Spatial–Spectral Hyperspectral Image Classification Using Random Multiscale Representation

Jianjun Liu, Member, IEEE, Zebin Wu, Member, IEEE, Jun Li, Member, IEEE, Liang Xiao, Member, IEEE, Antonio Plaza, Fellow, IEEE, and Jon Atli Benediktsson, Fellow, IEEE

Abstract—This paper presents a novel spatial–spectral classification method for remotely sensed hyperspectral images. First of all, a multiscale representation technique based on random projection, referred as random multiscale representation (RMSR), is proposed to extract the spatial features from the given scene. The idea behind RMSR is to properly model the spatial characteristics comprised by each pixel vector and its neighbors by some criteria computed at all reasonable scales, and then compress the implicit high-dimensional spatial features by using a very sparse measurement matrix that approximately preserves the salient spatial information. The entire process is explicitly performed by computing simple criteria (i.e., the first two moments) at rectangular scales of random bands, according to the nonzero entries of the sparse measurement matrix. Subsequently, a composite kernel framework is utilized to balance the extracted spatial features and the original spectral features in the classifier. Our proposed method is shown to be effective for hyperspectral image classification purposes. Specifically, our experimental results with hyperspectral images collected by the airborne visible/infrared imaging spectrometer and the reflective optics spectrographic imaging system demonstrate the effectiveness of the proposed method as compared to other state-of-the-art spatial–spectral classifiers.

Index Terms—Composite kernels, compressive sensing, hyperspectral, image classification, multiscale representation, random projection.

I. INTRODUCTION

SUPERVISED classification is an important task in hyperspectral image analysis [1], where pixels are assigned to one of the available classes according to a set of given training pixels. During the last decade, several classifiers have been extended for supervised classification of hyperspectral images, such as support vector machines (SVMs) [2], [3], multinomial logistic regression (MLR) [4], [5], and sparse representation classification [6], [7].

In order to improve the classification performance of these pixelwise classifiers, several approaches have included spatial-contextual information during the classification process. Generally, those spatial–spectral techniques can be roughly divided into two categories:

1) On one hand, some techniques extract the spatial features first, and then incorporate both the spatial and spectral features into the classification process. In this context, several techniques have been exploited, such as mean and median filters [8], [9], Gabor wavelets [10], morphological profiles [11], [12], and mean shift (MS) techniques [13]. After extracting the spatial features, composite kernels (CKs) [8], generalized CKs [14], and multiple kernels [15]–[17] have been used to perform the final classification by considering the spatial features in addition to the spectral features.

2) On the other hand, several other techniques integrate the spatial and spectral information separately. For instance, in [18] the probabilistic SVM is first used to estimate class conditional probability density functions, and then context-based class priors are estimated by using Markov random fields (MRF). In [19], a similar approach is proposed that can be solved in a Bayesian framework, which uses the MLR to learn the posterior probability distributions from the spectral information contained in the data, and then uses the MRF to include spatial-contextual information in the classifier. In [20]–[22], the original hyperspectral image is first classified by a pixelwise classifier and simultaneously segmented into some adaptive neighborhoods by using some segmentation techniques, such as partitional clustering [20], morphological watersheds [21], and minimum spanning forests [22], and then majority voting is adopted to integrate the pixelwise classification results and the obtained segmentation map.
This paper considers designing the spatial–spectral classification method by the first strategy, and mainly concentrates on the exploitation of spatial features that play a fundamental role in this classification process. Generally, the spatial characteristics of a pixel are related to its surrounding pixels. As a result, the spatial feature extraction techniques usually define a fixed spatial neighborhood by a sliding window [8] or create an adaptive spatial neighborhood relating to the parameter value of the filter considered [9], and then compute spatial criteria within the obtained neighborhoods/regions. In order to provide additional discriminant information related to the structures of different objects in a scene, multiscale representation techniques repeating the aforementioned process with different scales (i.e., neighborhoods) to perform a multiscale (or a multilevel) analysis, have been proposed in the literature. For instance, in [11] the derivative of the morphological profile is used to isolate different structures in the image, where the multiscale morphological profile is based on the repeated use of openings and closings with a structuring element of increasing size. In [23], a multilevel context-based system is used for the classification of very high resolution images. In [24], a so-called extended multiattribute profile (EMAP), which is based on the sequential application of different types of attribute filters to a hyperspectral image, is proposed, and the multilevel spatial features are obtained by stacking all filtered images together in the same data structure. In [25], a multiscale MS analysis approach is adopted to compute the multiscale representation of hyperspectral images by using different spatial bandwidths. Several other strategies have been proposed in the literature for multiscale classification [26] and modeling of spatial-contextual information [27] that are not further described here for space considerations.

In the aforementioned multiscale representation techniques, a sequence of scales are usually necessary in order to properly model the spatial information of different objects. Nevertheless, considering a large number of scales can result in high-dimensional feature vectors that will bring a heavy burden on computation and storage. In order to address the problems of high-dimensional features, most methods based on multiscale representation consider to select a limited number of optimal scales empirically [11], [24], [25]. It is obvious that these methods are both unsuitable for an automatic analysis of data and suboptimal since they might not provide a complete characterization of the spatial information [28]. Indeed, if there are two pixels that respectively belong to two similar objects that are not separated by the limited scales, it is difficult for these methods to distinguish the two pixels. In addition, there are other methods that consider feature extraction [29] and feature selection [28] to solve the problem of high-dimensional features. However, these methods cannot decrease the memory demands and computational complexity significantly, as the feature reduction process is generally data dependent and time consuming, and one needs to store the spatial features in memory throughout the entire multiscale representation process.

In view of the aforementioned issues, this paper first proposes a new multiscale representation technique based on random projection for extracting the spatial features of hyperspectral images, and then designs a spatial–spectral classification method by utilizing a CK framework [8] to balance the extracted spatial features and the original spectral features in the classification process. The main idea behind the proposed multiscale representation technique, referred to as random multiscale representation (RMSR), is to implicitly\(^1\) construct a dense multiscale representation (DMSR) by the first two moments computed at all reasonable rectangular scales in order to provide a complete characterization of the spatial information of the given image, and then compress the implicit high-dimensional spatial features by using a very sparse measurement matrix [30] that approximately preserves the salient spatial information. The entire process of RMSR is explicitly performed by computing the investigated criteria (i.e., the first two moments) at rectangular scales of random bands, according to the nonzero entries of the sparse measurement matrix, thus avoiding the selection of optimal scales and decreasing the memory demands and computational complexity significantly. Moreover, the computation of the first two moments can be made independent of the scale size by taking advantage of the integral image method in [31]. The novelty of the proposed method relies on the exploitation of the sparse measurement matrix to propose a new multiscale representation technique RMSR for extracting the spatial features of hyperspectral images, which presents several innovative contributions with regards to the existing multiscale representation methods.

1) First and foremost, RMSR neither relies on the empirical selection of optimal scales nor on the feature reduction of spatial features, thus providing an alternative solution for the multiscale representation of hyperspectral images.

2) Second, RMSR can efficiently handle a large number of scales (even dense scales) and, therefore, it can benefit from the complementary information obtained in dense scales, so that the spatial characteristics of a pixel can be properly captured and the simple criteria can be utilized to describe the spatial information within a given scale.

3) Finally, RMSR is general since the criteria available for it are not limited only to the investigated first two moments, and one can easily extend the criteria used in the existing multiscale representation methods to it.

The remainder of this paper is organized as follows. Section II first briefly introduces the very sparse random measurement matrix that will be used in the proposed method, and then proposes RMSR for extracting the spatial features as well as the final spatial–spectral classification method based on the CK technique. The effectiveness of the proposed method is demonstrated in Section III by a series of experiments with two real hyperspectral images, collected by the airborne visible/infrared imaging spectrometer (AVIRIS) and the reflective optics spectrographic imaging system (ROSIS). These experiments demonstrate the effectiveness of the proposed method as compared to other state-of-the-art spatial-spectral classifiers. Finally, Section IV concludes the paper with some remarks and hints at plausible future research lines.

\(^1\)The modifier ‘implicitly’ means the following process that just stands in theory does not exist in practice.
II. PROPOSED SPATIAL–SPECTRAL CLASSIFICATION METHOD

Fig. 1 shows a flowchart of the proposed spatial–spectral classification method, which consists of two main steps: 1) extraction of the spatial random multiscale features by using RMSR; and 2) integration of spatial and spectral information in a CK framework using the SVM as the baseline classifier. In the remainder of this section, we describe in more details the strategies adopted for RMSR and final spatial–spectral classification.

A. Very Sparse Random Measurement Matrix

This section briefly introduces the very sparse random measurement matrix that will be used in RMSR. In random projection, a high-dimensional feature vector \( a \in \mathbb{R}^m \) is projected to a lower dimensional feature vector \( v \in \mathbb{R}^n \) by using a random matrix \( R \in \mathbb{R}^{n \times m} \) as follows:

\[
v = Ra
\]

where the columns of \( R \) have unit length, and \( n < m \). The Johnson–Lindenstrauss (JL) lemma [32] indicates that if the dimensionality \( n \) is suitably high, there exists such a random matrix \( R \) that can provide a stable embedding. Specifically, given an integer \( d \) and let \( n \geq n_0 = O(\epsilon^{-2} \ln(d)) \) with \( \epsilon > 0 \), then for any two vectors \( a_i, a_j \) in a finite collection \( X \) of \( d \) vectors in \( \mathbb{R}^m \), we have

\[
(1 - \epsilon)||a_i - a_j||_2^2 \leq ||Ra_i - Ra_j||_2^2 \leq (1 + \epsilon)||a_i - a_j||_2^2
\]

The random Gaussian matrix \( R \in \mathbb{R}^{n \times m} \) is a typical matrix satisfying the JL lemma, where the entry \( R_{ij} \) follows a zero-mean and unit-variance Gaussian distribution. Since the random Gaussian matrix \( R \) is dense, it involves significant memory and computation requirements when \( m \) is large. To overcome this problem, in [33] a sparse random matrix \( R \) is introduced with its entries defined as follows:

\[
R_{ij} = \sqrt{\rho} \times \begin{cases} 1 & \text{with probability } \frac{1}{\rho} \\ 0 & \text{with probability } 1 - \frac{1}{\rho} \\ -1 & \text{with probability } \frac{1}{\rho} \end{cases}
\]

where \( \rho = 3.2 \) and it is proved that if \( n \geq (4 + 2\beta)(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln(d) \) with \( \beta > 0 \), the statement (2) holds true with probability at least \( 1 - d^{-\beta} \). This matrix \( R \) is easy to compute by using a uniform random generator. In [34], it is proved that, for \( \rho = O(m) \) (\( a \in \mathbb{R}^m \)) even \( \rho = m/\ln(m) \), the sparse random matrix \( R \) is almost as effective as the conventional random Gaussian matrix. Moreover, in [30] good results have been shown for visual tracking by setting \( \rho = m/c \) with a small \( c \), i.e., there are only about \( c \) nonzero entries in each row of \( R \). Although a large \( c \) is beneficial to improve the accuracy as shown in [34], we set \( c = 4 \) in our experiments as a tradeoff between accuracy and computational cost (see Section III-B). Therefore, the adopted random matrix \( R \) with \( \rho = m/4 \) is very sparse, and the computational cost is very low (only \( O(4n) \)). In addition, the memory requirements are also very light, as we only need to store the nonzero entries of \( R \).

B. Random Multiscale Representation

Recent works on multiscale representations for hyperspectral imagery have shown that it is necessary to include a large number of scales during the representation process [11]–[13], [24], [25]. Inspired by this observation, we intend to use dense rectangular scales to provide a complete characterization of the spatial information, i.e., for each pixel vector, all possible rectangular neighborhoods centered at the pixel should be ideally included. Intuitively, increasing the number of scales is beneficial for the classification performance, since the more scales are used, the more spatial characteristics are provided.

Let us denote a hyperspectral image by \( X \in \mathbb{R}^{L \times W 	imes H} \), with \( L \) being the number of bands and \( W \times H \) being the spatial dimensions. For each pixel vector \( x \in \mathbb{R}^L \) of \( X \), its DMSR is constructed in this paper by computing some criteria of its rectangular neighborhoods at dense scales \( \{F_{1,1}, F_{1,2}, ..., F_{w,h}\} \), i.e.,

\[
F_{w,h}(x) = f(\Omega_{x,w,h}^m)
\]

where \( f(\cdot) \) denotes the criterion function per spectral channel, \( \Omega_{x,w,h}^m \) is the rectangular neighborhood centered at \( x \),

\[
\text{where } \rho = 1 \text{ also holds true, but the random matrix } R \text{ is not sparse.}
\]

\[
The columns of the sparse random matrix \( R \) are unnormalized, thus the statement (2) should be rewritten as \( (1 - \epsilon)||a_i - a_j||_2^2 \leq (1/n)||Ra_i - Ra_j||_2^2 \leq (1 + \epsilon)||a_i - a_j||_2^2 \).
\]
and $2w + 1$ and $2h + 1$ are the width and height of $\Omega_{x,h}^w$, respectively. Here, we pad $X$ with minor reflections of itself to deal with the outliers. Then, we concatenate these vectors $F_{w,h}(x)$ to form a high-dimensional multiscale feature vector $a = [F_{1,1}(x); F_{1,2}(x); \ldots; F_{w,h}(x)] \in \mathbb{R}^m$ where $m = whL$. Fig. 2 shows a graphical illustration of the aforementioned procedure to construct the DMSR of the hyperspectral image.

Once the neighborhoods of each pixel vector have been defined, the DMSR features are extracted by computing some criteria associated to the considered neighborhoods. As discussed in [8], the simplest (but still powerful) spatial features are based on the application of moment criteria. For simplicity, the criteria investigated in the proposed method are the first two moments, i.e., mean and standard deviation. Note that, the integral image method can be used to accelerate the computation of the mean [31]. Specifically, for the given hyperspectral image $X \in \mathbb{R}^{L \times W \times H}$, the entry of its integral image $I \in \mathbb{R}^{L \times W \times H}$ at a spatial location $(x, y)$ represents the summation of all pixels in $X$ per spectral channel within a rectangular region formed by the origin and $(x, y)$

$$\bar{X}(l, x, y) = \sum_{i=1}^{x} \sum_{j=1}^{y} X(l, i, j), l = 1, 2, \ldots, L. \quad (5)$$

After computing the integral image, we just need three additions to calculate the summation of intensities over any rectangular region, as illustrated graphically in Fig. 3. Then, for a given neighborhood $\Omega_{x,h}^w$ centered at $x = X(\cdot, x, y) \in \mathbb{R}^L$, the mean function $f_{\bar{\mu}}(\cdot)$ can be calculated by the following expression:

$$f_{\bar{\mu}}(\Omega_{x,h}^w) = \frac{I(; x + w, y + h) - I(; x + w, y - h - 1) - I(; x - w - 1, y + h) + I(; x - w - 1, y - h - 1)}{N}$$

where $N = (2w + 1)(2h + 1)$. As for the standard deviation function $f_{\bar{\sigma}}(\cdot)$, it can be written as follows:

$$f_{\bar{\sigma}}(\Omega_{x,h}^w) = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x(i) - \bar{m})^2}$$

$$= \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} x(i)^2 - \frac{N}{N-1} \bar{m}^2} \quad (7)$$

where $x(i)$ is the $i$th pixel vector in $\Omega_{x,h}^w$, $\bar{m} = f_{\bar{\mu}}(\Omega_{x,h}^w)$ and $(\cdot)^2$ denotes the component-wise application of the square function. It is easy to see that the integral image method can also be used to solve (7) if we precompute the integral images of $X$ and $X^2$. Hence, the computational complexity of the first two moments is independent of the size of the given rectangular scale. This is important for the proposed method, as we intend to use dense scales in order to properly model the spatial-contextual information.

After obtaining the DMSR features for all pixel vectors, we can use them as spatial features for classification purposes. However, as recommended in the existing multiscale representation methods [11], [24], [25], [28], [29], it is impractical to directly use the obtained DMSR features since they lie in a high-dimensional space $\mathbb{R}^m$. As discussed in the Section I, the methods based on the manual setting of optimal scales might not provide a complete characterization of the spatial information, and the methods based on feature reduction cannot decrease the memory demands and computational complexity significantly. Thus, it is necessary to design an effective mechanism to address the problems of the DMSR features. Moreover, the dimensionality $m$ of the DMSR features is typically in the order of $10^6$ (if a given image contains 100 spectral bands with size of 200 x 200 pixels, then $m = whL \approx 100 \times 100 \times 100 = 10^6$) and, therefore, the feature reduction techniques adopted in the existing multiscale representation methods are impractical for the proposed DMSR due to the limited memory load and compute capacity.

In order to circumvent the aforementioned limitations, we construct a very sparse random measurement matrix $R \in \mathbb{R}^{n \times m}$ introduced in Section II-A to project every DMSR feature vector $a \in \mathbb{R}^m$ onto a low-dimensional vector $v \in \mathbb{R}^n$. The entire process composed by the DMSR and the random matrix $R$ is referred as RMSR. Since the random matrix $R$ is data independent and very sparse, RMSR can be organically performed by computing the investigated criteria (i.e., the first two moments) at random rectangular scales of random bands according to the nonzero entries of $R$ as shown in Fig. 4. Thus, we do not need to store the DMSR features, and the computational cost of RMSR is very light.
It is worth noting that, although the very sparse measurement matrix $R$ can provide a stable embedding that approximately preserves the salient spatial information, the dimensionality $n$ of it should be suitably high as stated in the JL lemma [32]. In [33], a JL-bound (introduced in Section II-A) has been given for the sparse measurement matrix. For hyperspectral images with about $d \approx 10^5$ pixels, $\epsilon = 0.2$, and $\beta = 1$, the lower bound for $n$ is approximately 4000. However, this JL-bound is conservative in many applications [34]. In [35], the results show that the JL-bound is much higher than that, which is sufficient to give good results on image and text data. As for hyperspectral images, this also holds true as shown in Section III-B.

To interpret the aforementioned issue, one can resort to the perspective of compressive sensing. In compressive sensing [36]–[38], it is shown that, for random projection (1), if $R$ satisfies the restricted isometry property (RIP) [36], [39] and $a$ is sparse, then the compressive measurement $v$ approximately preserves the salient information in $a$, i.e.,

$$
(1 - \epsilon)||a||_2^2 \leq ||Ra||_2^2 \leq (1 + \epsilon)||a||_2^2.
$$

(8)

In [40], it is proved that any random matrix $R$ satisfying the JL lemma can also satisfy the RIP in compressive sensing. The random matrix $R$ adopted in RMR asymptotically satisfies the JL lemma, thus holding true for the statement (8). Moreover, the statement (2) is equal to (8) by setting $a = a_i - a_j$. In other words, if $a_i - a_j$ is sparse, the bound $n \geq \kappa \beta \log(m/\beta)$ [30] (with constants $\kappa$ and $\beta$) derived from the RIP is available. It is apparent that the pairwise differences of the DMSR features are very sparse and highly compressible, since dense scales are used and related features are very similar to each other. In this paper, as recommend in [30], it is expected that $n \geq 50$ when $m = 10^6$, $\kappa = 1$ and $\beta = 10$, which is much tighter than the JL-bound.

Furthermore, we should highlight that the proposed RMR is general, since the criteria available for it are not limited only to the investigated first two moments, and other criteria that can be used to filter an image within a sliding window, such as median and singular value, are all available. In order to illustrate the generalization of RMR, the median criterion is also used for evaluation purposes in our experiments. Moreover, one can easily extend the adaptive filters used in the existing multiscale representation methods to RMR by replacing the rectangle neighborhoods with the adaptive neighborhoods, but this is beyond the scope of this paper.

Next, we graphically illustrate how the proposed RMR features separate the classes. Specifically, the AVIRIS Indian Pines dataset is tested, and three classes $C_2$, $C_3$, and $C_4$ (see Section III-A) are considered. Both the spectral features and the RMR features of the three classes are reduced by principal component analysis (PCA), and the first three principal components (PCs) are retained. The scatterplots depicting the reduced vectors of the spectral features and the RMR features are shown in Fig. 5(a) and (b), respectively. It is clear that the RMR features can increase the margin between the classes and improve the smoothness within class.
C. Spatial–Spectral Classification With CK

After performing the spatial feature extraction process of RMSR, each pixel has three kinds of features: spectral features \( \mathbf{x} \), which are the original values of each pixel vector's elements; and spatial features \( \mu_\mathbf{x} \) and \( \sigma_\mathbf{x} \), which are the projective mean and standard deviation values computed within each pixel’s neighborhoods, respectively. In this paper, we adopt the CK framework [8] to combine the spatial and spectral features for the final classification, where the used kernel is the Gaussian radial basis function (RBF)

\[
K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2\right), \gamma \in \mathbb{R}^+.
\]  

(9)

In our context, the spatial–spectral CK \( K \) is defined as

\[
K(i, j) = \nu K(\mathbf{x}_i, \mathbf{x}_j) + (1 - \nu)K\left(\left[\mu_\mathbf{x}_i, \sigma_\mathbf{x}_i\right], \left[\mu_\mathbf{x}_j, \sigma_\mathbf{x}_j\right]\right)
\]

(10)

where \( 0 \leq \nu \leq 1 \) controls the relative weight of the spatial and spectral information.

Once the spatial–spectral kernel \( K \) has been constructed, the next step is to perform the final classification procedure. Here, we adopt the SVM (embedded in the aforementioned CK framework) to conduct the final classification. We have selected the SVM because it is one of the most successful kernel classifiers and is intrinsically less sensitive to the high dimensionality of the feature space.

It is worth noting that in this paper, the integration of RMSR into a RBF-based CK has been taken with overall consideration of the following three issues:

1) Recent works [41] have highlighted the importance of kernel-based methods in the classification of hyperspectral images, thus we only consider investigating the proposed RMSR features by using kernel-based classifiers.

2) The incorporation of spectral features is beneficial to the classification accuracy of spatial features, and thus the CK, which is widely used to combine the spectral and spatial features [1], is adopted.

3) Only the pairwise \( l_2 \)-norm distance of the DMSR features is maintained by RMSR as shown in (2) and the RBF kernel is just the suitable one that is based on \( l_2 \)-norm.

III. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we first introduce the two hyperspectral datasets used in experiments, and then present a series of experimental results to validate the proposed method. As mentioned before, an SVM classifier (embedded in a CK framework for spatial–spectral classification) is used, where the LIBSVM software package [42] has been used for the specific implementation.\(^5\) The free parameters of the SVM have been carefully optimized via a cross-validation procedure. More specifically, the RBF-kernel parameter is varied in the range \( \gamma \in \{2^{-4}, ..., 2^{4}\} \), and the regularization parameter is varied in the range \( C \in \{10^0, ..., 10^5\} \). As for the free parameter \( \nu \) of the CK framework, it is varied in the range \([0, 1]\). A detailed explanation of the procedure followed to optimize parameter settings is given in [2] and [8], and is not repeated here for space considerations. To evaluate the performance of the classification methods, three widely used quality metrics are adopted, i.e., the overall accuracy (OA), the average accuracy (AA), and the Kappa coefficient (\( \kappa \)). In addition, the original data have been scaled in the range \([0, 1]\) before the experiments and, unless otherwise specified, the quantitative measures are obtained by averaging ten Monte Carlo runs. Besides the first two moments, another criterion (i.e., the median) is also used to implement the proposed method. In the following, the proposed method implemented with moment criteria is abbreviated by MOM and the one implemented using the median criterion is abbreviated by MED.

A. Hyperspectral Imagery Datasets

The first hyperspectral image dataset used in our experiments is an image collected by the AVIRIS over NW Indiana’s Indian Pines region in 1992. The AVIRIS sensor collects 220 bands, covering the wavelength range of 0.4–2.5 \( \mu \text{m} \), and the number of retained bands is 200 (after removing 20 water absorption bands). This image consists of 145 \( \times \) 145 pixels, with a nominal spectral resolution of 10 nm and moderate spatial resolution of 20 m by pixel. It contains 16 ground reference classes ranging from 20 to 2468 pixels in size and, for each class, we randomly choose around 5% of the labeled samples for training and use the remaining 95% for testing. The details of training and test sets are shown in Table I. For illustrative purposes, Fig. 6 shows a false color composite of the scene and the associated ground reference map.

<table>
<thead>
<tr>
<th>Class</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1 Alfalfa</td>
<td>3</td>
</tr>
<tr>
<td>C2 Corn-no till</td>
<td>72</td>
</tr>
<tr>
<td>C3 Corn-min till</td>
<td>42</td>
</tr>
<tr>
<td>C4 Corn</td>
<td>12</td>
</tr>
<tr>
<td>C5 Grass/pasture</td>
<td>25</td>
</tr>
<tr>
<td>C6 Grass/trees</td>
<td>38</td>
</tr>
<tr>
<td>C7 Grass/pasture-mowed</td>
<td>2</td>
</tr>
<tr>
<td>C8 Hay-windrowed</td>
<td>25</td>
</tr>
<tr>
<td>C9 Oats</td>
<td>2</td>
</tr>
<tr>
<td>C10 Soybeans-no till</td>
<td>49</td>
</tr>
<tr>
<td>C11 Soybeans-min till</td>
<td>124</td>
</tr>
<tr>
<td>C12 Soybean-clean till</td>
<td>31</td>
</tr>
<tr>
<td>C13 Wheat</td>
<td>11</td>
</tr>
<tr>
<td>C14 Woods</td>
<td>65</td>
</tr>
<tr>
<td>C15 Bldg-grass-tree drives</td>
<td>19</td>
</tr>
<tr>
<td>C16 Stone-steel towers</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>525 9841</td>
</tr>
</tbody>
</table>

\(^5\)In LIBSVM, a “one-against-one” strategy is implemented for multiclass classification.
The scene contains 610 × 340 pixels, with very high spatial resolution of 1.3 m per pixel. There are nine ground reference classes of interest, and we randomly choose 40 samples per class for training and use the rest for testing. The details of training and test sets are shown in Table II. In addition, a false color composite image and the associated ground reference map are shown in Fig. 7.

### B. Different Sparse Random Measurement Matrices

There are two tuning parameters in the construction of the very sparse random measurement matrix $\mathbf{R} \in \mathbb{R}^{n \times m}$: $c$ controls the number of the nonzero entries in each row of $\mathbf{R}$, and $n$ is the dimensionality of the projection vectors (i.e., the number of spatial features). The first set of experiments is designed to analyze the impact of these two parameters, and the criterion is the distortion error measured by comparing the $l_2$-norm distance between two projection data vectors to their $l_2$-norm distance in the original high-dimensional space, which is defined as

$$
\epsilon(a_i, a_j) = 1 - \frac{1}{\sqrt{n}} \cdot \frac{\|\mathbf{R}a_i - \mathbf{R}a_j\|_2}{\|a_i - a_j\|_2}.
$$

(11)

Here, we randomly choose 1000 pairs of data vectors per dataset for testing, and compute the error between members of a pair of data vectors, averaged over these pairs, for each $\mathbf{R}$. $c$ is varied to be 2, 4, 10, 50, and 100, and $n$ is varied in steps of 20 in the range [20, 1000]. At each pair $c$ and $n$, $\mathbf{R}$ is generated anew 100 times to obtain the average error. Fig. 8 shows the results for the two given datasets. As for the free parameters $w$ and $h$ in the multiple scales $\{F_{w, h}\}_{w, h=1, j=1}^W$, they were set equal to each other. More specifically, for Indian Pines dataset we used, $w = h = 50$, and for University of Pavia dataset we used, $w = h = 100$.

As shown by Fig. 8, it is clearly seen that random projection can yield very accurate results. When $n$ is analyzed, the performance obtained by using more features is better than that obtained using less features, but the differences observed between both cases are quite slight when $n > 100$. When $c$ is analyzed, the results obtained by using more nonzero entries are better than that obtained using less nonzero entries in most cases, but the gaps among the results of different $c$ are not significant. As a result, we can conclude that the construction of $\mathbf{R}$ is robust to parameter settings, and almost all the information provided by the DMSR features is preserved by RMSR. In the following, we assume that $c$ is fixed to 4 and $n$ is fixed to 200 as a tradeoff between accuracy and computational cost.

In the next set of experiments, we investigate the sparse random projection by statistically analyzing the errors of using different $\mathbf{R}$. Different from the first set of experiments, the
Fig. 8. Influence of the parameters of $R$. (a) AVIRIS Indian Pines scene. (b) ROSIS University of Pavia scene.

Fig. 9. Cumulative distribution probability of error.

suggested parameters $c$ and $n$ are used and $R$ is generated anew $10^4$ times to avoid any bias induced by random generator. Fig. 9 shows the cumulative distribution probability of error for the two given datasets. It is apparent that the sparse random projection can achieve a low error with a high probability. Thus, we can conclude that the sparse random projection process is robust.

C. Different Numbers of Scales

In this set of experiments, we investigate the impact of the input parameters on RMSR. There are only two parameters: $w$ and $h$ intended for the definition of the multiple scales $\{F_{i,j}\}_{i=1,j=1}$, which control the scale size of the largest rectangle neighborhood. The product of $w$ and $h$ gives the number of scales used by RMSR, and their values should not be more than half of the width and height of the given dataset, respectively. Several values for $w$ and $h$ have been tested in our experiments with the two considered hyperspectral datasets, and the obtained classification accuracies are reported in Fig. 10(a) and (b), respectively. In these experiments and the following, the parameters $w$ and $h$ are set equal to each other, and the error bars indicate the standard deviation for ten random samplings.

From Fig. 10, it can be seen that for both the two given datasets, when the number of scales is small, MED performs better than MOM in most cases; whereas when the number of scales is large, MOM outperforms MED consistently. As the number of scales increases, the classification accuracies of both MED and MOM increase almost monotonically. This indicates that a large number of scales have advantages in improving classification performance. However, the accuracies reach saturation asymptotically when the number of scales is relatively large. This is because the rectangle neighborhoods with large widths or heights might cover many objects that the computed criteria of different classes are almost the same. In view of the above issues, a moderate number of scales are used in the experiments, since a larger number of scales will increase the computational complexity of MED significantly and will not bring significant performance improvements. More specifically, we hereinafter set parameters $w$ and $h$ to a fixed value of 50 for the AVIRIS Indian Pines scene and to a fixed value of 100 for the University of Pavia scene. These empirical parameter settings were found appropriate after simple experimentation, which indicate that these parameters are not difficult to set in practice.

D. Influence of Kernel Weight

The kernel weight $\nu$ in (10) balances the spectral and spatial information in the construction of the proposed method. In this set of experiments, we examine the influence of it to the performance of the investigated MED and MOM by varying $\nu$ from 0 to 1. The classification results for the two given datasets are shown in Fig. 11. It can be seen that when $\nu$ is set to 0 or 1, i.e., only spatial information or spectral information is used, the investigated two methods do not perform very good results on both datasets. The optimal values of $\nu$ are different for the two given datasets, but they are all in the range $[0.1, 0.5]$. When $\nu$ goes from 0.6 to 1, almost all cases degrade performance apparently. All these indicate that the spatial features play an important role in the spatial–spectral classification, and the integration of the spectral features into the spatial features can improve the classification performance when compared with the cases only using the spatial features.
one is the SVMCK, where the spatial features are extracted by both the SVM classifier and the CK framework. The first three widely used spatial–spectral classification methods based on multiscale methods, PCA is adopted as the dimensionality reduction method, and the multiscale MS technique is utilized to extract the spatial features. As reported in [14], the first three PCs are retained, and the area and standard deviation are considered to be beneficial to RMSR. However, this is not the main goal of our approach as we intend to conduct the classification with all the information available at hand, without any prior dimensionality reduction. Thus, in this paper we do not consider using bases images for the construction of the proposed method.

\section{Influence of Base Images}

In traditional multiscale methods, it is suggested to use dimensionality reduction methods to generate the characteristic images and then use the obtained characteristic images as base images to construct the multiscale representations [11]–[13], [24], [25]. In this set of experiments, we analyze the impact of using and not using base images on RMSR. As with most multiscale methods, PCA is adopted as the dimensionality reduction method, and the \( q \) most significant PCs are used to produce the base images. In our experiments, several values for \( q \) were tested, and the corresponding methods for MED and MOM are denoted by MEDp and MOMp, respectively. As for RMSR without using base images, the original spectral features were used, and the corresponding methods are MED and MOM themselves.

A comparison of the results obtained using and not using base images for the classification of the two considered datasets is given in Fig. 12(a) and (b), respectively. From the results reported in Fig. 12, it is clear that using base images could be beneficial to RMSR. However, this is not the main goal of our approach as we intend to conduct the classification with all the information available at hand, without any prior dimensionality reduction. Thus, in this paper we do not consider using base images for the construction of the proposed method.

\section{Comparison of Different Classification Methods}

In this set of experiments, we compare the proposed MED, MOM, and DMSR (i.e., MOM without random projection) with three widely used spatial–spectral classification methods based on both the SVM classifier and the CK framework. The first one is the SVMCK, where the spatial features are extracted by computing the mean and standard deviation of the neighborhood pixels in a window per spectral signature, and more details can be seen in [8]. The second one is denoted by MSCK, since EMAP is used to build the spatial features. As reported in [14], the first three PCs are retained, and the area and standard deviation are considered to build the morphological attribute profiles. The corresponding parameters and codes of EMAP are provided by [12] and [24].
In order to emphasize the assessment of the proposed RMSR features, the kernel weights $\nu$ of the aforementioned methods are all set to 0.5 for a fair comparison of their spatial features. Moreover, the pixelwise SVM that only uses spectral features [2] is also included as a baseline classifier. The global and class-specific accuracies obtained by all the considered classification methods for the two given datasets are reported in Tables III and IV, and the corresponding classification maps are illustrated in Figs. 13 and 14, respectively. In addition, the processing time in seconds, measured in a 64-b quad-core Intel CPU 2.40-GHz processor, is also included for reference. Here, the processing time of DMSR is not included since it is far more than the others.

For the AVIRIS Indian Pines scene, as reported in Table III, all spatial–spectral methods yield higher classification accuracies when compared with the pixelwise SVM. Among these methods, MOM gives the highest global and most of the best class-specific accuracies, while DMSR can be regarded as the second-best method followed by MED. Note that, DMSR should ideally perform better than MOM, but the experimental results are just the opposite. This is because the dimensionality of DMSR features is so high that the classifier may suffer from the high dimensionality of the feature space. However, this impact is not serious, since the gap between MOM and DMSR is not significant. The classification maps reported in Fig. 13 confirm these observations conducted on Table III. It is also clear that the spatial–spectral methods provide smoother classification maps than the spectral-only SVM. The classification maps obtained by the proposed MOM, DMSR, and MED contain smoother regions than those produced by the other methods, and the misclassified pixels mainly concentrate in the corners that are difficult to be described by all compared spatial features, due to the information asymmetry among the classes lying in these irregular regions. Moreover, although MSCK and EMAPCK are also the methods based on multiscale spatial features, the classification maps of them are not as smooth as those of MOM, DMSR, and MED, and contain some block-like misclassified structures, especially in the flat regions (e.g., rectangle regions in Fig. 13), that are correctly classified by the proposed three methods.

From Table IV, it is also clear that for the ROSIS University of Pavia scene, all spatial–spectral methods perform better than SVM and MOM yields the best classification accuracies among all these methods. The numerical results reported on Table IV can be confirmed by visual inspection of the classification maps shown in Fig. 14.
G. Different Numbers of Training Samples

In this set of experiments, we analyze how the number of training samples affects the classification performance of the classification methods compared in Section III-F for the two considered datasets. Here, DMSR is not included since it is the coarse version of MOM. The parameters of these methods are fixed to be the same as those used in Section III-F. For the Indian Pines dataset, in each test we randomly choose 1–20% of the labeled pixels per class for training and the remaining pixels for testing. For very small classes, we take a minimum of two training samples per class. For the ROSIS University of Pavia scene, we build training sets by randomly choosing 10, 20, 40, 60, 80, and 100 training samples per class. Fig. 15(a) and (b) shows the classification accuracies obtained in this set of experiments with the AVIRIS Indian Pines and ROSIS University of Pavia scenes, respectively. From Fig. 15, it is obvious that the OA increases monotonically, and the standard deviation decreases as the number of training samples increases. The proposed MED and MOM consistently yield higher OAs than the other methods considered in experiments.

IV. CONCLUSION AND FUTURE LINES

This paper developed a RMSR technique for extracting the spatial information of hyperspectral images and hereby presented a spatial–spectral classification method. The RMSR technique allows using dense scales to represent the spatial characteristics of the hyperspectral images, so that we can benefit from the complementary information collected at various scales to capture the spatial information around a pixel. The spatial features are extracted by computing criteria on the dense rectangular scales. After obtaining the spatial features on dense scales, we concatenate them as high-dimensional multiscale features to keep all the spatial information extracted in the process. Then, a very sparse random measurement matrix is introduced to compress the high-dimensional spatial features into lower dimensional features without the loss of salient information. The entire process of RMSR is organically performed by computing criteria at random scales of random bands according to the nonzero entries of the very sparse measurement matrix. The criteria investigated in RMSR are the first two moments, which are simple and can be significantly accelerated by the integral image method. Thanks to these effective mechanisms, the computation load of RMSR becomes very light. The final classification is conducted by a CK-based approach that appropriately weights the spatial features with regards to spectral features. The proposed method has been tested on two widely used hyperspectral datasets. Our experiments indicate that use of dense scales by the proposed method can be very helpful to improve the obtained classification results, and RMSR is computationally manageable and highly effective. Although the results obtained by the proposed method are very encouraging, further enhancements such as the use of adaptive neighborhoods and additional criteria to describe spatial neighborhoods should be pursued in future developments.

ACKNOWLEDGMENT

The authors would like to thank Prof. D. Landgrebe from Purdue University for providing the Indian Pines dataset, and Prof. P. Gamba from the University of Pavia, Italy, for providing the ROSIS University of Pavia dataset.

REFERENCES

Antonio Plaza (M’05–SM’07–F’15) received the M.S. and Ph.D. degrees in computer engineering from the University of Extremadura, Badajoz, Spain.

He is an Associate Professor (with accreditation for Full Professor) at the Department of Technology of Computers and Communications, University of Extremadura, where he is also the Head of the Hyperspectral Computing Laboratory (HyperComp). He was the Coordinator of the Hyperspectral Imaging Network. He authored more than 400 publications, including 119 Journal Citation Reports (JCR) journal papers (71 in IEEE journals), 20 book chapters, and more than 240 peer-reviewed conference proceeding papers (94 in IEEE conferences). He was an Associate Editor for IEEE ACCESS, and a Member of the Editorial Board of the IEEE GEOSCIENCE AND REMOTE SENSING NEWSLETTER during 2011–2012 and the IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE in 2013. He was also a Member of the steering committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS and REMOTE SENSING in 2012. He was the Director of Education Activities for the IEEE GEOSCIENCE AND REMOTE SENSING SOCIETY (GRSS) from 2011 to 2012, and since November 2012, he has been the President of the Spanish Chapter of IEEE GRSS. Since January 2013, he has been the Editor-in-Chief of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING JOURNAL.

Dr. Plaza has guest edited seven special issues on JCR journals (three in IEEE journals). He has been the Chair for the IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, in 2011. He received the recognition of Best Reviewers of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2009 and the recognition of Best Reviewers of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2010, a journal for which he has served as an Associate Editor from 2007 to 2012.

Jon Atli Benediktsson (S’84–M’90–SM’99–F’04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984, and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

He is currently a Rector and a Professor of electrical and computer engineering at the University of Iceland. He is the cofounder of the biomedical startup company Oxymap. His research interests include remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing. He has published extensively in those fields. He is a Fellow of The International Society of Optics and Photonics and a Member of the Association of Chartered Engineers in Iceland (VFI) and Societas Scinetiarum Islandica. During 2011–2012, he was the President of the IEEE Geoscience and Remote Sensing Society (GRSS) and has been on the GRSS AdCom since 2000. He has been an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) since 1999, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003, and the IEEE ACCESS since 2013. He was an Editor of TGRS from 2003 to 2008. He was on the International Editorial Board of the International Journal of Image and Data Fusion and was the Chairman of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2007 to 2010.

Prof. Benediktsson received the Stevan J. Kristof Award from Purdue University as outstanding graduate student in remote sensing in 1991, the Icelandic Research Councils Outstanding Young Researcher Award in 1997, and the IEEE Third Millennium Medal in 2000. He was a corecipient of the University of Iceland’s Technology Innovation Award in 2004, the 2012 TGRS Paper Award, and the IEEE GRSS Highest Impact Paper Award in 2013. He also received the yearly research award from the Engineering Research Institute, University of Iceland in 2006, the Outstanding Service Award from the IEEE GRSS in 2007, and the IEEE/VFI Electrical Engineer of the Year Award in 2013. He is also a Member of Tau Beta Pi.