Matrix-Based Discriminant Subspace Ensemble for Hyperspectral Image Spatial–Spectral Feature Fusion

Renlong Hang, Qingshan Liu, Senior Member, IEEE, Huihui Song, and Yubao Sun

Abstract—Spatial–spectral feature fusion is well acknowledged as an effective method for hyperspectral (HS) image classification. Many previous studies have been devoted to this subject. However, these methods often regard the spatial–spectral high-dimensional data as 1-D vector and then extract informative features for classification. In this paper, we propose a new HS image classification method. Specifically, matrix-based spatial–spectral feature representation is designed for each pixel to capture the local spatial contextual and the spectral information of all the bands, which can well preserve the spatial–spectral correlation. Then, matrix-based discriminant analysis is adopted to learn the discriminative feature subspace for classification. To further improve the performance of discriminative subspace, a random sampling technique is used to produce a subspace ensemble for final HS image classification. Experiments are conducted on three HS remote sensing data sets acquired by different sensors, and experimental results demonstrate the efficiency of the proposed method.

Index Terms—Contextual spatial–spectral feature fusion, hyperspectral (HS) image classification, matrix-based discriminant analysis (MDA), random sampling, support vector machine (SVM).

I. INTRODUCTION

W

ith the rapid development of imaging spectroscopy technologies, current sensors are able to acquire hyperspectral (HS) data with high spatial and spectral resolutions simultaneously. For instance, the Reflective Optics System Imaging Spectrometer (ROSI) sensor covers a range of 115 spectral channels with the spatial resolution of 1.3 m. HS data can be presented as a 3-D cube consisting of two spatial dimensions and one spectral dimension, as shown in Fig. 1, which provides an avenue for accurate classification of land cover scenes.

HS image classification often suffers from two issues. The first one is that the widely used classification models based on pixelwise spectral information always lead to a salt and pepper thematic map because the intraclass spectral responses vary and interclass land covers may have very similar spectral natures in the study scene [1]. The other one is that classifiers often need to process high-dimensional data with a small number of training samples. A typical HS image contains hundreds of spectral bands. Directly using the spectral information without any preprocessing not only increases the computation complexity but also induces the Hughes phenomenon [2]. This means that, when the number of features exceeds a threshold, the classification accuracy starts to decrease.

In order to address the first issue, many approaches were proposed to incorporate spatial information into spectral information [3]–[5]. This is because the footprint of one kind of material or one object often contains more than one pixel, and thus, neighboring pixels spatially correlate with a high probability. In addition, exploring spectral and spatial information simultaneously is beneficial to reducing the outlier effects. The existing spatial–spectral feature fusion approaches can be roughly categorized into three classes: feature-level fusion, decision-level fusion, and regularization-based fusion.

For feature-level fusion, one often extracts the spectral features and the spatial features independently and then concatenate these features into a vector [6]–[10] or construct multiple kernel functions on their corresponding features [11]–[13], followed by a classifier. In [7], morphological profiles were utilized to extract the spatial information, which was then combined with spectral features. However, this direct concatenation will lead to a very high dimensional feature space, thereby resulting in a decreased classification performance or even leading to overfitting. To address this issue, kernel method is an alternative scheme. In [11] and [12], the authors constructed different kernel functions for different features, and each of them only needs to deal with a low-dimensional feature. Nevertheless, tuning the weight parameters of kernels for this type of method is time consuming. Recently, a new family of generalized composite kernel framework without any weight
parameters has been proposed, and the results show an efficient classification performance [13].

For decision-level fusion, multiple results are first derived from spatial and spectral information, respectively. In [14]–[16], the authors combined the results of pixelwise classification and the segmentation map by majority voting. These methods perform particularly well for images with large spatial structures. However, the results heavily depend on a higher accuracy of the initial segmentation results, which is a very difficult task for remote sensing data. In [17], three spatial features, including gray-level cooccurrence matrix and differential morphological profiles, as well as urban complexity index, were exploited to construct a support vector machine (SVM) ensemble. However, training an optimal classifier with multiple SVMs is very time consuming.

For regularization-based fusion, they incorporate a regularizer representing spatial information into the original object function. In [18] and [19], Markov random field (MRF) modeling the joint prior probabilities of each pixel and its spatial neighbors was incorporated into the Bayesian classifier as a regularizer. MRF achieved a great success in characterizing the spatial information. However, optimizing the objective function in MRF requires an iterative step, such as simulated annealing, which is extremely time consuming on data with high resolution. Recently, Zhou et al. assumed that the labels of pixels from the same spatial neighbor are identical, and thus proposed a regularizer to preserve the local spatial information in the reduced-dimensionality space [20]. Similar idea can be found in [21].

To deal with the second issue, dimensionality reduction (DR) technique [22] has been employed to reduce the redundant information and extract the most distinct features for classification. Popular DR techniques include unsupervised approaches, such as principal component analysis (PCA) [23] and independent component analysis (ICA) [24], as well as supervised approaches, such as Fisher linear discriminant analysis (LDA) [25]. The main difference between PCA and LDA is that PCA aims to an orthogonal set of vectors that maximize the variance of the projected vectors, whereas LDA aims to seek the discriminant vectors such that the ratio of the interclass distance to the intraclass distance is maximized. Accordingly, LDA and its variants have been extensively exploited in terms of remote sensing image classification. In [26], a regularized LDA was introduced to cope with a particular ill-posed problem where the ratio between the number of training samples and the number of spectral features is small. Nonparametric weighted feature extraction [27] is a creative extension of LDA. It proposed a new criteria to underling the separability between the class distribution boundaries. Recently, to make full use of unlabeled samples when there are no sufficient training samples, semisupervised LDA has been proposed [21], [28], [29]. In [29], Liao et al. designed an optimal projection matrix to preserve the local neighborhood information inferred from unlabeled samples, while simultaneously maximizing the class discrimination of the data inferred from the labeled samples. However, computing the high-dimensional affinity matrix representing the correlations between different unlabeled samples is very memory space consuming. To alleviate this phenomenon, Yuan et al. proposed a self-adaptive method to construct a local constraint only on the particular test sample rather than the whole unlabeled samples [21]. Therefore, each test sample had its unique transformation matrix for dimension reduction.

In this paper, we focus on the first class spatial–spectral feature fusion methods, i.e., the feature-level fusion. However, most of the existing methods in this class are based on vectorized representation by extracting the spectral and spatial features independently, which may lose some intrinsic relationships between them. To combine spectral and spatial features efficiently, a new feature fusion method is proposed. First, matrix-based spatial–spectral feature representation is designed for each pixel to capture the local spatial contextual and spectral information of all the bands, which can well preserve the spatial–spectral correlation. Then, matrix-based discriminant analysis (MDA) is adopted to learn the discriminative feature subspace for classification. To further improve the performance of discriminative subspace, a random sampling technique is used to produce a subspace ensemble for final HS image classification. In classification stage, SVM [30] is used due to its many significant merits for remote sensing image classification, such as less sensitivity to dimensionality, sparse representation of decision boundary, and robustness to noisy data. The main contributions of this paper can be summarized as follows.

1) MDA is first introduced to extract the spatial and spectral features of remote sensing images simultaneously.
2) Subspace ensemble method based on random sampling is used to derive diverse discriminant subspaces and, in some extent, alleviate the imbalanced data distribution problem in remote sensing image classification, thus promoting the classification performance.

The rest of this paper is structured as follows. In the following section, we introduce our proposed method in detail. The experiments are reported in Section III, followed by the conclusion in Section IV.

II. METHODOLOGY

A. Matrix-Based Spatial–Spectral Feature Representation

Suppose that an HS scene is denoted by a 3-D matrix \( X \in \mathbb{R}^{m \times n \times l} \) with \( m \times n \) pixels and \( l \) spectral bands. Assume that \( X_{ij}^{(k)} \) denotes the pixel at the spatial location \((i, j)\) in band \( k \), where \( i \in \{1, 2, \ldots, m\}, \quad j \in \{1, 2, \ldots, n\}, \quad k \in \{1, 2, \ldots, l\} \). Since the HS pixels within a small neighborhood usually consist of similar materials [31], [32], we exploit the spatial neighborhood to combine the spectral and spatial–contextual information. Suppose the spatial window size is \( \omega \times \omega \), which is an odd positive integer. Then the spectral and spatial–contextual feature of pixel \( X_{ij} \) in band \( k \) is \( \Omega(X_{ij}^{(k)}) = \{ X_{pq}^{(k)} | p \in i-b, \ldots, i+b, \quad q \in j-b, \ldots, j+b \} \), where \( b = (\omega - 1)/2 \). Thereafter, we concatenate \( \Omega(X_{ij}^{(k)}) \) into a row vector. Similar operations are executed on band 1 to band \( l \). Finally, all of these \( l \) row vectors are arranged to form a new matrix \( Y_{ij} \in \mathbb{R}^{l \times \omega^2} \).

B. Matrix-Based Discriminative Feature Subspace Learning

As discussed in Section II-A, \( Y \in \mathbb{R}^{l \times \omega^2} \) is a new matrix representation for one pixel. To reduce the redundant information
in this feature matrix and improve the discriminant ability of feature representation, we use MDA to map $Y$ into a new subspace $[33]$, in which the intraclass scatter is minimized and the interclass scatter is maximized. Assume that $L = [\mu_1, \mu_2, \ldots, \mu_r] \in \mathbb{R}^{L \times r}$ and $R = [\nu_1, \nu_2, \ldots, \nu_c] \in \mathbb{R}^{c \times L}$ are two transformation matrices. Then, the projection of $Y$ onto the $(r \times c)$-dimensional space $L \otimes R$ can be expressed as

$$Z = L^T Y R.$$  \hspace{1cm} (1)

Suppose there are $N$ training pixels and $C$ different classes to be classified. The $j$th training pixel is $Y_j$, where $j \in \{1, 2, \ldots, N\}$. $M$ and $\hat{M}_i$ denote the mean of all training pixels and training pixels in class $\Pi_i$, $i \in \{1, 2, \ldots, C\}$, respectively. $N_i$ is the number of training pixels in class $\Pi_i$. To get the optimal projection matrices, we maximize the ratio of the interclass scatter matrix and the intraclass scatter matrix for MDA as for LDA. Fortunately, the total scatter of the projected pixels can be measured by the trace of their covariance matrix. Therefore, the objective function can be described as follows:

$$J = \frac{S_B}{S_W}$$  \hspace{1cm} (2)

where $S_B$ denotes the interclass scatter matrix, and $S_W$ denotes the intraclass scatter matrix with the following definitions:

$$S_B = \text{tr} \left( \sum_{i=1}^{C} N_i (\hat{M}_i - \bar{M})(\hat{M}_i - \bar{M})^T \right)$$

$$S_W = \text{tr} \left( \sum_{i=1}^{C} \sum_{Z_k \in \Pi_i} (Z_k - \hat{M}_i)(Z_k - \hat{M}_i)^T \right).$$  \hspace{1cm} (3)

Since $\hat{M}_i = \frac{1}{N_i} \sum_{Z_k \in \Pi_i} Z_k = \frac{1}{N_i} \sum_{Y_k \in \Pi_i} L^T Y_k R = L^T M_i R$

thus (3) can be rewritten as

$$S_B = \text{tr} \left( \sum_{i=1}^{C} N_i L^T (M_i - M) R R^T (M_i - M)^T L \right)$$

$$S_W = \text{tr} \left( \sum_{i=1}^{C} \sum_{Y_k \in \Pi_i} L^T (Y_k - M_i) R R^T (Y_k - M_i)^T L \right).$$

The optimal transformations $L$ and $R$ will maximize (2). The same as in [33], we utilize an iterative algorithm to optimize $L$ and $R$. In particular, for a fixed $R$, (2) can be rewritten as

$$J = \frac{L^T S_B^R L}{L^T S_W^R L}$$  \hspace{1cm} (4)

where $S_B^R = \sum_{i=1}^{C} N_i (M_i - M) R R^T (M_i - M)^T$, and $S_W^R = \sum_{i=1}^{C} \sum_{Y_k \in \Pi_i} (Y_k - M_i) R R^T (Y_k - M_i)^T$. Hence, the optimal solution of $L$ consists of $r$ eigenvectors corresponding to $r$ maximal eigenvalues by computing an eigendecomposition on $(S_W^R)^{-1} S_B^R$. Subsequently, the optimal solution of $R$ can be obtained when $L$ is fixed. The whole process will iterate until a predefined convergence condition is arrived.

### C. Learning Subspace Ensemble With Random Sampling

Ensemble learning has been proved to be an efficient technique to improve the stability and accuracy of single weak classifier by training several different classifiers and combining their decisions [34], [35]. When forming ensemble, the
accuracy is mainly affected by the diversity of classifiers, which can be achieved via different methods [36], [37]. One of the most widely used methods is bagging [38], [39]. The core idea is to randomly sample a few training subsets from the original training set repeatedly with replacement. In [40], bagging has been used to construct an SVM ensemble for land cover classification, and the experiments suggest that an optimized ensemble method could lead to improved results. However, it costs so much time to train several SVM classifiers simultaneously. In [7], a feature ensemble scheme was utilized to combine discrimination information of two feature subspaces (i.e., spatial and spectral) by training only one SVM classifier.

In HS image classification field, we are often confronted with imbalanced data distribution problem, in which the number of samples in one class is much smaller than that in other classes [35]. For example, the number of available samples in Indian Pines (IP) data set ranges from 20 to 2468 for different classes. In this situation, the intraclass scatter matrix will be dominated by the majority classes. Thus, the dimensionality reduced subspace can preserve the structure information of majority classes while ignoring that of the minority classes, resulting in a decreased performance for subsequent classification. Random sampling method can alleviate this phenomenon to some extent by taking advantage of a part of the original samples instead of the whole data, which can be confirmed from the classification results shown in Table V.

Based on the preceding discussion, we propose a random-sampling-based subspace ensemble method for spectral–spatial feature fusion. On one hand, motivated by bagging methods, random sampling repeatedly derives a few different training subsets, leading to diverse discriminant subspaces. These subspaces ensemble will improve the classification capacity. On the other hand, random sampling can alleviate class imbalanced data problem, resulting in an enhanced subspace.

Fig. 2 shows the detailed flowchart of the proposed method. Suppose that we have a training set $Y$, and $Y_j$ represents the $j$th training pixel. First, $n$ training subsets are randomly sampled from $Y$ with replacement according to a certain proportion (e.g., 80%), and $Y^{(i)}$, $i \in \{1, 2, \ldots, n\}$ denotes the $i$th training subset. Then, for every subset $Y^{(i)}$, MDA is used to learn the optimal transformations $L_i$ and $R_i$ as discussed in Section II-B. Thereafter, for every couple $L_i$ and $R_i$, the projection matrix $Z_j^{(i)}$ of $Y_j$ can be achieved according to (1). Therefore, we can obtain $n$ projection matrices $\{Z_j^{(1)}, Z_j^{(2)}, \ldots, Z_j^{(n)}\}$. Finally, all of these matrices are reshaped to vectors, respectively. In feature fusion phase, we concatenate these vectors into one stacked vector, and PCA is used to reduce the redundancy in the vector.

### III. Experiments

#### A. Data Set

We test the proposed method on three HS remote sensing data sets, which are widely used for classification algorithm evaluation [1], [15], [31], [41].
TABLE II
NUMBER OF PIXELS FOR TRAINING/TESTING AND THE TOTAL NUMBER OF PIXELS FOR EACH CLASS IN THE PUS GROUND TRUTH MAP

<table>
<thead>
<tr>
<th>Class</th>
<th>Name</th>
<th>Total</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Asphalt</td>
<td>6641</td>
<td>548</td>
<td>6093</td>
</tr>
<tr>
<td>C2</td>
<td>Meadows</td>
<td>18649</td>
<td>540</td>
<td>18109</td>
</tr>
<tr>
<td>C3</td>
<td>Gravel</td>
<td>2099</td>
<td>392</td>
<td>1707</td>
</tr>
<tr>
<td>C4</td>
<td>Trees</td>
<td>3064</td>
<td>524</td>
<td>2540</td>
</tr>
<tr>
<td>C5</td>
<td>Painted metal sheets</td>
<td>1345</td>
<td>265</td>
<td>1080</td>
</tr>
</tbody>
</table>

Fig. 5. KSC data set. (a) False-color composite of the KSC with bands 31, 21, and 11. (b) Ground truth map containing 13 mutually exclusive land cover classes.

- Indian Pines (IP) data set: The first data set was acquired by the AVIRIS sensor over the Indian Pine test site in northwestern Indiana, USA, on June 12, 1992. The original data set contains 224 spectral bands. However, we utilize 200 of them after removing four bands containing zero values and 20 noisy bands affected by water absorption. The spatial size of the image was $145 \times 145$ pixels, and the spatial resolution was 20 m. As described in [32], we also used ten major classes for experiments. The false-color composite image and the ground truth map are shown in Fig. 3. The number of the total labeled pixels and the number of pixels for training and testing for each class are shown in Table I.

- Pavia University Scene (PUS) data set: The second data set was acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy, on July 8, 2002. The original image was recorded with 115 spectral channels ranging from $0.43$ to $0.86 \mu m$ and covering the visible and infrared spectrum. The utilized image contains 103 bands after removing noisy bands. The image size was $610 \times 340$ pixels with a spatial resolution of 1.3 m. A three-band false-color composite image and the ground truth map are shown in Fig. 4. In the ground truth map, there are nine classes of land cover with more than 1000 labeled pixels for each class. The number of pixels for training and testing and the total number of pixels for each class are shown in Table II.

- Kennedy Space Center (KSC) data set: The third data set was acquired by the AVIRIS sensor over Kennedy Space Center, Florida, on March 23, 1996. The original data set contains 224 spectral bands. However, we utilize 176 bands of them after removing bands with water absorption and low signal-to-noise ratio. The spatial size of the image was $512 \times 614$ pixels, and the spatial resolution was 18 m. Discrimination of land covers in this data set is difficult due to the similarity of spectral signatures among certain vegetation types. For classification purposes, 13 classes representing the various land cover types that occur in this environment were defined. Fig. 5 shows a three-band false-color composite image and the ground truth map. Table III shows the number of the total labeled pixels and the number of pixels for training and testing for each class.

B. Experimental Setup

To examine the performance of the proposed method, we compared with the following five commonly used approaches and four recently proposed spatial–spectral feature fusion methods: 1) the method directly based on pixelwise spectral features (PX); 2) the method by using PCA to extract the most relevant features from spectral feature vectors (PX+PCA); 3) the method by using LDA to extract the most informative features from spectral feature vectors (PX+LDA); 4) the method based on spectral and spatial–contextual feature, in which the feature vectors were built from the concatenation of spectral and spatial features and then LDA was used to extract the most discriminant features from these feature vectors (PT+LDA); 5) the method that first used the extended morphological profiles (EMPs) to characterize the spatial feature of each pixel as proposed in [7] and then used LDA to extract the most useful features from the feature vectors built by the concatenation of spectral and spatial features (EMP+LDA); 6) the method based on tensor discriminative locality alignment (TDLA) proposed in [32]; 7) the method based on spectral-spatial kernel discriminant analysis (SSKDA) proposed in [21]; 8) the method based on spatial and spectral regularized local discriminant embedding (SSRLDE) proposed in [20]; and 9) the method based on generalized composite kernels proposed in [13]. Since the realization of methods 6, 7, 8, and 9 is very complex, we directly used the classification results in these papers and compared it with our proposed method under the
same number of training samples from each class. In addition, to illustrate the effectiveness of random-sampling-based subspace ensemble method, we also compared the proposed method (i.e., PT+MDA+RS) with only matrix representation and MDA (i.e., PT+MDA).

In all experiments, we adopted the radial basis kernel function (RBF) for SVM [42]. The values for the regularization parameter $\rho$ and the variance $\gamma$ of the RBF kernel parameters were chosen by using a grid search from the given sets $\{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$ and $\{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$. Since remote sensing data often involves more than two classes, we employed the one-against-one strategy in this paper. In addition, the remained dimensionalities after LDA in methods 3, 4, and 5 were all $C - 1$, where $C$ is the number of classes to be classified.

The classification accuracy was evaluated by the overall accuracy (OA), the average accuracy (AA), the per-class accuracy, and the Kappa coefficient $\kappa$. OA is defined by the ratio between the number of correctly classified pixels to the total number of pixels in the testing set. AA refers to the average of accuracies in all classes, and $\kappa$ is the percentage of agreement corrected by the number of agreements that would be expected purely by chance.

C. Experiment 1: IP Data Set

In order to show the effectiveness of our proposed method, we quantitatively and qualitatively evaluate the classification performance by comparing with the aforementioned classification methods. For this data set, 30 available pixels were randomly chosen from each class as training set, and the rest were used as testing set. Before SVM training, all data were normalized to zero mean and unit variance to avoid scaling issues in computing the kernels. In order to reduce the effect of random selection, all experiments were repeated ten times. For the proposed method, the best window size was $9 \times 9$, the optimal parameter values for MDA were $r = 151$ and $c = 9$, the number of training subsets is 5, and every subset contains 80% samples from the original training set.

Table IV presents the quantitative results achieved by different methods, and bold values indicate the best results. From this table, we can observe that the accuracies of the spatial–spectral classification methods (PT+LDA, EMP+LDA, TDLA, PT+MDA, and PT+MDA+RS) are higher than those of the pixel-wise classification methods (PX, PX+PCA, and PX+LDA). In particular, the OA, AA, and $\kappa$ of EMP+LDA are significantly higher than those of PT+LDA because the spatial features extracted by morphological operators are more informative than the simple contextual information from neighbors. However, the PT+MDA method using the same information as PT+MDA but with different representation approach and feature extraction method can get higher OA, AA, and $\kappa$ than EMP+LDA. This indicates that the PT+MDA method can extract the most distinct spatial and spectral features simultaneously for supervised classification. More importantly, the performance of our proposed method PT+MDA+RS is higher than that of PT+MDA, which certifies the effectiveness of random-sampling-based subspace ensemble method. In addition, for each class accuracy, our proposed method gets the highest values in eight classes. Furthermore, it is interesting to see that our proposed method can achieve superior OA compared with the three recently proposed methods in Table V. Note that, in [13], the whole 16 classes of samples instead of the main ten classes were used for experiments and the number of available samples in each class ranges from 20 to 2468. However, our proposed method
method can still improve the OA significantly from 93.93 to 97.24. It indicates that our proposed method can effectively deal with the class imbalanced data classification problem.

Fig. 6 shows the classification maps of different methods. In each of these figures, different colors refer to different land covers. Compared with the ground truth map in Fig. 3, it can be seen that some of Corn-mintill were misclassified into Soybean-mintill pixels in all classification maps because these two materials have similar spectral features. Although the spatial information is addressed in the PT+LDA and PT+EMP methods, only a little improvement can be observed in Fig. 6(d) and (e), because the spatial–spectral correlation information was not fully employed in the vectorized representation. On the contrary, the results from our proposed method in Fig. 6(f) shows fewer misclassifications than other classification maps. In addition, the box plot of OA in Fig. 7 for two different methods demonstrates that PT+MDA+RS is more stable than PT+MDA, which is consistent with the variance from Table IV.

D. Experiment 2: PUS Data Set

As in [7], we randomly chose 3921 pixels as training set and the rest of pixels as testing set for the PUS data set. The optimal values for MDA were $r = 61$ and $c = 5$. The optimal window size was $5 \times 5$, which is because many small structure objects exist in this data set and large window sizes may induce extra information of neighboring classes, leading to lower classification accuracies.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>94.95 (97)</td>
<td>81.13±2.72 (83.12±1.99)</td>
<td>93.93±0.57 (97.24±0.75)</td>
</tr>
<tr>
<td>PUS</td>
<td>87.81 (94.33)</td>
<td>88.18±1.70 (87.38±1.01)</td>
<td>98.09 (99.29)</td>
</tr>
<tr>
<td>KSC</td>
<td>—</td>
<td>97.17±0.82 (97.02±0.49)</td>
<td>—</td>
</tr>
</tbody>
</table>

Fig. 6. Classification maps of six different methods on IP data. (a) PX. (b) PX+PCA. (c) PX+LDA. (d) PT+LDA. (e) EMP+LDA. (f) PT+MDA+RS.

Fig. 7. Box plot of OA of different methods on IP data set. Numbers in the horizontal axis correspond to PT+MDA (1) and PT+MDA+RS (2). We plot these boxes by doing ten different replications. The red line through the center of each box indicates the median value of the OA. The edges of boxes are the 25th and 75th percentiles.
Table VI describes the classification performance of different methods in terms of OA, AA, per-class accuracy, and $\kappa$. Similar to the IP data set, the spatial–spectral classification methods are all better than the pixelwise classification methods. In particular, our proposed method PT+MDA+RS reaches the highest OA, AA, and $\kappa$. Note that the improvement of our proposed method against PT+MDA on this data set is not obvious compared with the IP data set. This is, in part, because the PT+MDA method performs very well, and thus, the improvement space is limited. On the other hand, for another evaluation criterion per-class accuracy, our method can get the highest values in eight classes, which can also confirm the
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VII
OA, AA, PER-CLASS ACCURACY (%), \( \kappa \), AND STANDARD DEVIATION OF TEN RUNS ACHIEVED BY SEVEN METHODS ON THE KSC DATA SET

<table>
<thead>
<tr>
<th>Label</th>
<th>PX</th>
<th>PX+PCA</th>
<th>PX+LDA</th>
<th>PT+LDA</th>
<th>EMP+LDA</th>
<th>PT+MDA</th>
<th>PT+MDA+RS</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>94.19±1.75</td>
<td>95.01±1.53</td>
<td>95.59±1.80</td>
<td>98.54±1.35</td>
<td>98.04±1.38</td>
<td>98.70±0.81</td>
<td>99.12±0.38</td>
</tr>
<tr>
<td>C2</td>
<td>86.76±2.88</td>
<td>85.30±4.58</td>
<td>87.49±0.61</td>
<td>94.43±3.40</td>
<td>93.97±3.25</td>
<td>95.32±4.62</td>
<td>98.40±1.46</td>
</tr>
<tr>
<td>C3</td>
<td>85.04±4.86</td>
<td>89.57±3.19</td>
<td>88.35±4.43</td>
<td>86.52±5.61</td>
<td>88.35±5.15</td>
<td>98.42±0.73</td>
<td>97.12±2.33</td>
</tr>
<tr>
<td>C4</td>
<td>59.47±8.33</td>
<td>60.70±1.69</td>
<td>63.26±4.97</td>
<td>70.93±4.84</td>
<td>74.19±6.50</td>
<td>90.20±5.42</td>
<td>90.50±3.12</td>
</tr>
<tr>
<td>C5</td>
<td>50.90±4.51</td>
<td>48.41±5.12</td>
<td>67.72±5.04</td>
<td>65.93±4.84</td>
<td>90.07±8.10</td>
<td>83.02±5.84</td>
<td>92.33±3.88</td>
</tr>
<tr>
<td>C6</td>
<td>60.10±5.12</td>
<td>60.00±8.87</td>
<td>57.38±4.13</td>
<td>80.58±4.96</td>
<td>93.30±6.23</td>
<td>91.69±5.49</td>
<td>96.30±3.41</td>
</tr>
<tr>
<td>C7</td>
<td>83.62±7.35</td>
<td>88.09±9.35</td>
<td>88.09±5.85</td>
<td>88.51±5.83</td>
<td>96.17±5.58</td>
<td>97.47±2.49</td>
<td>98.27±1.71</td>
</tr>
<tr>
<td>C8</td>
<td>91.49±2.35</td>
<td>88.40±3.44</td>
<td>91.65±1.80</td>
<td>94.33±1.53</td>
<td>91.80±6.75</td>
<td>96.13±2.06</td>
<td>98.50±1.38</td>
</tr>
<tr>
<td>C9</td>
<td>96.57±1.82</td>
<td>97.82±1.34</td>
<td>97.74±1.65</td>
<td>98.82±0.07</td>
<td>99.24±1.45</td>
<td>99.68±0.30</td>
<td>100±0.00</td>
</tr>
<tr>
<td>C10</td>
<td>94.40±2.64</td>
<td>95.60±1.84</td>
<td>96.26±0.42</td>
<td>99.07±0.03</td>
<td>95.16±2.42</td>
<td>97.42±1.37</td>
<td>96.70±1.74</td>
</tr>
<tr>
<td>C11</td>
<td>95.38±1.07</td>
<td>96.34±1.60</td>
<td>98.36±1.03</td>
<td>97.82±2.10</td>
<td>96.71±5.84</td>
<td>99.07±0.58</td>
<td>99.83±0.13</td>
</tr>
<tr>
<td>C12</td>
<td>94.17±4.39</td>
<td>96.07±2.58</td>
<td>94.39±1.04</td>
<td>93.77±1.63</td>
<td>97.26±1.88</td>
<td>98.84±1.04</td>
<td>99.94±0.06</td>
</tr>
<tr>
<td>C13</td>
<td>99.88±0.02</td>
<td>99.90±0.08</td>
<td>99.18±0.00</td>
<td>99.78±0.00</td>
<td>100±0.00</td>
<td>99.34±0.36</td>
<td>99.78±0.16</td>
</tr>
<tr>
<td>OA</td>
<td>89.78±0.49</td>
<td>90.37±0.24</td>
<td>91.28±0.43</td>
<td>93.94±0.36</td>
<td>94.62±0.78</td>
<td>97.24±0.46</td>
<td>98.32±0.49</td>
</tr>
<tr>
<td>AA</td>
<td>83.98±0.96</td>
<td>84.71±0.91</td>
<td>86.57±0.97</td>
<td>90.00±0.38</td>
<td>92.91±1.47</td>
<td>97.59±0.72</td>
<td>97.44±0.73</td>
</tr>
<tr>
<td>( \kappa )</td>
<td>86.62±0.54</td>
<td>89.27±0.28</td>
<td>90.30±0.48</td>
<td>93.25±0.40</td>
<td>94.01±0.87</td>
<td>96.93±0.65</td>
<td>98.13±0.68</td>
</tr>
</tbody>
</table>

Fig. 9. Classification maps of six different methods on KSC data. (a) PX. (b) PX+PCA. (c) PX+LDA. (d) PT+LDA. (e) EMP+LDA. (f) PT+MDA+RS.

efficiency of our proposed method. In addition, our proposed method can do as well as, if not better than, the three recently proposed methods in Table V.

Fig. 8 shows the classification maps of six methods. From this figure, we can observe that the spatial–spectral classification methods [see Fig. 8(d)–(f)] significantly reduce outliers and generate more homogeneous regions compared with the pixelwise classification methods [see Fig. 8(a)–(c)], particularly for Meadows and Gravel.

E. Experiment 3: KSC Data Set

For the KSC data set, we randomly selected 10% available pixels from each class as training set and the rest of pixels as testing set. The best reduced dimensionalities are \( r = 91 \) and \( c = 17 \), and the best window sizes are \( 5 \times 5 \) and \( 3 \times 3 \) for PT+MDA+RS and PT+MDA, respectively. Table VII describes the classification performance of different methods in terms of OA, AA, per-class accuracy, and \( \kappa \). Similar to the other two data sets, the spatial–spectral classification methods are all better than the pixelwise classification methods. In particular, our proposed method PT+MDA+RS reaches the highest OA, AA, and \( \kappa \). For per-class accuracy, our method can get the highest values in 10 out of 13 classes. In addition, our proposed method can get similar OA as SSRLDE in Table V. Fig. 9 shows the classification maps of six methods.

F. Parameter Analysis

In the proposed method, there are two important parameters, i.e., the reduced dimensionalities \( r \) and \( c \) in MDA and the window size in matrix representation; we thus examine their effects on the classification accuracy here by conducting experiments in the IP and PUS data sets. For simplicity, we fix the window size when we tune the reduced dimensionalities and vice versa.

1) Results for Different Reduced Dimensionalities: To show the effects of reduced dimensionalities \( r \) and \( c \) on the
classification accuracy, we plot a 3-D diagram of OA against different reduced dimensionalities for the proposed method. By setting the window size as $9 \times 9$ for IP data and $5 \times 5$ for PUS data, respectively, the result is shown in Fig. 10, where the $x$-axis denotes the reduced dimensionality $r$, the $y$-axis the reduced dimensionality $c$, and the $z$-axis the OA. From Fig. 10(a), we can observe that, at the initial phase, when $r$ is small, the OA is very low. As $r$ increases, the OA increases, and it will get the maximal value when $c = 9$ and $r = 151$. Similar results can be found in Fig. 10(b) for PUS data, and the maximal value appears at $c = 5$ and $r = 61$.

2) Results for Various Window Sizes: Based on the facts that remote sensing data sets are featured by different spatial resolutions and objects with various sizes exist in one data set, we thus evaluated the effects of different window sizes on the classification performance. By fixing the values of other parameters, we changed the window size from $3 \times 3$ to $9 \times 9$. Fig. 11 shows the OA of PT-LDA, PT-MDA, and PT-MDA+RS versus different window sizes. From Fig. 11(a), we can see that our proposed method PT-MDA+RS always outperforms the other two methods in the same window size and the OA of these three different methods increase as window size increases. To get the best classification results, we select $9 \times 9$ as the optimal window size for IP data. From Fig. 11(b), we can see that, when the window size exceeds $5 \times 5$, the OA decreases. Hence, the optimal window size for PUS data is $5 \times 5$.

### Table VIII

<table>
<thead>
<tr>
<th>Data</th>
<th>Time Consuming (in Seconds) for Different Methods on the Three Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>PX</td>
</tr>
<tr>
<td></td>
<td>2.12</td>
</tr>
<tr>
<td>PUS</td>
<td>14.45</td>
</tr>
<tr>
<td>KSC</td>
<td>0.48</td>
</tr>
</tbody>
</table>

### G. Computational Complexity

The computational complexity of the proposed method mainly concentrates on MDA. Assume that $Y \in R^{m_1 \times m_2}$ is a matrix representation for a pixel in remote sensing data; then, we find two optimal transformation matrices to get a low-dimensional representation of $Y$ in discriminant analysis. For ease of understanding, assume that the matrix has uniform dimension numbers in two directions, i.e., $m_1 = m_2 = m$. The complexity of LDA is $O(m^6)$. On the other hand, in MDA, the complexity of computing the scatter matrices is $O(2 \times m^3)$, and the complexity for optimization is $O(2 \times m^3)$ for each loop, which is much lower than that of LDA. To further demonstrate this point, we quantitatively compared the computation time of different methods on three data sets. All the experiments were carried out on a personal computer (Intel Core 3.4-GHz processor with 16-GB random access memory). The software implementation was performed using MATLAB (MathWorks, Inc.). The time comparison is shown in Table VIII. From this table, we can see that the time cost of our proposed method...
is similar to that of PT+MDA but less than that of PT+LDA. Note that it costs little time in test stage for our proposed method because MDA only needs matrix product operations, which is very fast in MATLAB.

IV. Conclusion

In this paper, we have proposed a novel matrix-based discriminant subspace ensemble method for HS image spatial–spectral feature fusion. In the proposed method, matrix-based spatial–spectral feature representation is designed for each pixel to capture the local spatial contextual and the spectral information of all the bands, which can well preserve the spatial–spectral correlation. Then, MDA is adopted to learn the discriminative feature subspace for classification. To further improve the performance of discriminative subspace, a random sampling technique is used to produce a subspace ensemble for final HS image classification. By conducting experiments on three data sets collected by different instruments (AVIRIS and ROSIS), we compared the proposed method with the classical pixelwise methods and the vectorized spectral–spatial fusion methods. The experimental results indicate that using spatial information improves the classification performance and results in more homogeneous regions in classification maps compared with only using spectral information. In addition, the proposed method can improve the OA, AA, and \( \kappa \) on three data sets when compared with the vectorized feature fusion methods. We also evaluated the influences of different window sizes and reduced dimensionalities on the classification performance. In general, the best window size depends on the spatial resolution and the size of objects in the data set, whereas the optimal reduced dimensionalities rely on the specific data set.

ACKNOWLEDGMENT

The authors would like to thank Prof. M. Crawford for providing the KSC data set, Prof. P. Gamba for providing the University of Pavia data set, Prof. D. Landgrebe for providing the Indian Pines data set, and Prof. C. Lin for providing the LIBSVM toolbox. The authors would also like to thank the handling editor and two anonymous reviewers for their detailed and constructive comments and suggestions, which greatly helped them to improve the quality of this paper.

REFERENCES


Renlong Hang

Renlong Hang received the B.S. and M.S. degrees from the Nanjing University of Information Science and Technology, Nanjing, China, in 2011 and 2014, respectively. He is currently working toward the Ph.D. degree in the School of Atmospheric Science, Nanjing University of Information Science and Technology. His current research interests include machine learning and pattern recognition.

Qingshan Liu (M’05–SM’07) received the M.S. degree from Southeast University, Nanjing, China, in 2000 and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2003. From 2010 to 2011, he was an Assistant Research Professor with the Department of Computer Science, Computational Biomedicine Imaging and Modeling Center, Rutgers, The State University of New Jersey, Piscataway, NJ, USA. Before he joined Rutgers University, he was an Associate Professor with the National Laboratory of Pattern Recognition, Chinese Academy of Sciences. During June 2004 and April 2005, he was an Associate Researcher with the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong. He is currently a Professor with the School of Information and Control Engineering, Nanjing University of Information Science and Technology, Nanjing. His research interests are image and vision analysis. Dr. Liu was a recipient of the President Scholarship of the Chinese Academy of Sciences in 2003.

Huihui Song received the B.S. degree in technology and science of electronic information from the Ocean University of China, Qingdao, China, in 2008, the Master’s degree in communication and information system from the University of Science and Technology of China, Hefei, China, in 2011, and the Ph.D. degree in geography and resource management from The Chinese University of Hong Kong, Hong Kong, in 2014. She is currently a Professor with the Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing, China. Her research interests include remote sensing image processing and image fusion.

Yubao Sun received the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2010. He is currently an Assistant Professor with the School of Information and Control Engineering, Nanjing University of Information Science and Technology. His current research interests are sparse representation theory and application, low-rank representation, video analysis and understanding, and graph and hypergraph learning.