Multi-feature Hyperspectral Image Classification with Local and Non-local Spatial Information via Markov Random Field in Semantic Space

Xiangrong Zhang, *Senior Member, IEEE*, Zeyu Gao, Licheng Jiao, *Senior Member, IEEE,*

Huiyu Zhou

***Abstract*—Hyperspectral images (HSIs) provide invaluable information in both spectral and spatial domains for image classification tasks. In this paper, we use semantic representation as a middle-level feature to describe image pixels' characteristics. Deriving effective semantic representation is critical for achieving good classification performance. Since different image descriptors depict characteristics from different perspectives, combining multiple features in the same semantic space makes semantic representation more meaningful. First, a probabilistic support vector machine is used to generate semantic representation based multi-features. In order to derive better semantic representation, we introduce a new adaptive spatial regularizer which well exploits the local spatial information, while a non-local regularizer is also used to search for global patch-pair similarities in the whole image. We combine multiple features with local and non-local spatial constraints using an extended Markov random field model in the semantic space. Experimental results on three hyperspectral data sets show that the proposed method provides better performance than several state of the art techniques in terms of region uniformity, overall accuracy, average accuracy, and Kappa statistics.**

***Index Terms*—Hyperspectral image classification, semantic representation, Markov random field (MRF), non-local spatial constraint**

I. Introduction

In the past few decades, hyperspectral images (HSIs) have been frequently used in earth observation. Hyperspectral imaging sensors capture images at hundreds of spectral bands with a spatial resolution ranging from 0.75 to 20 m per pixel for airborne sensors (e.g. AVIRIS from NASA), and 5 to 506 m per pixel for satellite sensors (e.g. EO-1 Hyperion from NASA and PROBA CHRIS from ESA). High spectral resolutions provide useful information for discriminating different materials and objects, and thus HSIs have a variety of applications in various areas, such as precision agriculture [1], environmental monitoring [2], and military operations [3]. Classification is one of the most popular applications in the analysis of HSIs. High dimensionality of HSIs makes it possible to achieve accurate object identification and classification but also causes a challenging problem as training samples can be limited, which is known as the Hughes effect [4].

A variety of spectral pixel-wise classifiers [5]-[8] have been proposed to solve this problem. Among these established methods, support vector machine (SVM) classifiers [5], to some degree, have shown promising success in HSI classification and gained large attention due to their robust performance in a high-dimensional feature space with the ability of dealing with a small number of training samples. Recently, many spectral-spatial classification techniques have been proposed based on the assumption that image pixels from a local region usually belong to the same class. There are many ways to impose spatial information, such as post-processing techniques [9], [10], composite kernel [11]-[13], joint sparsity model [14]-[16], and Markov random fields [17]-[23]. These methods can significantly enhance the classification accuracy in the applications.

MRFs are commonly used by incorporating spatial information into a Bayesian framework and have shown consistent performance in HSI classification. For instance, in [17] Tarabalka *et al.* integrated a probabilistic SVM with MRF and achieved good performance. Other MRF-based methods for HSIs can be found in [21], [22]. However, the over-smoothness problem is a fatal drawback of the traditional MRF-based methods, which has attracted many researchers to continuously work. Xia *et al.* proposed using rotation forests with a specific constraint to learn the posterior probability and then combined this approach with MRF [18]. In [19], an adaptive MRF was proposed where a weighting mechanism was used. In [20], a set of segmentation techniques were incorporated into a MRF-based framework in order to take advantage of the boundary information. These methods incorporate spatial priors through a regularised term which models the relationship between a pixel and its neighborhood with the discrete-value labels. It leads to a difficult and discrete optimization problem. One of the solutions to this problem is using graph-cut techniques [24], which is time-consuming. In [23], the spatial prior was modeled as a MRF on implicit marginal probabilities instead of discrete-value labels. In [27], the semantic representation of image patches was used to model MRF spatial priors instead of discrete labels for image classification tasks. Due to this modification, both methods shown in [23] and [27] allowed one to effectively solve the optimization problem with the smoothness constraint and to

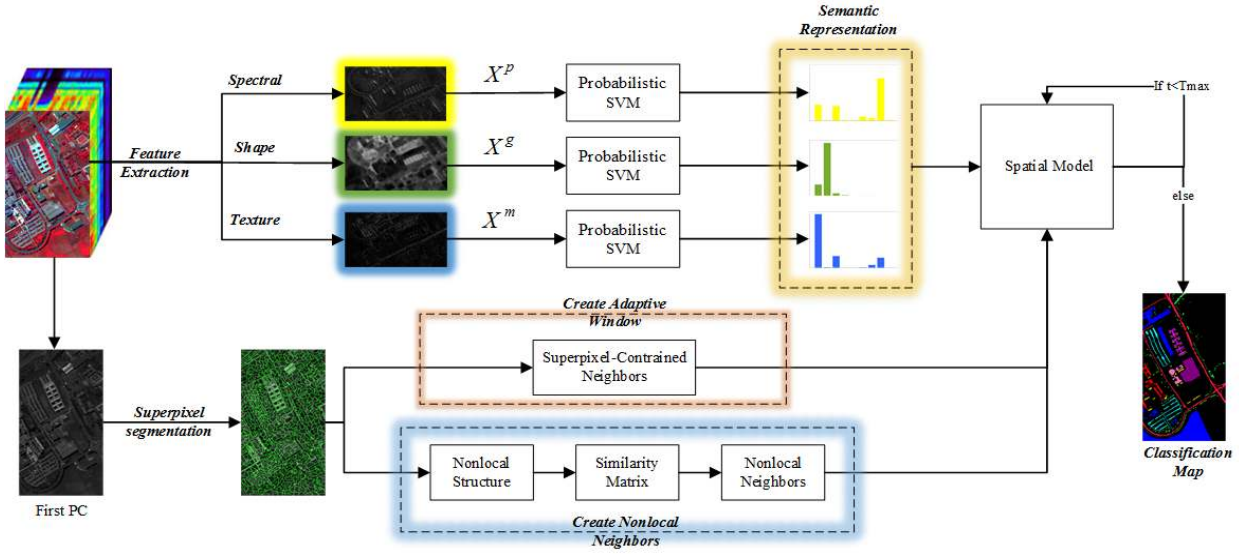


Fig. 1. Flowchart of the proposed framework (NE-MFAS).

achieve better performance. Similar to [27], semantic representation or semantic multinomial representation for computer vision applications [25]-[26] represents the probability of a given patch or image belonging to a specific category. To correctly classify image pixels in a HSI, it is reasonable to consider the semantic representation of a pixel as a probability of whether or not the pixel belongs to a specific class. So posterior probability or implicit marginal probability which reflects the relationship between pixels and labels in the Bayesian framework can be treated as an approximation of the semantic representation. On the other hand, multiple features and non-local spatial information can be exploited to further improve the performance of MRF-based methods.

Since different types of features depict HSIs from different perspectives, multiple feature fusion approaches have been used to enhance the discrimination capability [16], [28]-[30]. However, different types of features usually lie on different feature spaces. Combining these features directly using methods such as vector stack (VS) has witnessed limited performance [29]. In [16] and [30], representation-based methods were presented to combine multiple features. Recently, a MRF-based multiple feature fusion method for HSI classification was proposed [25], which combines multiple features whilst using MRF to incorporate spatial information. Feature fusion and spatial constraints are conducted separately. Unfortunately, the two-part optimization scheme cannot ensure global optimization.

In order to incorporate spatial information and multiple features into a MRF model simultaneously, we here propose a novel MRF model for HSI classification. As we have known, many image pixels belong to the same class with the same pattern but locate at different regions in the image. Therefore, non-local information may be exploited to enhance the discriminability of each pixel in a HSI.

Inspired by the work reported in [23] and [27], we here propose a new approach to combine multiple features with adaptive spatial constraints (MFAS), which employs local adaptive spatial information whilst combining multiple features via an extended MRF model. Firstly, different types of features on different feature spaces are mapped to the same semantic space via a probabilistic SVM classifier, where multiple features lead to various semantic representations. Then various semantic representations and local spatial information are integrated into the MRF model. Specifically, local information is exploited in a superpixel-based method proposed in this paper. Finally, the proposed MRF model is transformed to a global energy minimization process. Removing the modular "Create Nonlocal Neighbors" marked in blue in Fig.1, this framework is a simplified MFAS. Furthermore, in order to extract spatial information from the non-local regions, a non-local regularizer is proposed and added to the MFAS, called the non-local extension of MFAS (NE-MFAS). The non-local information is extracted by searching for similar patches using a K nearest neighbor (KNN) method. In MFAS and NE-MFAS, all the information including local, non-local and multi-feature information are integrated in one single MRF model which is optimized by minimizing the MRF energy function. The proposed framework NE-MFAS is illustrated in Fig.1.

The MRF model proposed in this paper is different from the conventional MRF models presented in [17]-[23] and [27]. The approach shown in [27] incorporates all the semantic representations of image patches via Kullback-Leibler (KL) divergence and then combines the spatial context with a MRF model. However, in our method, we combine multiple features with spatial constraints in an objective function. Furthermore, we measure the similarity between two image patches incorporating non-local information with the basic model MFAS. To our knowledge, this is the first time that multiple features and non-local information are fused in one single MRF model. More importantly, by modeling the MRF spatial prior on the semantic space with the geodesic distance, a manifold distance for measuring the similarity between two probability simplexes [31], our model is finally transformed into a convex and derivable problem which can be solved using a gradient descent method. Comparably, the approaches presented in

[17]-[22] used graph-cut techniques whilst the work shown in [23], modeling a spatial prior with l_1-norm, used the alternating direction method of multipliers (ADMM) to solve the optimization problem.

The rest of this paper is organized as follows. Section II shows the basic spatial context MRF model. In Section III, the proposed basic method MFS combining multiple features with local and spatial constraints in the semantic manifold is presented, and a new approach to construct superpixel-constrained neighborhoods is illustrated and used in our model, namely MFAS. Finally, the non-local extension of MFAS, NE-MFAS, is proposed. In Section IV, the performance of the proposed method is compared with several state of the art techniques on three hyperspectral datasets and the used parameters of the proposed method will be fully evaluated. Finally, Section V gives concluding remarks of this paper.

II. Basic Spatial Context Model

A. Notations

Let $H \equiv \{1,2,...,N\}$ be the indexes of the N pixels of a hyperspectral image. Let $\mathbf{X} \equiv \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ be a HSI dataset including image pixels, where $\boldsymbol{x}_n \in R^d$ represents the spectral bands of the n-th pixel, and d is the number of the spectral bands, and $\mathbf{X}_{train} \equiv \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_M\}$ represent the labeled samples, $Y_{train} \equiv \{y_1, y_2, ..., y_M\} \in K^M$ represent the corresponding class labels of the samples, where M is the number of the labeled pixels, $K = \{1,...,L\}$ is the set of the class labels, and L is the number of the classes.

The goal of the supervised classification of HSIs is: We design a model with $\{\mathbf{X}_{train}, Y_{train}\}$, then predict a label y_n to each pixel $n \in H$ based on $\boldsymbol{x}_n$, given the label set of the unlabeled pixels. Finally, the classification map Y can be obtained.

B. Basic MRF model

The posterior probability $p(y_n \mid \boldsymbol{x}_n)$ of the n-th pixel obtained by a pixel-wise probabilistic classifier often leads to misclassification in homogeneous areas. Considering the relations between neighbors, MRF have been successfully used to exploit the spatial context of a HSI to classify image pixels. In other words, the spatial prior of MRF (if a random field has a Gibbs distribution [39]) can be used here for improving system performance.

In a Bayesian framework, the label $\hat{Y}$ of a HSI can be obtained by maximizing the posterior of Y, given $\mathbf{X}$.

$$\hat{Y} = \underset{Y \in K^N}{\operatorname{argmax}}\, P(Y \mid \mathbf{X}) = \underset{Y \in K^N}{\operatorname{argmax}} \{P(\mathbf{X} \mid Y) P(Y)\} \quad (1)$$

where $P(Y \mid \mathbf{X})$ and $P(Y)$ represent the class-conditional probability distribution and the prior probability of the classes, respectively. In a MRF framework, maximum a posteriori (MAP) decision rules can be used to solve a global energy function minimization problem. Then this function can be simplified based on the assumption of class-conditional independence of the pixels.

$$\hat{Y} = \underset{Y \in K^N}{\operatorname{argmin}} \left\{ \sum_{n=1}^{N} \left(-\log p(y_n \mid \boldsymbol{x}_n) + \log p(y_n) - \log p(\boldsymbol{y}) \right) \right\} \quad (2)$$

where $p(y_n \mid \boldsymbol{x}_n)$ is modeled using a probabilistic SVM [17] or a multinomial logistic regression [26], and $p(y_n)$ is omitted as we assume all the classes have the same contribution. $p(\boldsymbol{y})$ is usually modeled to impose the spatial prior of labels $\boldsymbol{y}$ in the HSI classification. The global energy of Eq.(2) in HSIs can be formulated as:

$$E = \sum_{n=1}^{N} \left\{ -\log \boldsymbol{s}_n - \mu \sum_{(n,m) \in Ne} \delta(y_n - y_m) \right\} \quad (3)$$

where $\boldsymbol{s}_n = p(y_n \mid \boldsymbol{x}_n)$, the first term represents the spectral energy function and the second one characterizes the spatial energy function. $(n,m) \in Ne$ donates that the n-th pixel and m-th pixel are connected in the MRF, μ is the parameter of the smoothness level, and $\delta(\cdot)$ is the Dirac unit impulse function.

The solution of a MRF-based approach is the global minimum of the energy function which is a challenging optimization problem. Sun *et al.* [23] used the implicit marginal probability instead of the label y to model the spatial prior. In [27], a MRF with semantic representation is used to exploit the spatial contexts of image patches. Both the approaches use the posterior probability of the image pixels instead of the labels to model smoothness and derivable energy functions, which can be solved by ADMM and gradient descent, respectively. The

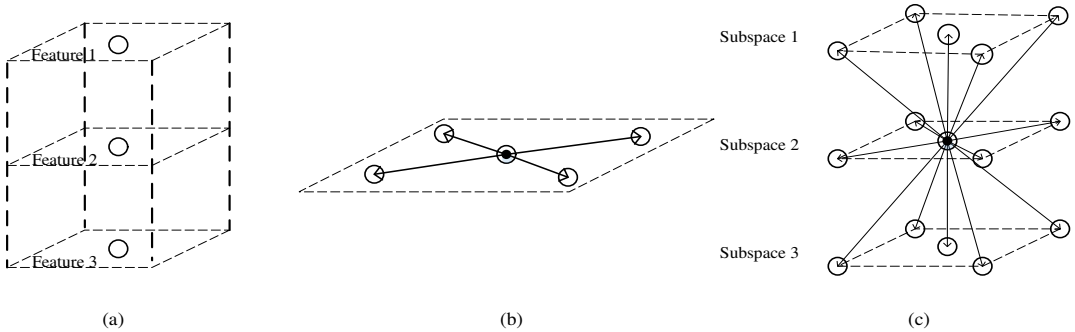


Fig. 2. (a) Multiple features lie in different space, (b) basic spatial context, (c) joint multiple feature spatial context.

basic MRF model [27] with 4-connected pixels is shown in Fig. 3(a), where the original semantic representation of each pixel is the input of the MRF, and the neighborhood is the spatial regular term to constrain the smoothness. The objective function of the basic model can be defined as:

$$E(\boldsymbol{s}_1,...,\boldsymbol{s}_N,\overline{\boldsymbol{s}}_1,...,\overline{\boldsymbol{s}}_N)=\sum_{n=1}^{N}\left\{g(\overline{\boldsymbol{s}}_n,\boldsymbol{s}_n)+\mu\sum_{(n,m)\in Ne}g(\overline{\boldsymbol{s}}_n,\overline{\boldsymbol{s}}_m)\right\} \quad (4)$$

where $\overline{\boldsymbol{s}}_n$ is an unknown denoised semantic representation (DSR) of pixel *n*. The first term is the spectral energy function which measures the similarity between $\overline{\boldsymbol{s}}_n$ and the original semantic representation $\boldsymbol{s}_n$. The second term as the spatial constraint represents the spatial energy function which describes the relationships between $\overline{\boldsymbol{s}}_n$ and the DSRs of the neighbors. The energy can be modeled as the distance between two different semantic representations. The geodesic distance ($g(\bar{\boldsymbol{s}},\boldsymbol{s})=2\arccos\left(\left\langle\sqrt{\bar{\boldsymbol{s}}},\sqrt{\boldsymbol{s}}\right\rangle\right)$) where $\sqrt{\boldsymbol{s}}$ represents element-wise square root) is a suitable choice [31] for measuring the distance between two semantic representations, i.e., probability vectors, and it has been proved to be effective in a semantic manifold [40].

To minimize the global energy, we resort to the iterative conditional modes (ICM) method [41], which iterates over every pixel whilst minimizing the local energy related to one particular variable and keeping others fixed. Then the global optimization problem shown in Eq. (4) is transformed to several sub-problems, and the energy function of *n*-th (*n*=1...*N*) pixel can be written as:

$$E(\overline{\boldsymbol{s}}_n;B_n)=g(\overline{\boldsymbol{s}}_n,\boldsymbol{s}_n)+\alpha\frac{1}{|B_n|}\sum_{m\in B_n}g(\overline{\boldsymbol{s}}_n,\boldsymbol{s}_m) \quad (5)$$

where B_n is the spatial neighborhood of pixel *n*. In Fig. 3(a), B_n contains four neighbors. It is normalized by the size of the neighborhoods $|B_n|$ for convenience, and α is the penalty value for the spatial item.

III. Proposed Method

A. Probabilistic SVM classification

In our application, semantic representation of image pixels in HSIs can be regarded as the probability of whether or not the pixels belong to individual classes. We adopt the popularly used probabilistic SVM [17][21][29] to describe the semantic representation of each pixel.

Given a pixel $\boldsymbol{x}_n$, to obtain the semantic representation of this pixel, we need to map $\boldsymbol{x}_n$ to a middle-level space i.e., the probability space.

For this purpose, we need to know the posterior probability of the pixel belonging to each class *k*. Originally, the SVM classifier does not provide class probability estimates, and it only provides a decision value $d(k\,|\,\boldsymbol{x}_n)$ that indicates the distance between the pixel and the separating hyperplane of class *y*. In [37], $p(k\,|\,\boldsymbol{x}_n)$ is calculated using a sigmoid function.

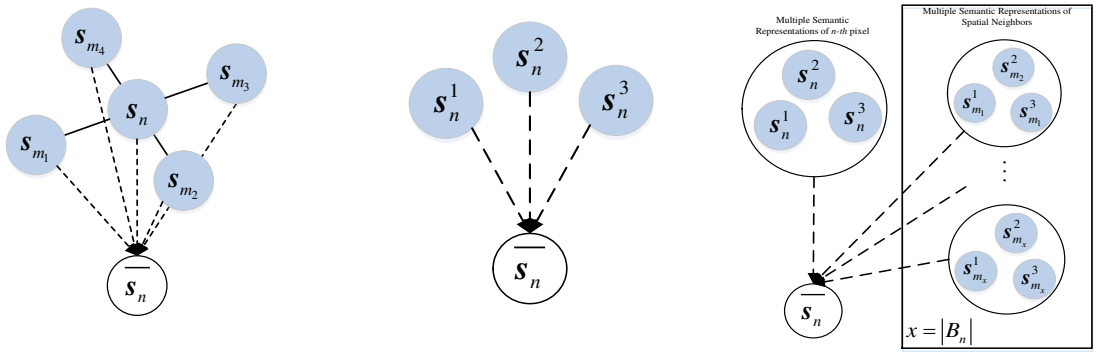
$$p(k\,|\,\boldsymbol{x}_n)=\frac{1}{1+\exp(A_c\cdot d(k\,|\,\boldsymbol{x}_n)+B_c)} \quad (6)$$

where A_c and B_c are estimated by the SVM classifier. Here we use the LIBSVM library [38]. Then we use $\boldsymbol{s}_n=p(k\,|\,\boldsymbol{x}_n)$ as the semantic representation of the *n*-th pixel.

B. Multiple Features of HSI

The combination of multiple features can enhance the discriminability and positively supports the classification task [16], [28]-[30]. In order to extract meaningful information of HSIs, we use low-level feature extraction methods. The first one is the original spectral feature which can preserve the original information of HSIs. The second one is the use of Gabor features [32][33] which describe the texture information of HSIs. In order to exploit the shape information of HSIs, we extract differential morphological profiles (DMP) features [29]. Three features and their results are shown in Table I. We obtain the first three principal components of the HSI using principal component analysis (PCA). Then we extract Gabor and DMP features on the first three PCs based images. For the Gabor feature extraction, we implement Gabor filtering with eight angles and five wavelengths individually on each PC. For the DMP feature extraction, the structure sizes of the morphological opening and closing operation are set to be 2, 4, 6, 8, and 10 respectively. Then we calculate the difference of the morphological processing results between the adjacent structure sizes.

Each type of features for a pixel is represented as a vector,



(a) (b) (c)

Fig. 3. (a) Basic model with 4-connected pixels, (b) multi-feature combination, (c) MFS model.

$\boldsymbol{x}_n^v \in R^{d_v}$ $v=(1,2,...,V)$ are the v-th type of features of the n-th pixel, V and d_v are the number of feature types and the dimension of the v-th type of the features, respectively. The parameters for each type of the features are shown in Table I.

One of the traditional approaches for integrating multiple features is to use vector stacking (VS) which concatenates multiple features directly before they are sent to a classifier. Although VS has been widely used for multi-feature fusion [34], the study reported in [35] shows that the accuracy of a classifier, e.g. SVM with combined features, may decline significantly and the hyperdimensional features may cause the over-fitting problem. In addition, since each type of features and the feature-specific spatial contexts lie in different feature spaces (see Fig.2), combining different types of features is not the best option.

In our study, we firstly map these low-level features to a middle-level space namely semantic space via a probabilistic SVM classifier which has been introduced in the previous section, and then combine these middle-level features in the semantic space. We expect to take advantage of multiple features and overcome the Hughes effect and the over-fitting problem.

TABLE I
INTRODUCE OF BASIC FEATURES

Features	Parameters	Dimension
Spectral feature	All the bands of HSI	the number of bands
Gabor texture feature	Base image : PC1,PC2,PC3 Num. of angles : 8 Num. of wavelengths : 5	3×8×5=120
DMP feature	Base image : PC1,PC2,PC3 Size of structure elements : 2,4,6,8,10	3×2×4=24

C. Multiple-Feature with Spatial Constraint

For the n-th pixel with multiple features $\boldsymbol{x}_n^v$ $v=(1,...,V)$, the multiple semantic representation $\boldsymbol{s}_n^v$ can be obtained using the probabilistic SVM. Different from the low-level features which lie on individual feature spaces, the semantic representations lye on the same middle-level feature space. We can combine these middle-level features through various ways. The work shown in [29] reveals a SVM classifier to fuse probability estimations. It is also shown that minimizing the Kullback-Leibler (KL) divergence led to better classification performance [27]. Here, we use the MRF model shown in Eq. (4) to fuse these middle-level features whilst combining all the local spatial information. Then we can easily transform the objective function as follows:

$$E(\overline{\boldsymbol{s}_n^{(1)}};B_n)=\frac{1}{|V|}\sum_{v=(1,...,V)} g(\overline{\boldsymbol{s}_n^{(1)}},\boldsymbol{s}_n^v) + \alpha\frac{1}{|V|\cdot|B_n|}\sum_{m\in B_n,v=(1,...,V)} g(\overline{\boldsymbol{s}_n^{(1)}},\boldsymbol{s}_m^v) \quad (7)$$

where $\overline{\boldsymbol{s}_n^{(1)}}$ represents the first order DSR of pixel n. Notice that for a training set, if the pixel i is a training sample belonging to class k, the value of $\boldsymbol{s}_i^v$ is a L-dimensional vector with $\boldsymbol{s}_{i,k}^v$ (the kth element of $\boldsymbol{s}_i^v$) being one and the other elements being zeros. If B_n contains the training sample i, $\overline{\boldsymbol{s}_{n,k}^{(1)}}$ (the kth element of $\overline{\boldsymbol{s}_n^{(1)}}$) will be bigger. Similarly, the label information will be integrated together at the same time for better performance. The label of the unlabeled pixel can be determined using the maximum element of $\overline{\boldsymbol{s}_n^{(1)}}$.

Let $\overline{\boldsymbol{S}^{(1)}}\equiv\{\overline{\boldsymbol{s}_1^{(1)}},\overline{\boldsymbol{s}_2^{(1)}},...,\overline{\boldsymbol{s}_N^{(1)}}\}$ be the set of the first order DSRs, and $\boldsymbol{S}\equiv\{\boldsymbol{s}_1^v,\boldsymbol{s}_2^v,...,\boldsymbol{s}_N^v\}$, $v=(1,2,...,V)$ be the set of the original semantic representations. Since the optimization of the solution is carried out by an ICM method, a number of iterations are necessary. It is reasonable for us to use $\overline{\boldsymbol{S}^{(1)}}$ as the input, and go through every pixel. Then we can obtain a reliable and precise second order DSRs $\overline{\boldsymbol{S}^{(2)}}$. Continuously, we can also derive $\overline{\boldsymbol{S}^{(t)}}$ through $\overline{\boldsymbol{S}^{(t-1)}}$, ($t$ = 2, 3, 4,…). Notice that multiple semantic representations' fusion has been performed in the process of deriving $\overline{\boldsymbol{S}^{(1)}}$. So we simplify the objective function as:

$$E(\overline{\boldsymbol{s}_n^{(t)}};B_n)=g(\overline{\boldsymbol{s}_n^{(t)}},\overline{\boldsymbol{s}_n^{(t-1)}}) + \alpha\frac{1}{|B_n|}\sum_{m\in B_n} g(\overline{\boldsymbol{s}_n^{(t)}},\overline{\boldsymbol{s}_m^{(t-1)}}) \quad (8)$$

where $\overline{\boldsymbol{S}^{(t)}}\equiv\{\overline{\boldsymbol{s}_1^{(t)}},\overline{\boldsymbol{s}_2^{(t)}},...,\overline{\boldsymbol{s}_N^{(t)}}\}$ is the t-th order DSRs. The above process is iterated t_{max} times to obtain the final results. The final DSRs are expected to be more robust with less noise. The above method namely multiple-features with spatial constraints (MFS) is shown in Fig.3(c).

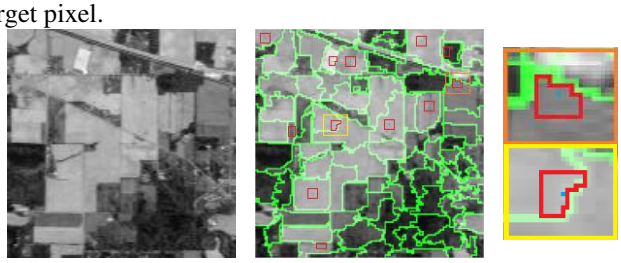
Particularly, α is the penalty factor to control the importance of the spatial neighbors and $|B_n|$ stands for the size of the neighborhood shown in Eqs. (7) and (8). We define $\alpha=|B_n|$ for simplification, indicating that the weight of each neighbor is the same as that of the center pixel in this MRF model.

D. Superpixel-Constrained Neighborhood

All the MRF-based and spatial-spectral approaches are based on the assumption that image pixels from the same local region belong to the same class. But the investigated window usually contains some pixels which belong to other classes. Researchers find various ways to overcome this issue, e.g. using weight coefficients [19][30][46], superpixel segmentation [16][42], and adaptive neighborhood construction such as anisotropic local polynomial approximation intersection of confidence intervals (LPA-ICI) [43].

In this paper, a superpixel segmentation method based on entropy rates [44] is used to generate a 2-D superpixel map. Specifically, this is a graph-based clustering algorithm which can generate compact, homogeneous and balanced superpixels. Also it only has one parameter to be tuned, which controls the

number of the superpixels in the base image. Here, the first principal component (PC) of HSIs by PCA is used to generate the base image for the superpixel segmentation (shown in Fig. 4(a)). Having superpixels, some superpixel-based methods [16][42] treat each superpixel as a unit based on the result of the image segmentation. Different from these methods, our superpixel-based approach is developed using the segmentation result to constrain the neighborhood that helps us to include the neighboring pixels which belong to the same class as that of the target pixel. It also enhances the identification of the target pixel. As shown in Fig. 4(b), the blue points are the target pixels, and the red window represents the neighboring window of each target pixel.



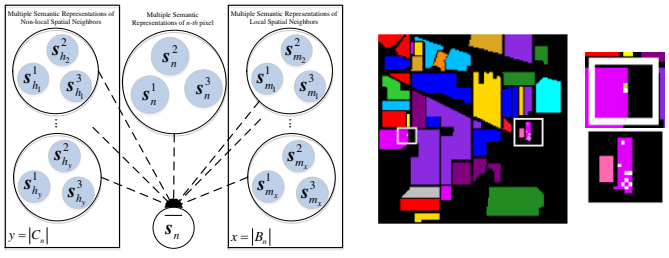
(a) (b)

Fig. 4. (a) The first PC of Indian Pines image. (b) The segment map with some instance of superpixel-based neighborhood in the Indian Pines image.

Finally, a large window with detailed information can be extracted with less mis-classified pixels, whilst the boundary information can be well preserved. Taking superpixel-based neighborhoods, we can transform our basic model (MFS) to a new model namely multiple feature with adaptive spatial constraints (MFAS).

E. Non-local Extension of MFAS

Non-local information has demonstrated its importance in HSI analysis [46]-[49]. The motivation for extending our model to the non-local one is because of the high degree of redundancy in each HSI. Traditional spatial-spectral methods hold the assumption that image pixels in a local region belong to the same class, but we also observe that pixels belonging to the same class can be found in different regions. Non-local techniques can be used to explore this similarity. To utilize the non-local spatial information, we measure the similarity between the patch centered with the concerned pixel and the other non-local patch. Adding the non-local similarity information, we propose a non-local extension method NE-MFAS (see Fig. 5(a)).



(a) (b)

Fig. 5. (a) NE-MFAS model.(b) Illustration of non-local neighbors.

Spatially neighboring pixels tend to belong to the same class and a pixel can be represented by its local neighbors. Therefore,

Algorithm 1: NE-MFAS for HSI classification

Input: 1) HSI data set with labels of training samples

2) The number of superpixels L

3) The number of non-local neighbors C, local and non-local structure window size W_{local} and $W_{non\text{-}local}$

Step 1: Extract multiple features of each pixel from the HSI

Step 2: Map multiple features to the same semantic space through probabilistic SVM

Step 3: For each test sample:

Step 3.1: Construct the superpixel-constrained neighborhood

Step 3.2: Calculate C-nearest non-local neighbors and corresponding weight by Eqs. (9) and (10)

Step 3.3: Obtain the first order DSR by Eq. (11)

Step 3.4: Initialize the highest order of DSR $t_{\max}$, and $t \leftarrow 2$

While $t \neq t_{\max}$, do

(1) Obtain $\overline{s_n^{(t)}}$ through $\overline{S^{(t-1)}}$ by Eq. (12)

(2) $t \leftarrow t + 1$

End while

Step 3.5: Let y_n equal to the index of maximum element of $\overline{s_n^{(t_{\max})}}$

Output: Predict labels of testing samples Y_{test}

the similarity of two pixels can be approximated by the similarity of two image patches. Our superpixel-based neighborhood method can be used to construct a patch structure.

Given an image patch, non-local means (NLM) algorithms [45]-[48] expect to find the patches being spatially far from this given patch but having structures similar to that of the given patch. The non-local constraint can be applied to this classification, and we can extract the spatial information from both local and non-local neighbors.

To measure the similarity between two patches, we use a variation of the standard KNN method. Firstly, the average pooling strategy is applied to each patch exploiting the most significant information of the patch. Then, the similarity of the two patches is measured using geodesic distance (GD) which can capture the manifold structure of the HSI. We define the similarity of pixels n and n' using GD as:

$$SGD(n,n') = g\left(\frac{1}{|U_n|}\sum_{m\in U_n} \boldsymbol{x}_m^p, \frac{1}{|U_{n'}|}\sum_{m'\in U_{n'}} \boldsymbol{x}_{m'}^p\right) \quad (9)$$

where $\boldsymbol{x}_m^p$ stands for the spectral feature of pixel m and U_n represents the superpixel-based neighbors of pixel n. Finally, C-nearest pixels are selected to form a group which is called the non-local neighbors of pixel n. Fig. 5(b) illustrates an example

of the non-local neighbors of pixel n on the Indian Pines data set. The green point in the white rectangle is the target pixel and the other white points are the neighbors. From this map, we observe that these image pixels all belong to the same class but lie in different regions.

In order to reflect the difference between the neighbors with different similarities, we define the weight coefficient of the non-local neighbors as:

$$\omega_{(n,n')} = \exp\left(\frac{-\left|SGD(n,n')\right|}{\gamma}\right) \tag{10}$$

where n' denotes one of the non-local neighbors of pixel n and γ is the scale parameter. By incorporating the non-local spatial information, the objective function shown on Eq. (7) can be extended to:

$$E(\overline{\boldsymbol{s}_n^{(1)}};(B_n,C_n)) = \frac{1}{|V|}\left(\sum_{v=(1,\ldots,V)} g(\overline{\boldsymbol{s}_n^{(1)}},\boldsymbol{s}_n^v) + \sum_{m\in B_n}\sum_{v=(1,\ldots,V)} g(\overline{\boldsymbol{s}_n^{(1)}},\boldsymbol{s}_m^v) + \sum_{h\in C_n}\sum_{v=(1,\ldots,V)} g(\overline{\boldsymbol{s}_n^{(1)}},\omega_{(n,h)}\cdot\boldsymbol{s}_h^v)\right) \tag{11}$$

where C_n is the non-local neighborhood of pixel n, the second term represents the local spatial energy and the third term stands for the non-local spatial energy. Then, the corresponding t-th order objective function becomes:

$$E(\overline{\boldsymbol{s}_n^{(t)}};(B_n,C_n)) = g(\overline{\boldsymbol{s}_n^{(t)}},\overline{\boldsymbol{s}_n^{(t-1)}}) + \sum_{m\in B_n} g(\overline{\boldsymbol{s}_n^{(t)}},\overline{\boldsymbol{s}_m^{(t-1)}}) + \sum_{h\in C_n} g(\overline{\boldsymbol{s}_n^{(t)}},\omega_{(n,h)}\cdot\overline{\boldsymbol{s}_h^{(t-1)}}) \tag{12}$$

We can use gradient descent to minimize Eqs. (12) and (13) for each pixel n. The gradient of pixel n can be derived as (k is the index of the element):

$$\frac{\partial E(\overline{\boldsymbol{s}_n};(B_n,C_n))}{\partial \overline{\boldsymbol{s}_{n,k}}} = \frac{1}{|V|}\sum_{v=(1,\ldots,V)} f(\overline{\boldsymbol{s}_{n,k}},\boldsymbol{s}_{n,k}^v) + \frac{1}{|V|}\sum_{m\in B_n}\sum_{v=(1,\ldots,V)} f(\overline{\boldsymbol{s}_{n,k}},\boldsymbol{s}_{m,k}^v) + \frac{1}{|V|}\sum_{h\in C_n}\sum_{v=(1,\ldots,V)} f(\overline{\boldsymbol{s}_{n,k}},w_{(n,h)}\cdot\boldsymbol{s}_{h,k}^v) \tag{13}$$

where

$$f(x,y) = \frac{\partial g(x,y)}{\partial x} = -\frac{\sqrt{y}}{2\sqrt{x}\sqrt{1-(\sqrt{x}\sqrt{y})^2}} \tag{14}$$

The complete process of the proposed NE-MFAS algorithm for the HSI classification is summarized in Algorithm 1.

IV. Experimental Results

In this section, we show the effectiveness and efficiency of the proposed NE-MFAS on three recognized hyperspectral datasets. The classification results are compared with those of several state-of-the-art methods.

A. Data sets

The first hyperspectral dataset was acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines region in Northwestern Indiana on June 12, 1992, each of which consists of 145×145 pixels and 220 spectral reflectance bands in the wavelength ranging from 0.4

TABLE II
TRAINING AND TEST SETS FOR IN AVIRIS INDIAN PINES

No	Class	Train	Test
1	Alfalfa	3	43
2	Corn-notill	72	1356
3	Corn-min	42	788
4	Corn	12	225
5	Grass/Pasture	25	458
6	Grass/Trees	37	693
7	Grass/Pasture-mowed	2	26
8	Hay-windrowed	24	454
9	Oats	1	19
10	Soybeans-notill	49	923
11	Soybeans-min	123	2332
12	Soybeans-clean	30	563
13	Wheat	11	194
14	Woods	64	1201
15	Bldg-grass-trees-drives	20	366
16	Stone-steel towers	5	88
Total		520	9729

TABLE III
TRAINING AND TEST SETS FOR IN ROSIS PAVIA UNIVERSITY

No	Class	Train	Test
1	Asphalt	332	6299
2	Meadows	933	17716
3	Gravel	105	1994
4	Trees	154	2910
5	Metal sheets	68	1277
6	Bare Soil	252	4777
7	Bitumen	67	1263
8	Bricks	185	3497
9	Shadows	48	899
Total		2144	40632

TABLE IV
TRAINING AND TEST SETS FOR IN AVIRIS SALINAS

No	Class	Train	Test
1	Broccoli green weeds 1	21	1988
2	Broccoli green weeds 2	38	3688
3	Fallow	20	1956
4	Fallow rough plow	14	1380
5	Fallow smooth	27	2651
6	Stubble	40	3919
7	Celery	36	3543
8	Grapes untrained	113	1115
9	Soil vineyard develop	63	6140
10	Corn senesced weeds	33	3245
11	Lettuce romaine 4 weeks	11	1057
12	Lettuce romaine 5 weeks	20	1907
13	Lettuce romaine 6 weeks	10	906
14	Lettuce romaine 7 weeks	11	1059
15	Vineyard untrained	73	7195
16	Vineyard vertical trellis	19	1788
Total		549	53580

to 2.5 μm. We remove 20 water absorption bands, leaving 200 radiance channels to be used. The spatial resolution of this image is about 20m per pixel. This image contains 16 ground-truth classes as shown in Table II. The false color map and the corresponding reference map are shown in Fig. 6(a) and (b).

The second one is an urban image acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) over Pavia University, northern Italy. It includes 610×340 pixels and 115 spectral reflectance bands ranging from 0.43 to 0.86 μm and has a spatial resolution of 1.3 m per pixel. We select 103 of the bands with 12 noisy bands removed. This image contains 9 classes of ground-truth data as shown in Table III. The false color map and the corresponding reference map are shown in Fig. 7(a) and (b).

The final data set was acquired over Salinas Valley, California, in 1998, by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor. The original data set is composed of 224 bands with a spectral range from 0.4 to 2.5 μm. The image has a size of 512×217 pixels with a spatial resolution of 3.7m. We also remove 20 water absorption bands with 204 bands left. This image contains 16 ground-truth classes which are described in Table IV. The false color map and the corresponding reference map are shown in Fig. 8(a) and (b).

B. Experimental Setting

To investigate the effectiveness of the proposed method for hyperspectral classification, our method NE-MFAS is compared against the basic models of our framework (SVM with each type of features, MFS, MFAS) to investigate the effectiveness of our method. Then we compare our method with the other state of the art techniques, including SRC [14], SVM-CK [11], SVM-MRF [17], and SOMP [14].

In our method, the SVM classifier uses the Gaussian kernel and the parameters are obtained by cross-validation. The parameter setting for the other SVM classifiers is also selected via cross-validation. The parameters of SRC and SOMP are the same as that shown in [14]. The experiments are conducted ten times and the average values of the experimental results are recorded in order to avoid the bias induced by random sampling.

To measure the classification performance, overall accuracy (OA), average accuracy (AA), accuracy for each class and kappa coefficient are calculated. All the results are averaged over ten times' run to reduce possible biases induced by random sampling.

C. AVIRIS Indian Pines data set

In this experiment, about 5% samples are randomly selected from each class for training (total 520 samples) and the rest samples (total 9729 samples) for testing (see Table II). Table V shows the classification results of SVM with different types of features, MFS, MFAS, NE-MFAS and the other similar methods, in which the best results are marked in bold.

For MFS, MFAS and NE-MFAS, the local window size is set to be 5×5 (W_{local} = 5) and the number of the iteration is set to 3. The non-local window size and the number of non-local

TABLE V
CLASSIFICATION ACCURACY (%) FOR THE INDIAN PINES IMAGE ON THE TEST SET

Class	SVM (Spectral)	SVM (Gabor)	SVM (DMP)	SRC (Spectral)	SVM-MRF	SVM-CK	SOMP	MFS	MFAS	NE-MFAS
1	8.37	42.56	0	49.62	0	63.64	55.00	96.50	**100**	**100**
2	71.11	68.60	69.11	59.10	81.32	90.48	85.45	95.64	95.27	**96.65**
3	58.65	72.51	75.62	55.66	67.80	91.86	75.07	95.56	**97.86**	97.73
4	21.16	58.36	62.18	36.32	40.84	83.81	81.17	**90.71**	90.04	88.44
5	83.12	83.91	84.83	85.79	89.43	88.41	88.92	92.93	94.69	**95.80**
6	92.74	89.67	93.46	93.61	99.27	97.75	98.54	99.15	99.96	**99.98**
7	0	0	0	64.80	0	73.33	7.20	95.00	91.54	**99.52**
8	98.90	96.59	97.78	98.24	99.63	98.90	99.74	**100**	**100**	**100**
9	0	0	0	42.11	0	54.73	6.40	54.74	**75.79**	75.68
10	59.92	75.74	80.48	66.28	72.60	86.43	70.17	94.24	92.34	**96.90**
11	82.41	83.48	87.23	74.35	95.39	92.29	93.50	98.92	99.10	**99.56**
12	48.31	62.61	52.42	47.50	74.36	83.94	70.89	94.40	95.60	**97.51**
13	91.91	94.43	95.83	97.92	98.82	98.97	98.71	96.08	99.48	**99.48**
14	95.77	95.15	96.19	94.93	97.99	96.82	98.52	99.80	**99.93**	99.59
15	45.00	81.50	71.58	34.13	60.35	89.21	90.55	96.23	96.15	**99.49**
16	79.89	86.93	47.50	87.03	90.67	91.69	91.43	93.30	98.98	**99.01**
OA	74.63	80.12	80.89	72.07	84.91	91.59	87.55	96.87	97.24	**98.20**
AA	58.58	68.25	64.66	67.96	66.80	86.39	75.30	93.30	95.42	**96.46**
Kappa	0.7073	0.7722	0.7672	0.6805	0.8257	0.9042	0.8572	0.9642	0.9685	**0.9794**

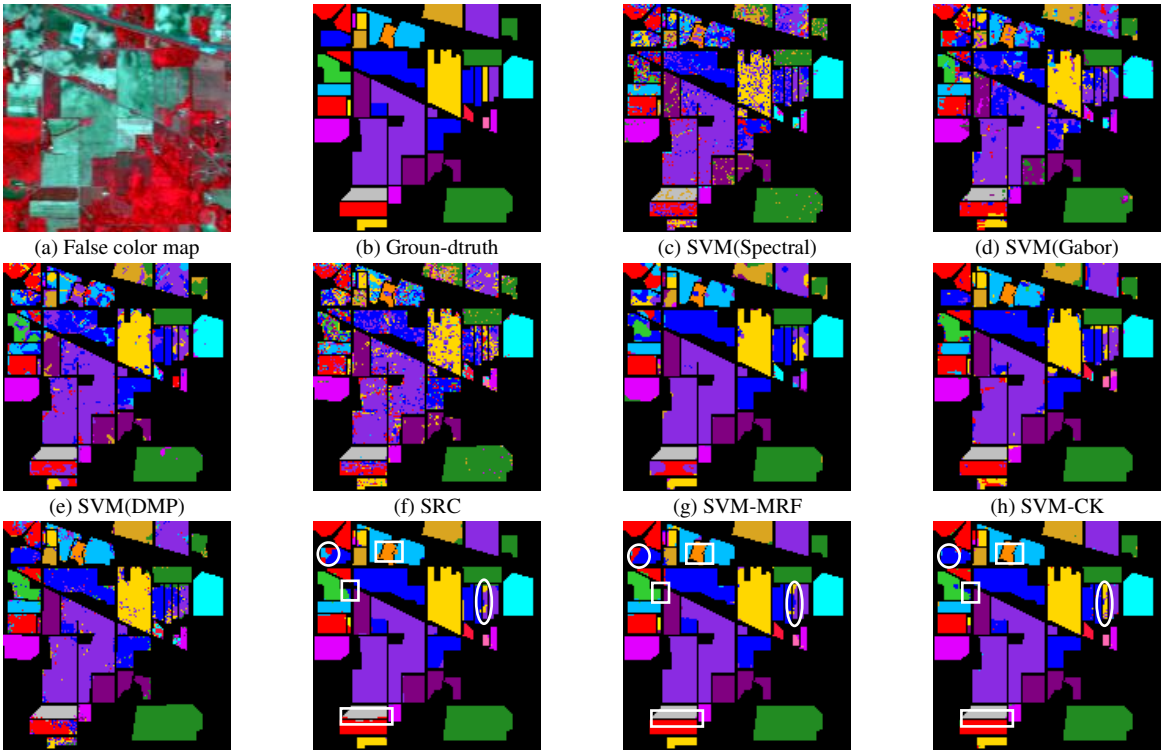


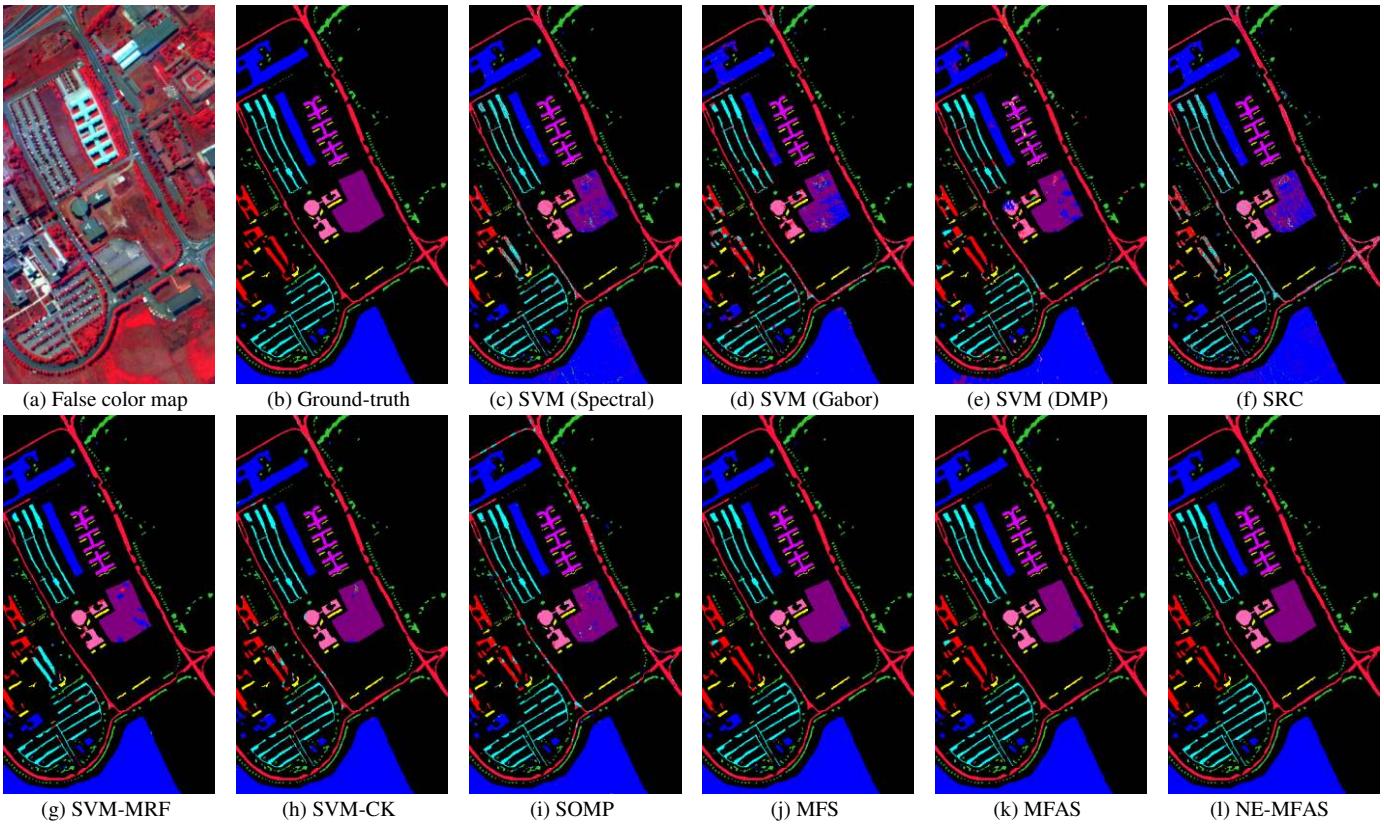
Fig. 6. Indian Pines image. (a) false color map; (b) ground-truth; (c) SVM with original spectral feature(OA=72.64%); (d) SVM with Gabor texture feature (OA=78.79%) (e) SVM with DMP feature (OA=79.62%) (f) SRC (OA=73.11%); (g) SVM-MRF (OA=88.27%); (h) SVM-CK (OA=92.26%); (i) SOMP (OA=89.38%); (j) MFS (OA=97.03%); (k) MFAS (OA=97.64%); (l) NE-MFAS (OA=98.52%).

neighbors are set to be 21×21 ($W_{non\text{-}local}$ = 21) and 40 (C=40) for NE-MFAS. For MFAS and NE-MFAS, the number of superpixels is set to 100.

From Table V, we clearly see that the methods with spatial information (MRF prior, CK and SOMP) have better performance compared to the pixel-based classifiers (SVM and SRC). It is also shown that SVMs with different types of features result in different performance. The DMP-based SVM leads to better performance than the others. However, it is a complicated scenario. For instance, SVM with spectral features has the best performance on classes 2 and 8; the Gabor texture feature leads to the best performance on classes 1, 12, 15 and 16; the DMP feature results in the best results on the rest of the classes. Our methods, MFS, MFAS and NE-MFAS, which exploit local spatial information whilst combining multiple features, give better performance than the other methods. Especially, NE-MFAS produces the highest overall accuracy and the best class classification accuracy for most of the classes (twelve of the total sixteen classes). Meanwhile, by incorporating the adaptive window constraint, MFAS achieves about 0.5% better than MFS in OA. Moreover, combining non-local spatial information, NE-MFAS makes the system performance 1% better than MFAS. It is obvious that the other methods fail to identify the samples of the Alfalfa class. The Alfalfa class has only 46 samples in total with merely 3 for training which only covers a small region. For the pixel-based classifiers (SVM and SRC), only 3 training samples cannot be used to learn an effective model for the rest samples. For the SVM-MRF, without any constraint, the over-smoothing effect causes the Alfalfa region to be erode by the regions around it (Corn-notill and Soybeans-notill). For SOMP and SVM-CK, with the fixed window (7×7), most the neighbors of each Alfalfa pixel are belonging to two adjacent classes (Corn-notill and Soybeans-notill). Assuming each Alfalfa pixel has a pattern similar to that of its neighbors, the Alfalfa pixels could be classified as Corn-notill and Soybeans-notill. Better classification accuracy is achieved with the MFS using a relatively small size (5×5) and incorporating multiple features. MFAS and NE-MFAS achieve correct identification of all the Alfalfa pixels by exploiting adaptive local spatial information and non-local spatial information. The same conclusion can be made in Oats and Grass/Pasture-mowed classes, reflecting that the proposed method not only achieves good performance in large homogenous regions but also effectively identifies small objects.

TABLE VI
CLASSIFICATION ACCURACY (%) FOR THE UNIVERSITY OF PAVIA IMAGE ON THE TEST SET

Class	SVM (Spectral)	SVM (Gabor)	SVM (DMP)	SRC (Spectral)	SVM-MRF	SVM-CK	SOMP	MFS	MFAS	NE-MFAS
1	90.49	90.25	90.98	74.32	98.43	98.13	86.83	99.41	99.46	**99.85**
2	96.53	97.04	96.09	94.52	99.83	99.60	99.50	**100**	99.99	**100**
3	70.34	68.58	81.97	59.54	71.45	88.00	90.77	96.16	96.52	**99.91**
4	92.11	92.90	79.37	81.68	90.14	97.82	89.34	99.02	97.71	**99.04**
5	99.18	97.34	91.10	99.62	99.78	99.69	99.98	**100**	99.71	99.89
6	76.03	65.13	85.16	54.59	84.58	98.19	92.15	96.10	96.27	**99.92**
7	76.28	69.83	66.14	76.56	79.81	96.06	94.59	94.26	96.98	**99.62**
8	85.28	84.99	90.50	75.66	95.99	93.52	95.48	99.48	99.47	**99.76**
9	98.93	99.71	89.43	89.84	98.44	98.18	90.20	**99.81**	99.15	99.47
OA	90.12	88.29	90.41	81.93	94.75	97.85	94.82	99.04	99.09	**99.85**
AA	87.24	84.98	85.64	78.48	90.94	96.58	93.20	98.34	98.51	**99.72**
Kappa	0.8680	0.8431	0.8724	0.7568	0.9295	0.9714	0.9312	0.9873	0.9879	**0.9980**



(a) False color map (b) Ground-truth (c) SVM (Spectral) (d) SVM (Gabor) (e) SVM (DMP) (f) SRC

(g) SVM-MRF (h) SVM-CK (i) SOMP (j) MFS (k) MFAS (l) NE-MFAS

Fig. 7. The University of Pavia image. (a) false color map; (b) ground-truth; (c) SVM with original spectral feature(OA=90.32%); (d) SVM with Gabor texture feature (OA=87.40%) (e) SVM with DMP feature (OA=90.25%) (f) SRC (OA=82.23%); (g) SVM-MRF (OA=95.42%); (h) SVM-CK (OA=97.72%); (i) SOMP (OA=95.13%); (j) MFS (OA=99.20%); (k) MFAS (OA=99.35%); (l) NE-MFAS (OA=99.77%).

Fig. 6(b) shows the ground-truth. The classification maps of the different methods are shown in Fig. 6(c)-(l) with the fixed training samples. In Fig. 6, the methods with spatial information have more accurate results than the others. Nevertheless, it also can be seen that SVM-MRF, SVM-CK, SOMP yield misclassification at the boundaries or small objects, as these methods do not consider boundary information and only use one type of features. MFS gives better uniformity and boundary location results than the traditional spectral-spatial methods. In addition, from the three white rectangles shown in Fig. 6(j) and (k), we observe that MFAS leads to clearer boundaries than MFS. This is because the fact that an adaptive window helps to preserve boundary information. From the two white circles shown in Fig. 6(j),(k)

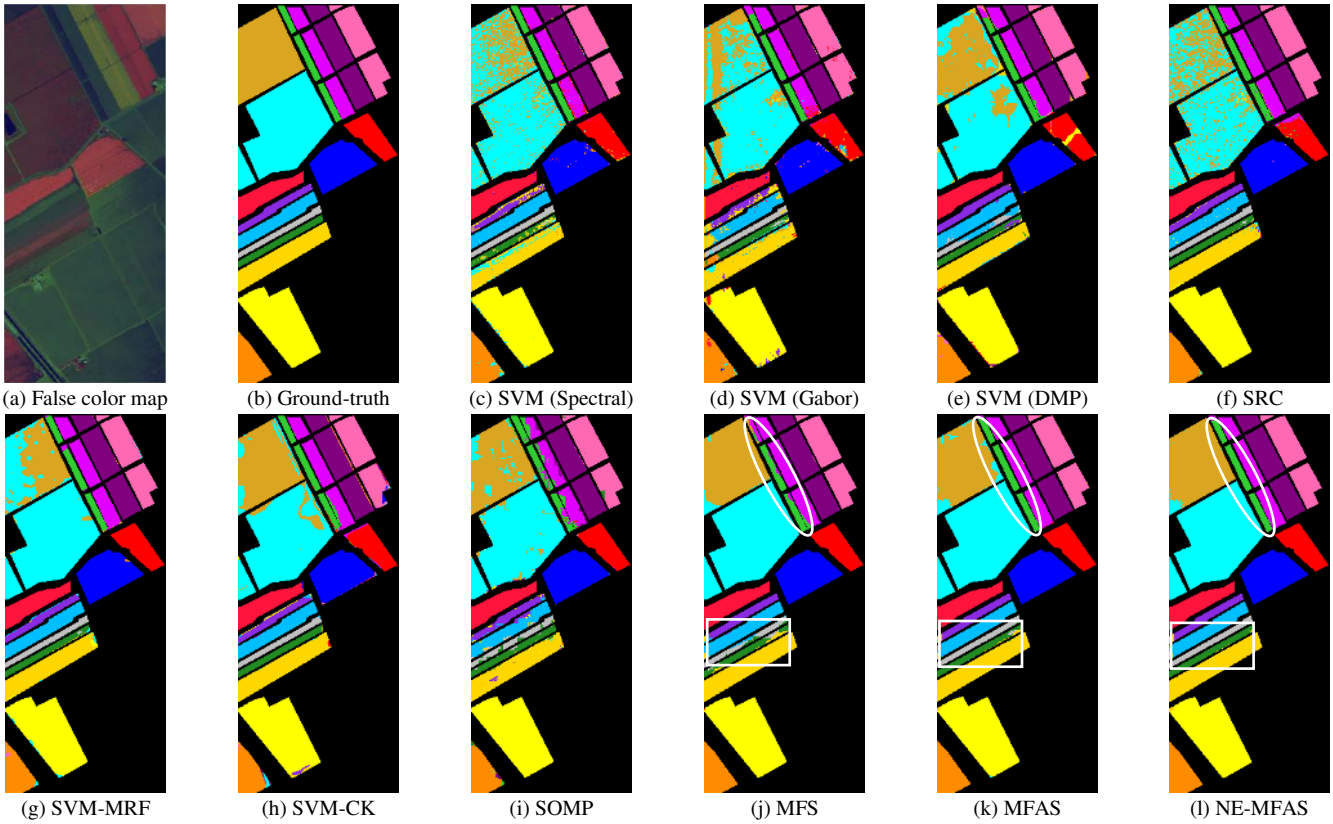
and (l), we notice that the classification map of NE-MFAS not only has clearer boundaries but also has better spatial consistency in the image, which shows the contribution of the non-local information. Overall, the proposed method NE-MFAS achieves better classification performance than the others.

D. ROSIS University of Pavia data set

In this section, we evaluate the proposed methods on the ROSIS University of Pavia dataset while comparing it with the aforementioned methods. Around 5% samples from each class

TABLE VII
CLASSIFICATION ACCURACY (%) FOR THE SALINAS IMAGE ON THE TEST SET

Class	SVM (Spectral)	SVM (Gabor)	SVM (DMP)	SRC (Spectral)	SVM-MRF	SVM-CK	SOMP	MFS	MFAS	NE-MFAS
1	98.20	97.03	95.03	99.92	99.98	97.91	**100**	**100**	**100**	**100**
2	98.56	98.61	97.14	99.71	99.47	98.01	99.87	**100**	99.93	**100**
3	87.57	81.41	85.46	92.72	98.90	96.19	96.90	**99.96**	98.13	97.83
4	98.20	95.02	93.90	**99.13**	96.05	97.14	89.42	83.88	99.00	98.73
5	96.11	96.36	92.71	93.74	**99.53**	95.66	82.29	96.95	99.25	99.47
6	99.48	99.70	98.46	99.97	99.82	97.70	99.81	99.88	99.99	**100**
7	99.23	96.44	95.47	**99.93**	99.56	97.15	98.89	99.17	98.92	98.74
8	87.25	86.91	88.85	78.46	97.78	93.11	95.37	**99.66**	99.11	99.26
9	98.84	96.73	98.45	99.53	98.39	97.69	99.86	**99.90**	99.57	99.70
10	86.24	80.55	94.16	93.85	98.38	98.87	95.43	**99.80**	99.12	99.56
11	81.44	57.11	91.75	97.78	95.18	87.24	91.11	91.84	95.61	**98.58**
12	99.09	81.33	90.81	98.17	**99.97**	97.35	88.97	96.64	97.15	97.17
13	96.35	87.95	63.58	97.07	97.08	**99.12**	74.86	83.31	95.85	94.92
14	89.14	81.11	90.39	95.56	95.28	**96.56**	93.44	80.96	91.03	95.37
15	61.55	53.65	79.39	56.58	59.74	89.79	80.23	98.65	95.86	**99.39**
16	95.88	91.79	96.21	98.57	97.42	96.14	96.62	**100**	99.96	99.97
OA	89.46	85.68	91.12	88.35	92.98	95.31	93.49	98.16	98.53	**99.33**
AA	92.07	86.36	90.73	93.81	95.47	95.98	92.69	95.66	98.10	**98.83**
Kappa	0.8822	0.8398	0.9010	0.8702	0.9214	0.9477	0.9274	0.9794	0.9836	**0.9926**



(a) False color map (b) Ground-truth (c) SVM (Spectral) (d) SVM (Gabor) (e) SVM (DMP) (f) SRC
(g) SVM-MRF (h) SVM-CK (i) SOMP (j) MFS (k) MFAS (l) NE-MFAS

Fig. 8. The Salinas image. (a) false color map; (b) ground-truth; (c) SVM with original spectral feature(OA=87.73%); (d) SVM with Gabor texture feature (OA=85.78%) (e) SVM with DMP feature (OA=91.57%) (f) SRC (OA=88.86%); (g) SVM-MRF (OA=92.95%); (h) SVM-CK (OA=94.75%); (i) SOMP (OA=94.18%); (j) MFS (OA=97.42%); (k) MFAS (OA=98.06%); (l) NE-MFAS (OA=99.01%).

are randomly chosen for training and the remaining samples for testing (see Table III). Table VI shows the classification results of each method, and the best results are marked in bold. For MFS, MFAS and NE-MFAS, the size of the local window is set to be 7×7 (W_{local} = 7) and the number of the iterations is set to 2. The number of superpixels is set to 800 for the superpixel-based neighborhood. For NE-MFAS, the non-local window size and the number of non-local neighbors are set to be 25×25 ($W_{non\text{-}local}$ = 25) and 60 (C = 60) respectively.

From Table VI, we can draw the same conclusion as that shown in the last section: Using spatial information is beneficial. Different from the Indian Pines dataset, University of Pavia dataset is an urban area without much homogeneity, so it is hard to find an adaptive window for each pixel. MFAS only achieves about 0.05% better than MFS in OA, and about 0.2% higher in AA. NE-MFAS is about 0.8% higher than MFS, and NE-MFAS achieves the best accuracy among all the methods in OA, AA and Kappa, because of the impact of the non-local spatial constraint.

Fig.7(c)-(l) illustrates the classification maps of the different methods. The spectral-spatial methods (e.g., SVM-MRF, SOMP) result in much better classification maps than the pixel-based methods (e.g., SVM and SRC). Although all the methods with spatial information obtain satisfactory results, our methods (MFS, MFAS and NE-MFAS) achieve better consistency in large regions and better identification in the small regions. From the white rectangle and the two white circles shown in Fig. 7(j)-(k) we witness that MFAS and NE-MFAS outline boundaries better than MFS. NE-MFAS shows good performance in detecting small regions such as trees in the white circle (the top one), which have less local spatial information but more non-local information. The non-local regularizer helps NE-MFAS fuse spatial information from those non-local regions, and improves the accuracy of classifying these tree pixels. Indeed, the proposed method NE-MFAS has achieved promising performance in terms of region uniformity and boundary outlining.

E. AVIRIS Salinas data set

For this data set, we randomly select 1% samples from each class to form the training set and the rest of the samples for testing (see Table IV). The classification results of each method have been shown in Table VII, and the best results are marked in bold.

Due to the large homogeneity of this image, the size of local and non-local windows is larger than those used for the Indian Pines and Pavia University datasets. The local window size is set to be 15×15 (W_{local} = 15) for MFS, MFAS, NE-MFAS. For NE-MFAS, the size of non-local windows and the number of non-local neighbors is set to be 33×33 ($W_{non\text{-}local}$ = 33) and 100 (C = 100) respectively. The number of superpixels is set to be 500 in experiments.

For this dataset, all the methods have good classification results, but our methods perform better. MFAS is only about 0.3% higher than MFS in OA, but 2.5% higher than that in AA. Even though MFS obtains the best results in some classes, the AA of MFS is worse than that of MFAS due to low accuracy in classes 13 and 14. This is due to the fact that MFAS can identify the boundary between Lettuce romaine 6 weeks and Lettuce romaine 7 weeks' samples. NE-MFAS achieves the highest accuracy. It is about 0.8% and 0.7% higher than MFAS in OA and AA respectively. These results further confirm the effectiveness of our MRF model and the significance of the adopted adaptive window and the non-local regularizer.

Fig. 8(b) shows the ground-truth of this dataset, and the classification maps of the different methods are shown in Fig. 8(c)-(l). Fig. 8 illustrates that our methods result in more accurate classification maps than the others. The white rectangle and the white circle regions shown in Fig. 8(j)-(l) illustrate better uniformity of using NE-MFAS than MFS and MFAS. It is evident that incorporating adaptive windows can make boundary outlining more accurate, but fixed windows may lead to better performance in large homogenous regions. The integration of non-local information in NE-MFAS leads to better performance in most of the regions.

F. Parameter Analysis

In this section, we discuss the effects of changing some parameters involved in our model on the system performance using two datasets, the Indian Pines and the University of Pavia. We still use 5% samples per class for training and the remaining for testing. Meanwhile, the experiments are conducted for five

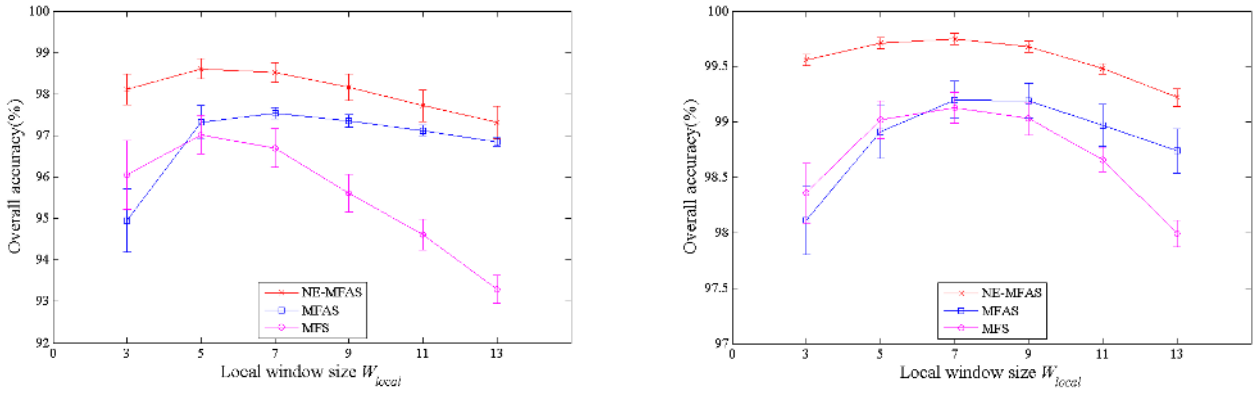
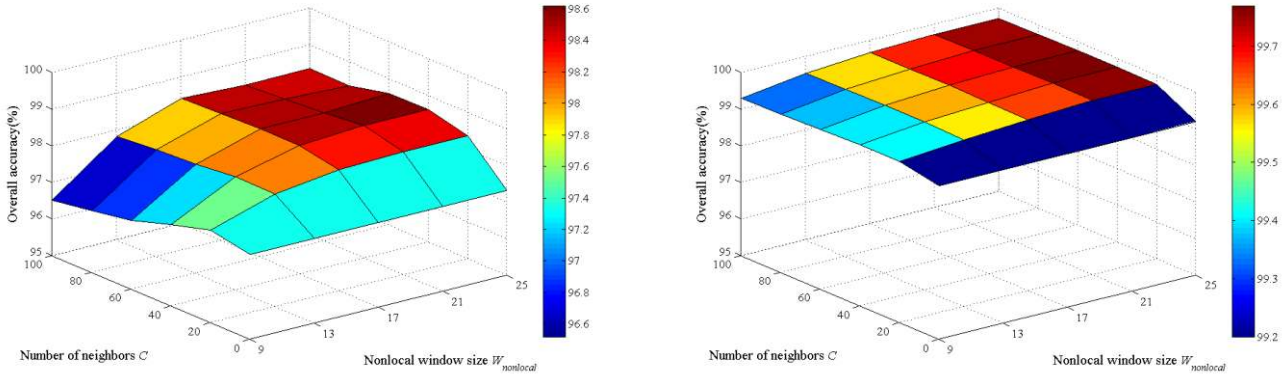


Fig. 9. Impact of the local window size W_{local} for (a) the Indian Pines and (b) the Pavia University.



(a) (b)

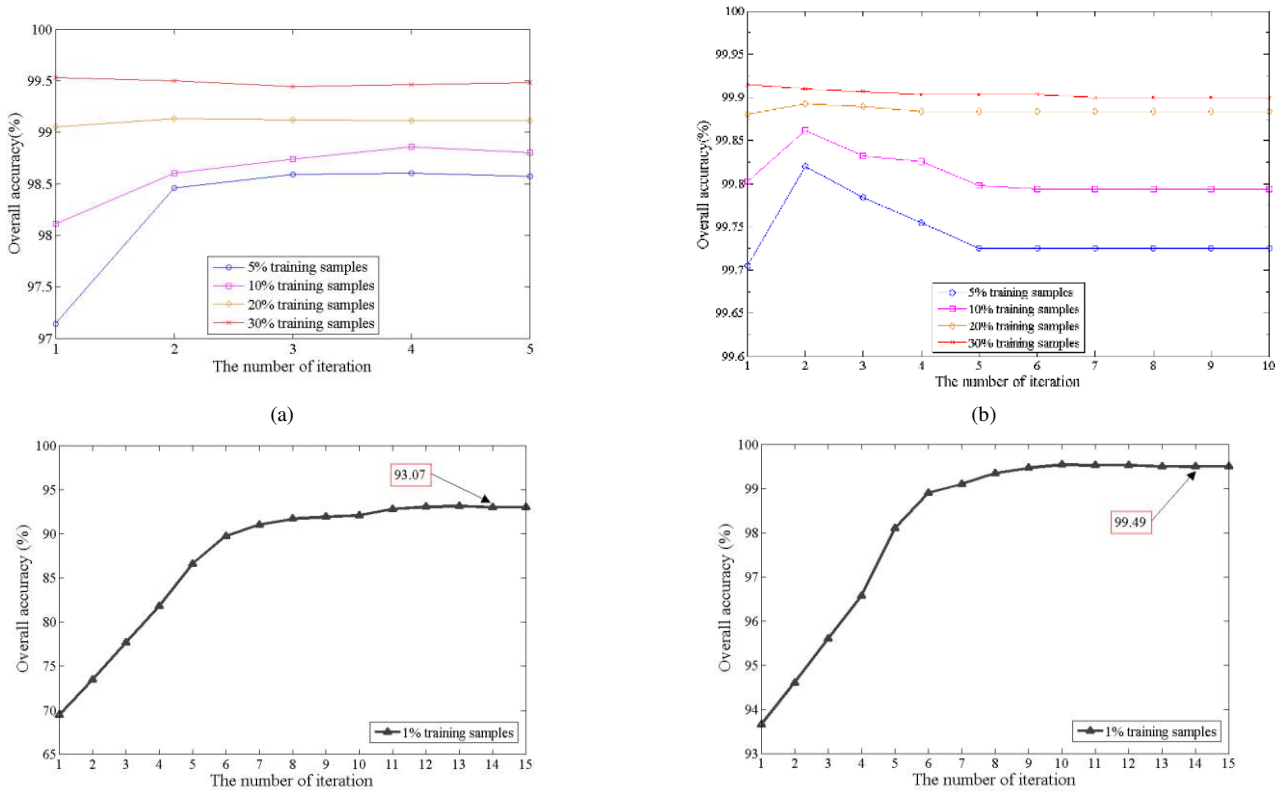
Fig. 10. Effects of non-local window size $W_{non\text{-}local}$ and the number of neighbors C using NE-MFAS (a) the Indian Pines and (b) the Pavia University.

times and the average results are recorded to avoid the bias induced by random sampling.

1) Effect of local window size

Firstly, we investigate the effect of the local window size W_{local} on the performance of MFS, MFAS and NE-MFAS using the two data sets. The local window size W_{local} ranges from 3×3 (W_{local} = 3) to 13×13 (W_{local} = 13) with the other parameters fixed. The OA values with the standard deviations for the two datasets are shown in Fig. 9, where the X-axis stands for the window size W_{local} and the Y-axis represents the OA.

From Fig. 9(a) and (b), we can see that the local window size W_{local} has certain impacts on the classification performance of the proposed methods. Compared with MFS, the OA results of NE-MFAS and MFAS are consistently higher than those of MFS, especially when the size of W_{local} is large. The fixed window used in MFS cannot change adaptively according to different situations and thereby contains false neighbors. MFS is very sensitive to the size of W_{local}. The performance decreases significantly with the increase of the window size. MFAS performs much better than MFS except for the case of W_{local} of 3×3. This demonstrates that our superpixel-based neighborhood method is effective for his classification. NE-MFAS is not sensitive to this parameter and has the best



(a) (b)

(c) (d)

Fig. 11. Effects of the number of iterations and the number of samples, (a) the Indian Pines, (b) the Pavia University,

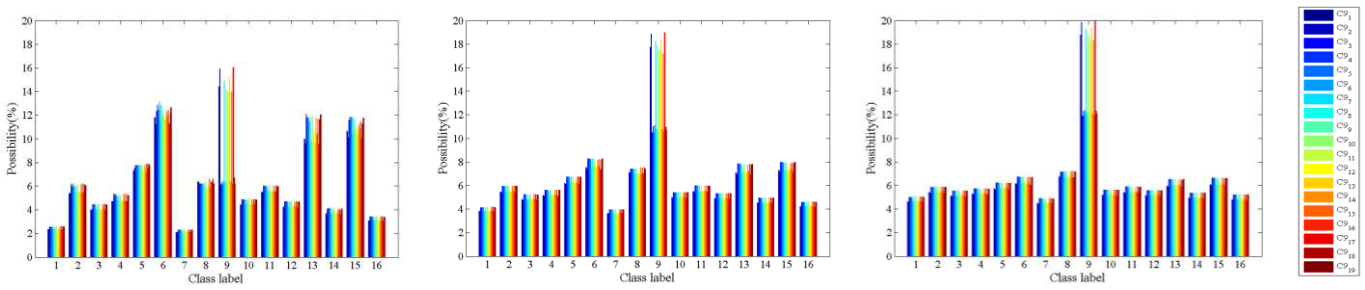


Fig. 12. The semantic representations of samples in 9-th class in Indian Pines dataset
(a) First order DSRs (b) Second order DSRs (c) Third order DSRs (d) legend

performance, even with a small window size of 3×3. According to the above experiments, we set W_{local} to be 5, 7 and15 for the Indian Pines, the University of Pavia and the Salinas, respectively.

2) Effect of non-local structure window size and the number of non-local neighbors

We examine the effects of the non-local structure window size $W_{non\text{-}local}$ and the number of the non-local neighbor C for NE-MFAS. For Indian Pines and University of Pavia datasets, the non-local structure window size $W_{non\text{-}local}$ varies from 9×9 to 25×25, and the number of the non-local neighbor C ranges from 0 to 100 with 20 as the interval. The OA values are shown in Fig. 13, where the X-axis is the non-local structure window size $W_{non\text{-}local}$, the Y-axis is the number of the neighbors C and the Z-axis stands for the OA.

Fig.10(a) and (b) reveal that NE-MFAS with a large size of $W_{non\text{-}local}$ (17 to 25, 21 for Indian Pines data set and 25 for Pavia University dataset) can lead to satisfactory classification accuracy at different numbers of neighbor C, except that C is equal to zero. For a fixed $W_{non\text{-}local}$, the number of non-local neighbors C has limited effects on the OA results. It favors a middle value (40 for Indian Pines dataset and 60 for Pavia University dataset) in consideration of a good balance between the OA results and computational efficiency. It is noticed that a large number of non-local neighbors may lead to poor results in a small number of classes. This is because the number of pixels belonging to these classes may be less than C. Many pixels belonging to the other classes are treated as non-local neighbors.

3) Effect of the number of training samples and the number of iterations

Finally, we discuss the effect of the number of samples and the number of iterations. For Indian Pines and University of Pavia datasets, the number of iterations ranges from 1 to 5 and about 5%, 10%, 20%, 30% of the labeled samples from each class are selected randomly as the training set. The time-consumption increases dramatically with the increasing of the number of the iterations. The OA results are illustrated in Fig. 11(a) and (b), where the X-axis is the number of the iterations and different colors stand for different numbers of the training samples. 1% training samples of the two datasets are shown in Fig. 11 (c) and (d) respectively.

As shown in Fig. 11, the performance of NE-MFAS is improved and the rate of convergence increases with the increased number of the training samples. We use the posteriori results to guide us to choose the number of iterations. For Indian Pines and Pavia University datasets with 5% training samples, the numbers of the iterations are set to be 3 and 2 respectively. We also can select the number of iterations by some convergence criteria (for example, the change rate of the class set of the test samples) which will cost more time due to the increase of the number of the iterations. The convergence result is the same in most cases. Even if there may have a tendency of decrease in some cases such as Fig. 11(b), the convergence result is 0.1% lower than the best result at most. Fig. 11(c) and (d) show that NE-MFAS can achieve satisfactory classification accuracy with 1% training samples for the two datasets, but the computational complexity due to the large number of iterations may be a drawback.

The 9-th class of Indian Pines is the smallest class with only twenty samples, and we have only one single training sample in our experiments, which always result in poor OA performance. We choose this class to illustrate the impact of the iterations on the performance of NE-MFAS. In Fig. 12, the semantic representation of every pixel by NE-MFAS in the 9-th class of the Indian Pines data set is shown, where the X-axis represents the class label, and the Y-axis is the probability of the pixel belonging to each class and different colors in the histogram stand for different pixels (19 testing samples in total). $C9_i$ in the legend represents the i-th sample of the 9-th class.

Fig. 12(a)-(c) are the first, second and third order DSRs respectively. The DSRs vary from a disorganized situation to a reasonable and stable settlement. With the increase of the number of the iterations, the result of DSRs converges to a stable value, finally. It is noticed that during this process, the probabilities of pixels belonging to the 9-th class increase, while the probabilities of pixels belonging to other classes decrease. Fig.12(c) shows that most of the pixels are classified to the 9-th class.

G. Computational time analysis

Time consumption in computation as an indicator of the algorithm performance is discussed. The computational time of all the considered methods on the Indian Pines data set is given in Table VIII. The experiments of all the methods are implemented using MATLAB R2014b on a 4.00 GHz Intel

CPU with 8GB of RAM.

As shown in Table VIII, the proposed MFAS is a faster than MFS, even though the superpixel segmentation takes time to accomplish. NE-MFAS is much slower than MFS and MFAS. Without considering the iterations of ICM, the computational complexity of MFS and MFAS are linear with respect to u and l where u is the number of the unlabeled pixels and l is the average number of the neighbors of the unlabeled pixels, i.e., $O(l \cdot u)$. For MFS, l is equal to W_{local} squared. However, l is less than W_{local} squared for MFAS due to the superpixel constraint, resulting in less time used for MFAS. In the same case, the computational complexity of NE-MFAS is $O((l+C)\cdot u)$ where C is the number of non-local neighbors of each pixel. Also there is a precomputation for the non-local similarity matrix, and hence the time-consumption is higher than the other two methods.

Usually u is much larger than l and C. Taking into account the ICM iteration, the computational complexity of our methods MFS, MFAS and NE-MFAS are similar to that of ICM which is proportional to $t_{\max}$ and u, i.e., $O(t_{\max}\cdot u)$ where $t_{\max}$ is the iteration times. Usually $t_{\max}$ is far less than u, and the computational complexity of the proposed methods is $O(u)$.

TABLE VIII
COMPUTATIONAL TIME ON INDIAN PINES DATA SET (SECOND)

METHOD	SVM	SRC	SVM-MRF	SVM-CK
TIME	28.75	58.11	108.93	43.17
METHOD	SOMP	MFS	MFAS	NE-MFAS
TIME	409.25	227.07	207.81	450.91

V. CONCLUSION

In this paper, we have proposed a novel method which could obtain semantic representation of each pixel with more detailed information and less noise for hyperspectral image classification. Firstly, different types of features were extracted to host comprehensive information of HSIs. Secondly, the probabilistic SVM was used to map these features which lie on different spaces to the same semantic space. Thirdly, in order to incorporate spatial information as well as multiple-semantic information, the modified MRF model has been applied, and also, in order to better describe the structure of the HSI, a new approach to construct adaptive windows has been proposed and used in our model. Furthermore, due to the redundancy of non-local spatial information, the non-local neighbors also have been exploited by a variation of KNN and incorporated in our MRF model at the same time, which makes the semantic representation of each pixel more meaningful. Finally, our model was transformed to a single optimization problem which can be solved by gradient descent. The experimental results on the three data sets have proved that our method outperforms the other state-of-the-art methods, and also can achieve good performance with small training samples.

The model proposed in this paper not only can combine multiple features but also incorporate different classifiers i.e. SVM, SRC, and MLR which obtain the probability results of each pixel and our model can assemble weak classifiers to become a stronger one.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (nos. 61272282, 61377011 and 61373111), and the Program for New Scientific and Technological Star of Shaanxi Province (No. 2014KJXX-45). H. Zhou is supported by UK EPSRC under Grants EP/N508664/1, EP/R007187/1 and EP/N011074/1, and Royal Society-Newton Advanced Fellowship under Grant NA160342.

REFERENCES

[1] P. K. Goel, S. O. Prasher, R. M. Patel, J. A. Landry, and R. B. Bonnell, "Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn," *Comput. Electron. Agric.*, vol. 39, no. 2, pp. 67-93, May 2003.

[2] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480-491, Mar. 2005.

[3] M. T. Eismann, A. D. Stocker, and N. M. Nasrabadi, "Automated Hyperspectral Cueing for Civilian Search and Rescue," *Proceedings of the IEEE* vol. 97, no. 6, pp. 1031-1055, June 2009.

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

[5] P. K. Gotsis, C. C. Chamis, and L. Minnetyan, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778-1790, Aug. 2004.

[6] X. Zhang, Y. He, N. Zhou, and Y Zheng, "Semisupervised dimensionality reduction of hyperspectral images via local scaling cut criterion," *IEEE Geosci. Remote Sens. Lett.,* vol. 10, no. 6, pp. 1547-1551, 2013.

[7] J. Li, J. M. Bioucas-Dias and A. Plaza, "Semisupervised Hyperspectral Image Classification Using Soft Sparse Multinomial Logistic Regression," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 318-322, Mar. 2013.

[8] F. Ratle, G. Camps-Valls and J. Weston, "Semisupervised Neural Networks for Efficient Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2271-2282, May 2010.

[9] 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.

[10] Y. Tarabalka, J. A. Benediktsson, J. Chanussot and J. C. Tilton, "A multiple classifier approach for spectral-spatial classification of hyperspectral data," *Geosci. Remote Sens. Symposium (IGARSS), 2010 IEEE International*, Honolulu, HI, 2010, pp. 1410-1413.

[11] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias and J. A. Benediktsson, "Generalized Composite Kernel Framework for Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816-4829, Sep. 2013.

[12] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93-97, Jan. 2006.

[13] M. Fauvel, J. Chanussot and J. A. Benediktsson, "Adaptive pixel neighborhood definition for the classification of hyperspectral images with support vector machines and composite kernel," *2008 15th IEEE International Conference on Image Processing*, San Diego, CA, 2008, pp. 1884-1887.

[14] Y. Chen, N. M. Nasrabadi and T. D. Tran, "Hyperspectral Image Classification Using Dictionary-Based Sparse Representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973-3985, Oct. 2011.

[15] X. Zhang; Q. Song; Z. Gao; Y. Zheng; P. Weng; L. C. Jiao, "Spectral-Spatial Feature Learning Using Cluster-Based Group SparseC oding for Hyperspectral Image Classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4142 - 4159, Sep. 2016.

[16] J. Li, H. Zhang and L. Zhang, "Efficient Superpixel-Level Multitask Joint Sparse Representation for Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5338-5351, Oct. 2015.

[17] 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.

[18] J. Xia, J. Chanussot, P. Du and X. He, "Spectral–Spatial Classification for Hyperspectral Data Using Rotation Forests With Local Feature Extraction and Markov Random Fields," *IEEE Trans. on Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2532-2546, May 2015.

[19] 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, Sept. 2011.

[20] M. Golipour, H. Ghassemian and F. Mirzapour, "Integrating Hierarchical Segmentation Maps With MRF Prior for Classification of Hyperspectral Images in a Bayesian Framework," *IEEE Trans. on Geosci. Remote Sens.*, vol. 54, no. 2, pp. 805-816, Feb. 2016.

[21] G. Moser and S. B. Serpico, "Combining Support Vector Machines and Markov Random Fields in an Integrated Framework for Contextual Image Classification," *IEEE Trans. on Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2734-2752, May 2013.

[22] P. Ghamisi, J. A. Benediktsson and M. O. Ulfarsson, "The spectral-spatial classification of hyperspectral images based on Hidden Markov Random Field and its Expectation-Maximization," *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*, Melbourne, VIC, 2013, pp. 1107-1110.

[23] L. Sun, Z. Wu, J. Liu, L. Xiao and Z. Wei, "Supervised Spectral–Spatial Hyperspectral Image Classification With Weighted Markov Random Fields," *IEEE Trans. on Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1490-1503, Mar. 2015.

[24] Y. Boykov, O. Veksler and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222-1239, Nov. 2001.

[25] Q. Lu, X. Huang, J. Li and L. Zhang. "A Novel MRF-Based Multifeature Fusion for Classification of Remote Sensing Images." *IEEE Geosci. Remote Sens. Lett.,* vol. 13, no .4, pp. 1-5, Apr. 2016.

[26] S. Z. Li and S. Singh, "Markov Random Field Modeling in Image Analysis," vol. 3. Berlin, Germany: Springer-Verlag, 2009.

[27] X. Song, S. Jiang and L. Herranz, "Joint multi-feature spatial context for scene recognition in the semantic manifold," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1312-1320.

[28] L. Zhang, L. Zhang, D. Tao and X. Huang, "On Combining Multiple Features for Hyperspectral Remote Sensing Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 879-893, Mar. 2012.

[29] X. Huang and L. Zhang, "An SVM Ensemble Approach Combining Spectral, Structural, and Semantic Features for the Classification of High-Resolution Remotely Sensed Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 257-272, Jan. 2013.

[30] E. Zhang, X. Zhang, L. Jiao, B. Hou, "Weighted Multifeature Hyperspectral Image Classification via Kernel Joint Sparse Representation." *Neurocomputing*, vol. 178, pp. 71-86, Nov. 2015.

[31] D. Zhang, X. Chen, and W. S. Lee. "Text classification with kernels on the multinomial manifold." *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval ACM*, pp. 266-273, 2015.

[32] M. Shi and G. Healey, "Hyperspectral texture recognition using a multiscale opponent representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 5, pp. 1090-1095, May 2003.

[33] T. C. Bau, S. Sarkar and G. Healey, "Hyperspectral Region Classification Using a Three-Dimensional Gabor Filterbank," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3457-3464, Sep. 2010.

[34] X. Huang and L. Zhang, "Comparison of Vector Stacking, Multi-SVMs Fuzzy Output, and Multi-SVMs Voting Methods for Multiscale VHR Urban Mapping," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 261-265, Apr. 2010.

[35] 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.

[36] B. Waske and J. A. Benediktsson, "Fusion of Support Vector Machines for Classification of Multisensor Data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3858-3866, Dec. 2007.

[37] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *Advances in Large Margin Classifiers*, vol. 10, pp. 61-74, Jun. 2000.

[38] Chang, ChihChung, and C. J. Lin, "LIBSVM: A library for support vector machines." *Acm Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 389-396, Apr. 2011.

[39] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721-741, Nov. 1984.

[40] R. Kwitt, N. Vasconcelos, and N. Rasiwasia. "Scene Recognition on the Semantic Manifold," *European Conference on Computer Vision (ECCV)* Springer-Verlag, vol. 7575, pp. 359-372, 2012.

[41] J. Besag, "On the Statistical-Analysis of Dirty Pictures," *Journal of the Royal Statistical Society*, vol. 48, no. 3, pp. 259-302, 1986.

[42] R. Roscher and B. Waske, "Superpixel-based classification of hyperspectral data using sparse representation and conditional random fields," *2014 IEEE Geoscience and Remote Sensing Symposium - IGARSS*, Quebec City, QC, pp. 3674-3677, 2014.

[43] A. Foi, V. Katkovnik and K. Egiazarian, "Pointwise Shape-Adaptive DCT for High-Quality Denoising and Deblocking of Grayscale and Color Images," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1395-1411, May 2007.

[44] M. Y. Liu, O. Tuzel, S. Ramalingam and R. Chellappa, "Entropy rate superpixel segmentation," *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, Providence, RI, pp. 2097-2104, 2011.

[45] A. Buades, B. Coll, and J. M. Morel. "A Review of Image Denoising Algorithms, with a New One." *Siam Journal on Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490-530, Jan. 2005.

[46] H. Zhang, J. Li, Y. Huang and L. Zhang, "A Nonlocal Weighted Joint Sparse Representation Classification Method for Hyperspectral Imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2056-2065, Jun. 2014.

[47] Y. Zhong, R. Feng and L. Zhang, "Non-Local Sparse Unmixing for Hyperspectral Remote Sensing Imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1889-1909, Jun. 2014.

[48] Y. Qian and M. Ye, "Hyperspectral Imagery Restoration Using Nonlocal Spectral-Spatial Structured Sparse Representation With Noise Estimation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 499-515, Apr. 2013.

[49] M. Jia, M. Gong, E. Zhang, Y. Li and L. Jiao, "Hyperspectral Image Classification Based on Nonlocal Means with a Novel Class-Relativity Measurement," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 7, pp. 1300-1304, Jul. 2014.