizing Composite Objective Function*. Louvain-la-Neuve, Belgium: CORE, 2007.

[35] X. Chen, W. Pan, J. T. Kwok, and J. G. Carbonell, "Accelerated gradient method for multi-task sparse learning problem," in *Proc. IEEE 9th Int. Conf. Data Mining*, Miami, FL, USA, 2009, pp. 746–751.

[36] F. Fernandez-Navarro, A. Riccardi, and S. Carloni, "Ordinal regression by a generalized force-based model," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 844–857, Apr. 2015.

[37] A. De Paola, S. Gaglio, G. L. Re, and F. Milazzo, "Adaptive distributed outlier detection for WSNs," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 888–899, May 2015.

[38] C. Yu, Q.-G. Wang, D. Zhang, L. Wang, and J. Huang, "System identification in presence of outliers," *IEEE Trans. Cybern.*, Doi: 10.1109/TCYB.2015.2430356.

[39] A. Quattoni, X. Carreras, M. Collins, and D. Trevor, "A efficient projection for $l_{1,\infty}$ regularization," in *Proc. IEEE Int. Conf. Mach. Learn.*, Montreal, QC, Canada, 2009, pp. 857–864.

[40] R. Neher and A. Srivastava, "A Bayesian MRF framework for labeling terrain using hyperspectral imaging," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1363–1374, Jun. 2005.

[41] J. Xu *et al.*, “The generalization ability of SVM classification based on Markov sampling,” *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1169–1179, Jun. 2015.

[42] H. Elliott, H. Derin, R. Cristi, and D. Geman, “Application of the Gibbs distribution to image segmentation,” Defense Tech. Inf. Center, Fort Belvoir, VA, USA, Tech. Rep. ADA133406, 1983.

[43] S. G. Beaven, D. Stein, and L. E. Hoff, “Comparison of Gaussian mixture and linear mixture models for classification of hyperspectral data,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 4. Honolulu, HI, USA, 2000, pp. 1597–1599.

[44] R. Szeliski *et al.*, “A comparative study of energy minimization methods for Markov random fields with smoothness-based priors,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008.

[45] Y. Xiao, H. Wang, and W. Xu, “Parameter selection of Gaussian kernel for one-class SVM,” *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 927–939, May 2015.

[46] A. Rakotomamonjy, “Surveying and comparing simultaneous sparse approximation or group-lasso algorithms,” *Signal Process.*, vol. 91, no. 7, pp. 1505–1526, 2011.

[47] D. Needell and R. Vershynin, “Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit,” *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 310–316, Apr. 2010.

[48] D. Böhning, “Multinomial logistic regression algorithm,” *Ann. Inst. Stat. Math.*, vol. 44, no. 1, pp. 197–200, 1992.

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

[50] W. Dai and O. Milenkovic, “Subspace pursuit for compressive sensing signal reconstruction,” *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 2230–2249, May 2009.

[51] S. Yu, S. De Backer, and P. Scheunders, “Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for hyperspectral satellite imagery,” *Pattern Recognit. Lett.*, vol. 23, nos. 1–3, pp. 183–190, 2002.

[52] L. Ma, M. M. Crawford, and J. Tian, “Local manifold learning-based-nearest-neighbor for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[53] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, “Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques,” *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.

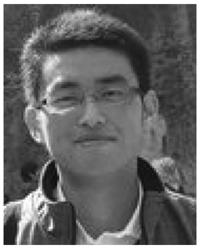
[54] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, “Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[55] K. Bernard, Y. Tarabalka, J. Angulo, J. Chanussot, and J. A. Benediktsson, “Spectral–spatial classification of hyperspectral data based on a stochastic minimum spanning forest approach,” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2008–2021, Apr. 2012.

Yuan Yuan (M’05–SM’09) is currently a Full Professor with the Chinese Academy of Sciences, Beijing, China. Her current research interests include visual information processing and image/video content analysis. She has authored or co-authored over 150 papers, including about 100 in reputable journals such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION as well as conference papers in CVPR, BMVC, ICIP, and ICASSP.

Jianzhe Lin (S’15) received the B.E. degree in optoelectronic information engineering and the B.A. degree in English from the Huazhong University of Science and Technology, Wuhan, China, in 2013. He is currently pursuing the master’s degree with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, China.

His current research interests include computer vision and machine learning.



Qi Wang (M’15–SM’15) received the B.E. degree in automation and Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently an Associate Professor with the School of Computer Science and the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi’an, China. His current research interests include computer vision and pattern recognition.