ISPRS Journal of Photogrammetry and Remote Sensing 94 (2014) 25–36

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Column-generation kernel nonlocal joint collaborative representation for hyperspectral image classification

Jiayi Li, Hongyan Zhang, Liangpei Zhang *

The State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, PR China

ARTICLE INFO

Article history:
Received 4 June 2013
Received in revised form 10 April 2014
Accepted 15 April 2014

Keywords:
Kernel method
Hyperspectral image classification
Joint collaboration model
Column generation

ABSTRACT

We propose a kernel nonlocal joint collaborative representation classification method based on column generation for hyperspectral imagery. The proposed approach first maps the original spectral space to a higher implicit kernel space by directly taking the similarity measures between spectral pixels as a feature, and then utilizes a nonlocal joint collaborative regression model for kernel signal reconstruction and the subsequent pixel classification. We also develop two kinds of specific radial basis function kernels for measuring the similarities. The experimental results indicate that the proposed algorithms obtain a competitive performance and outperform other state-of-the-art regression-based classifiers and the classical support vector machines classifier.

1. Introduction

With the rich spectral information, which spans the visible to infrared spectrum in hundreds of continuous narrow spectral bands, hyperspectral imaging has rapidly become an effective remote sensing technology for analyzing a variety of materials (Nidamanuri and Zbell, 2011; Plaza et al., 2009; Zhang et al., 2012a). Among the various applications, pixel classification aiming at categorizing pixels in a scene into a specific class has been an important task for the subsequent analysis and processing (Senthil et al., 2010; Zheng et al., 2013). To date, various classification techniques have been proposed. Zhong and Wang (2010) formulated a conditional random field (CRF) model (Lafferty et al., 2001) which utilizes the strong dependencies across spatial and spectral neighbors for the classification of hyperspectral images (HSIs). A novel classification framework based on spectral unmixing concepts was introduced by Dópido et al. (2012). Li et al. (2012) constructed a multinomial logistic regression based classification with a new family of generalized composite kernels when combining the spectral and the spatial information contained in hyperspectral data.

In real analysis scenarios, the supervised classification of such a high-dimensional dataset is still a difficult task. For example, the Hughes phenomenon appears as the dimensionality increases, and the collection of labeled training samples is generally difficult, expensive, and time-consuming. Another challenge is that the high-dimensional feature for an HSI often tends to be linearly inseparable. Therefore, many different techniques have been developed to deal with these obstacles. Among the various methods, support vector machines (SVM) with kernel tricks (Boser et al., 1992; Vapnik, 1999; Melgani and Bruzzone, 2004), which maps the original feature space into a higher dimensional kernel feature space to deal with the nonlinear problem, exploits the partial meaningful training samples as support vectors to construct an optimal separating nonlinear hyperplane. In this context, SVM, which aims at discriminating two different materials, has shown an excellent performance in supervised HSI classification. Tsang et al. (2006) presented a core vector machine (CVM) method which can optimize both the time and space complexities of the standard SVM algorithm for a large-scale dataset. Fauvel et al. (2008) incorporated the spatial information into a spectral-only SVM classifier by the fusion of the morphological information and the original hyperspectral data. Demir and Erturk (2010) utilized the empirical mode decomposition (EMD) of HSIs to increase the classification accuracy when using an SVM-based classification (Gualtieri and Cromp, 1999; Melgani and Bruzzone, 2004).

In recent years, a novel collaborative linear regression approach for object recognition has been introduced into high-dimensional classification tasks (Zhang et al., 2012b; Waqas et al., 2012; Yang et al., 2012), where the use of collaborative representation (CR) often leads to high computational efficiency and a desirable performance. The CR technique has also been applied to HSI classification (Li et al., 2013a,b), relying on the observation that the hyperspectral test pixel

* Corresponding author. Tel.: +86 13995556225.
E-mail address: zlp62@whu.edu.cn (L. Zhang).

http://dx.doi.org/10.1016/j.isprsjprs.2014.04.014
0924-2716/© 2014 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.

can be approximately represented by a given dictionary constructed from training samples. In the CR-based classification procedure, the training samples belonging to the same class contribute most to the test pixel in the linear representation, while the rest of the training samples act as collaborative assistants. Compared with the conventional SVM-based classifier, the CR-based classifier works from a reconstruction point of view and is more suitable for a multi-class classification task, while the SVM-based classifier is essentially a binary classifier. Moreover, the CR-based classifier can conduct classification tasks with a dictionary constructed from training samples, without an explicit learning stage; however, the sparse support vector for discriminating hyperplane construction must be trained before classifying the test data in the SVM classifier. With the aid of the rest of the training samples, the CR-based classifier can also work effectively in the case of a lack of samples (Zhang et al., 2012b; Li et al., 2013c).

In an HSI scene, it is natural that the pixels in a homogeneous region consist of similar materials (i.e. pixels with high spectral correlation) and can be used to improve the classification performance. Previous studies have shown the importance of incorporating the spatial neighborhood information into the classification (Fauvel et al., 2008). Several methods have been implemented by either combining the contextual and spectral information in the classification stage or post-processing, with the decisions obtained from individual pixels by spatial filtering. For the CR-based approaches, Waqas et al. (2012) presented a neighboring smooth constraint classification scheme, and Chen et al. (2011) applied the joint sparsity model (JSM) (Duarte et al., 2005) with contextual information for HSI classification. In addition, Zhang et al. (2013) proposed to utilize the nonlocal neighborhood information in the JSM classification model.

Since a hyperspectral dataset is not linearly separable, the linear regression based models cannot cope well with the nonlinear classification problem. Recently, a number of methods have been proposed to deal with this limitation. Qian et al. (2012) mixed several linear sub-models constructed in the corresponding partial input feature space with a final output classifier for the nonlinear task. Li et al. (2013d) developed a new framework for generalized composite kernel machines for HSI classification. Among these techniques, the kernel methods (Kwon and Nasrabadi, 2006) that implicitly exploit the high-order structure of the given data that cannot be captured by a linear version are often utilized and can show a significant improvement (Chen et al., 2013).

In this paper, we propose a kernel nonlocal joint collaboration model via a column-generation (CG) technique (Bi et al., 2004; Yuan et al., 2012) for HSI classification. This method first maps the original spectral space to a higher implicit kernel space by directly taking the similarity measures between spectral pixels as a feature, and then utilizes a nonlocal joint collaborative regression model for the kernel signal reconstruction and the subsequent pixel classification. After the explicit kernel dictionary and the explicit kernel signal are obtained, the linear model can be directly extended to a kernel version. Unlike the kernel trick used in various other approaches (Chen et al., 2013; Kwon and Nasrabadi, 2006), the CG strategy is easy to implement and does not require the explicit inner product structure in the regression analysis solution. We also develop two kinds of specific radial basis function (RBF) kernels for measuring the similarities, which are proved to be effective in the experiment section. The proposed method is aimed at dealing with the nonlinear phenomenon in HSI classification and achieving an improved performance. Experiments with several different hyperspectral datasets that have been widely used as public evaluation data confirm the effectiveness of the two proposed kernel algorithms.

The remaining parts of this paper are organized as follows. Section 2 introduces the nonlocal joint collaborative representation classification method. Section 3 defines the kernel joint collaboration model via column generation and proposes two specific RBF kernels for hyperspectral imagery. The experimental results of the proposed classification algorithms with two hyperspectral datasets are given in Section 4. Finally, Section 5 summarizes the paper, with a discussion on the findings and our ideas for extending the work.

2. Nonlocal joint collaborative representation classification

In this section, we first review the classical collaborative representation classification (CRC), and we then introduce a joint collaboration model which can be considered as a matrix-oriented extended version. Finally, we incorporate the discriminated spatial neighborhood information in the matrix-oriented extended version by constructing a nonlocal joint signal matrix, which consists of the highly correlated pixels in the neighboring window of the test pixel.

2.1. Classical collaborative representation classification

In this paper, every hyperspectral pixel can be denoted as a B-dimensional vector, where B refers to the number of bands of the HSI. For classification, suppose we have M distinct classes and N_i ($i = 1,\ldots, M$) training samples for each class. In the classical collaboration model, training samples from the ith class act as columns of a sub-dictionary $\mathbf{A}_i = \left[\mathbf{a}_{i,1}, \mathbf{a}_{i,2}, \ldots \mathbf{a}_{i,N_i}\right] \in \mathbb{R}^{B\times N_i}$. The collaborative dictionary $\mathbf{A} \in \mathbb{R}^{B\times N}$ with $N = \sum_{i=1}^{M} N_i$ is then constructed by combining all the sub-dictionaries $\{\mathbf{A}_m\}_{m=1,\ldots,M}$. Thus, a test pixel $s \in \mathbb{R}^B$ which belongs to the jth class can be written as a collaborative linear combination of all of the training samples as:

$$\begin{aligned}\mathbf{s} &= \mathbf{A}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} = \mathbf{A}_1\boldsymbol{\alpha}_1 + \ldots + \mathbf{A}_j\boldsymbol{\alpha}_j + \ldots + \mathbf{A}_M\boldsymbol{\alpha}_M + \boldsymbol{\varepsilon} \\ &= \mathbf{A}_j\boldsymbol{\alpha}_j + \sum_{k=1,k\neq j}^{M} \mathbf{A}_k\boldsymbol{\alpha}_k + \boldsymbol{\varepsilon} \in \mathbb{R}^B \end{aligned} \quad (1)$$

where the whole space constitutes a dominant low-dimensional subspace spanned by $\mathbf{A}_j$, and a complementary subspace set spanned by the rest of the training samples, which can be considered as an external collaborative partner to the dominant subspace. $\alpha \in \mathbb{R}^N$ is a coefficient vector and ε is random noise. For high-dimensional data classification, Zhang et al. (2012b) suggested that the ℓ_2-norm regularization could ensure a stable and discriminative representation coefficient for (1) to reconstruct the test pixel. In an HSI with Gaussian noise, the collaborative coefficient vector $\boldsymbol{\alpha}$ can be easily obtained by solving the following optimization problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \|\mathbf{s} - \mathbf{A}\boldsymbol{\alpha}\|_2 + \lambda\|\boldsymbol{\alpha}\|_2 \right\} \quad (2)$$

where λ makes a tradeoff between the data fidelity term and the penalty constraint term. For the classification, $\hat{\boldsymbol{\alpha}}_i$ is the coding vector associated with class i, and the ℓ_2-norm $\|\hat{\boldsymbol{\alpha}}_i\|_2$ also brings some discriminative information. The classification rule for CRC via regularized least squares, which is also referred to as CRC_RLS (Zhang et al., 2012b), is denoted as:

$$\mathbf{class}(\mathbf{s}) = \arg\min_{i=1,\ldots,M} \|\mathbf{s} - \mathbf{A}_i\hat{\boldsymbol{\alpha}}_i\|_2 / \|\hat{\boldsymbol{\alpha}}_i\|_2 \quad (3)$$

2.2. Joint collaboration model (JCM)

Considering the spatial consistency of the HSI, pixels in a spatial neighborhood can be simultaneously represented to assist with the classification of the center test pixel in the spatial window. In view of this, let $\mathbf{S} = \{\mathbf{s}_t\}_{t=1,\ldots,T}$ be T pixels in a spatial neighborhood centered at test pixel s_1. These pixels can then be represented by:

$$\mathbf{S} = [\mathbf{s}_1 \ \mathbf{s}_2 \ldots \ \mathbf{s}_T] = [\mathbf{A}\boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}_1 \ \mathbf{A}\boldsymbol{\alpha}_2 + \boldsymbol{\varepsilon}_2 \ldots \ \mathbf{A}\boldsymbol{\alpha}_T + \boldsymbol{\varepsilon}_T]$$
$$= \mathbf{A}\underbrace{[\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2 \ldots \boldsymbol{\alpha}_T]}_{\boldsymbol{\Psi}} + \boldsymbol{\Sigma} = \mathbf{A}_j\boldsymbol{\Psi}_j + \sum_{k=1,k\neq j}^{M} \mathbf{A}_k\Psi_l + \boldsymbol{\Sigma} \quad (4)$$

where $\boldsymbol{\Psi}$ is the set of all the coefficient vectors $\{\boldsymbol{\alpha}_t\}_{t=1,\ldots,T}$, and $\boldsymbol{\Psi}_j$ is a sub-set of $\boldsymbol{\Psi}$, which corresponds to all the pixels in the neighboring window. It is assumed that all the neighboring pixels share the same low-dimensional dominant subspace $\mathbf{A}_j$ with different coefficients. $\boldsymbol{\Sigma}$ is the model noise matrix corresponding to the joint signal matrix. In this way, the labeling process can be denoted as:

$$\mathbf{class}(\mathbf{s}_1) = \arg\min_{j=1,\ldots,M}\left\{\|\mathbf{S} - \mathbf{A}_j\hat{\boldsymbol{\Psi}}_{(j)}\|_F / \|\hat{\boldsymbol{\Psi}}_{(j)}\|_F\right\} \quad (5)$$

where $\mathbf{A}_j$ is a sub-part of $\mathbf{A}$ corresponding to class j, $\hat{\boldsymbol{\Psi}}_{(j)}$ denotes the corresponding portion of the recovered collaborative coefficients of the jth class, and $\|\cdot\|_F$ denotes the Frobenius norm.

2.3. Nonlocal joint collaborative representation classification (NJCRC)

While similar pixels tend to be clustered in an image spatial neighborhood, there will still be pixels with low correlation in heterogeneous areas, especially around the image edges. In order to address this problem, we select a large neighborhood window centered at test pixel $\boldsymbol{s}_1$, with a size of $\sqrt{T} \times \sqrt{T}$, and consider that only the $K - 1(K \leqslant T)$ neighboring pixels which are similar to $\boldsymbol{s}_1$ should be stacked into the nonlocal joint signal matrix $\mathbf{S}_K \in \mathbb{R}^{B\times K}$, while the other pixels in the neighborhood window should be discarded. Here, we use the K-NN (K-nearest neighbors) method (Keller et al., 1985) to construct the joint signal sub-matrix, as follows. We select the first K neighboring pixels, referred to as $\{\boldsymbol{s}_k\}_{k=1,\ldots,K}$, from all the T pixels, which are reordered by the spectral correlation between the host pixel with $\boldsymbol{s}_k$, and we stack them as a new nonlocal joint signal matrix $\mathbf{S}_K$. It is believed that these K pixels $\{\boldsymbol{s}_k\}_{k=1,\ldots,K}$ share a "common collaboration pattern" (Li et al., 2013a) as they are selected by the measure of the correlations between the central test pixels$_1$, not the spatial distance.

In this way, the nonlocal version of (4) can be extended by solving the following joint collaborative recovery via a Frobenius norm optimization problem:

$$\hat{\boldsymbol{\Psi}}_K = \arg\min_{\boldsymbol{\Psi}_K}\{\|\mathbf{S}_K - \mathbf{A}\boldsymbol{\Psi}_K\|_F^2 + \lambda\|\boldsymbol{\Psi}_K\|_F^2\} \quad (6)$$

where $\boldsymbol{\Psi}_K$ refers to the coefficient matrix corresponding to the nonlocal joint signal matrix. The classification rule should also be modified as follows:

$$\mathbf{class}(\mathbf{s}_1) = \arg\min_{j=1,\ldots,M}\left\{\|\mathbf{S}_K - \mathbf{A}_j\hat{\boldsymbol{\Psi}}_{K(j)}\|_F / \|\hat{\boldsymbol{\Psi}}_{K(j)}\|_F\right\} \quad (7)$$

where $\mathbf{A}_j$ is a sub-part of $\mathbf{A}$ corresponding to class j, and $\hat{\boldsymbol{\Psi}}_{K(j)}$ denotes the corresponding portion of the recovered collaborative coefficients of the ith class.

3. Kernel collaborative representation via column generation

In this section, we first introduce the kernel function and the two distance measurements for the RBF kernel, and we then map the HSI into a kernel feature space, in which the classes are assumed to be linearly separable.

3.1. Kernel function

Although spanning the visible to infrared spectrum in hundreds of continuous narrow spectral bands, HSIs are well known to be linearly inseparable, and should not be represented as fixed-size spectral feature vectors. The kernel method (Li et al., 2013d), a commonly used approach to deal with such nonlinear problems, is to assume that we have some way of measuring the similarity between pixels that does not require them to be preprocessed into a feature vector format (Murphy, 2012). We consider the similarity measurement as a real-value function of two arguments denoted as a pair of pixels. The real-value function $\kappa : \mathbb{R}^B \times \mathbb{R}^B \mapsto \mathbb{R}$ is defined as the inner product:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle \quad (8)$$

where $\boldsymbol{x}_i$ is the spectral pixel at location i in an HSI, and $\boldsymbol{x}_j$ is the one at location $j.\phi(\boldsymbol{x})$ is a function of the spectral vector. Commonly used kernels include the RBF kernel (Suykens and Vandewalle, 1999), the linear kernel (Yang et al., 2009), the string kernel (Leslie et al., 2002), and so on. As the feature space of the RBF kernel has an infinite number of dimensions, and the value of the RBF kernel decreases with distance, and ranges between [0, 1], it can be readily interpreted as a similarity measure (Philippe et al., 2004). We utilize the RBF kernel $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \times dist(\boldsymbol{x}_i, \boldsymbol{x}_j))$, with $\gamma > 0$ controlling the width of the RBF kernel, where $dist(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the distance measurement.

Zhang et al. (2007) noted that different RBF kernel functions can be fixed with a specific distance measurement $dist(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The first distance measurement used in this paper is the Euclidean distance, and this kernel function, which is denoted as "KE", can be rewritten as:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/\sigma\right) \quad (9)$$

where $\sigma = 1/\gamma$. This Euclidean distance focuses on the absolute difference between a pixel pair. Another generalized RBF kernel function is constructed with the chi-squared distance: $\chi^2(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2}\sum_{b=1}^{B}\frac{(\mathbf{x}_i(b)-\mathbf{x}_j(b))^2}{\mathbf{x}_i(b)+\mathbf{x}_j(b)}$, which can reflect the relative difference between corresponding spectral sub-regions. The second kernel function, which is denoted as "KC", can be represented as:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\chi^2(\mathbf{x}_i, \mathbf{x}_j)/\mu\right) \quad (10)$$

where μ is set to the mean value of the pairwise chi-squared distance and is adaptive to the training set.

3.2. The column-generation technique

Column generation (Nash and Sofer, 1996) has been widely used in linear programming since the 1950s. The kernel mapping in this paper, which directly takes the signal in the kernel space as a feature (Yuan et al., 2012), is similar to the simplified column-generation strategy for CG-Boost in multiple kernel leaning (Bi et al., 2004; Gehler and Nowozin, 2009).

Denote $\mathbf{s} \in \mathbb{R}^B$ as the data point of interest and $\mathbf{s}' \in \mathbb{R}^N$ as its representation in the kernel feature space. The kernel collaborative representation of test pixel s in terms of all the training pixels can then be formulated as:

$$\mathbf{s}' = [\kappa(\mathbf{a}_1,\mathbf{s})\cdots\kappa(\mathbf{a}_N,\mathbf{s})]^T = \underbrace{\begin{pmatrix}\kappa(\mathbf{a}_1,\mathbf{a}_1)\cdots\kappa(\mathbf{a}_1,\mathbf{a}_N)\\ \vdots \\ \kappa(\mathbf{a}_N,\mathbf{a}_1)\cdots\kappa(\mathbf{a}_N,\mathbf{a}_N)\end{pmatrix}}_{\kappa(\mathbf{A})}\underbrace{[\alpha_1'\cdots\alpha_N']^T}_{\boldsymbol{\alpha}'} = \kappa(\mathbf{A})\boldsymbol{\alpha}' \quad (11)$$

where the columns of $\kappa(\mathbf{A})$ are the representation of the training samples in the feature space , and $\boldsymbol{\alpha}'$ is assumed to be a $N \times 1$ kernel representation vector.

For the nonlocal joint representation model, we can also extend the nonlocal joint signal matrix $\mathbf{S}_K \in \mathbb{R}^{B\times K}$ into the kernel feature space. We first map all the T pixels into the kernel feature space, then select the first K signals in the kernel space with the K-NN approach, as described in Section 2.3, and we finally stack these

K signals as $\mathbf{S}'_K$, considering that they share a "common collaborative pattern" in the feature space. In this way, the nonlocal kernel joint signal matrix can be collaboratively represented in the kernel feature space as:

$$\mathbf{S}'_K = [\mathbf{s}'_{(1)} \cdots \mathbf{s}'_{(K)}] = \underbrace{\begin{pmatrix} \kappa(\mathbf{a}_1, \mathbf{a}_1) \cdots \kappa(\mathbf{a}_1, \mathbf{a}_N) \\ \vdots \\ \kappa(\mathbf{a}_N, \mathbf{a}_1) \cdots \kappa(\mathbf{a}_N, \mathbf{a}_N) \end{pmatrix}}_{\kappa(\mathbf{A})} \underbrace{\begin{pmatrix} \alpha'_{1,(1)} \cdots \alpha'_{1,(K)} \\ \vdots \\ \alpha'_{N,(1)} \cdots \alpha'_{N,(K)} \end{pmatrix}}_{\mathbf{\Psi}'_K} = \kappa(\mathbf{A})\mathbf{\Psi}'_K \tag{12}$$

where $\mathbf{\Psi}'_K$ is the kernel collaborative coefficient matrix, and the function (6) can be extended as :

$$\hat{\mathbf{\Psi}}'_K = \arg\min_{\mathbf{\Psi}'_K} \{\|\mathbf{S}'_K - \kappa(\mathbf{A})\mathbf{\Psi}'_K\|_F^2 + \lambda\|\mathbf{\Psi}'_K\|_F^2\} \tag{13}$$

The solution of (13) can be easily and analytically derived as:

$$\hat{\mathbf{\Psi}}'_K = \left(\kappa(\mathbf{A})^T\kappa(\mathbf{A}) + \lambda \cdot I\right)^{-1} \kappa(\mathbf{A})^T\mathbf{S}'_K \tag{14}$$

Once the coefficient matrix $\hat{\mathbf{\Psi}}'_K$ is obtained, the classification rule, which is analogous to (7), is denoted as:

$$\mathbf{class}(\mathbf{s}_1) = arg \min_{j=1,\ldots,M} \|\mathbf{S}'_K - \kappa(\mathbf{A}_j)\hat{\mathbf{\Psi}}'_{K,j}\|_F^2 / \|\hat{\mathbf{\Psi}}'_{K,j}\|_F^2 \tag{15}$$

where $\kappa(\mathbf{A}_j)$ is a sub-part of $\kappa(\mathbf{A})$ in class j, and $\hat{\mathbf{\Psi}}'_{K,j}$ denotes the portion of the recovered kernel collaborative coefficients corresponding to the entire training samples in the jth class.

3.3. Computational complexity analysis

Suppose the size of dictionary A is $B \times N$, the neighborhood size is T, and the amount of all the test pixels is n. The computational complexity of the proposed algorithms consists of three parts, as follows. First, the original spectral feature is mapped into the feature space. Second, the K-NN approach is introduced to select the nonlocal kernel feature signal, the computational complexity of which is $O(T(T-1)/2)$. Although the dominant computational cost of the algorithms comes from the closed-form solution of the coding coefficient matrix, as shown in Eq. (14), we also find that the projection matrix $\mathbf{P} = \left(\kappa(\mathbf{A})^T\kappa(\mathbf{A}) + \lambda \cdot \mathbf{I}\right)^{-1}\kappa(\mathbf{A})^T$ can be computed offline, which can accelerate the computation of the coding, and the second term $\mathbf{S}' \in \mathbb{R}^{N \times K}$ is acquired in the CG way, which costs $O(KN)$. In view of this, the final complexity for the whole hyperspectral dataset is $O(n(KN + KN^2))$, where $K \leqslant T$ is the number of neighboring pixels which are similar to the test pixel in the kernel feature space.

3.4. Procedure of the proposed kernel algorithm

By incorporating the nonlocal spatial structure information, the implementation details of the proposed kernel NJCRC algorithm are summarized in Table 1.

4. Experiments and discussion

4.1. Experimental design and datasets

The goal of the experiments is to investigate the effectiveness of the proposed algorithms in the classification of hyperspectral datasets. The classifications of standard SVM and SVM-NS (which combine the spectral and contextual neighborhood information by stacking) with the RBF kernel in the conventional reproducing kernel Hilbert space (RKHS) (Melgani and Bruzzone, 2004; Plaza et al., 2009), SRC (Wright et al., 2009) with an improved ℓ_1-norm algorithm named Lasso (Tibshirani, 2011), and JSRC with a greedy pursuit algorithm (referred to as SOMP in Tropp et al. (2006)) are used as benchmarks in this paper. Moreover, we also compare the proposed nonlinear CR-based algorithms on two specific kernel functions with the corresponding linear versions, including CRC, JCRC, and NJCRC. The parameters for SVM, including γ and σ, and those for SVM-NS, including γ, σ, and the size of the neighborhood, are obtained by cross-validation.

The two real-world hyperspectral datasets used for the experiments are briefly described as follows.

This scene gathered by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor in June 12, 1992 over the Indian Pines test site in north-western Indiana consists of 145 × 145 pixels and 224 spectral reflectance bands in the wavelength range 0.4–2.5 μm. The false color composite of the Indian Pines image is shown in Fig. 1(a). We also reduced the number of bands to 200 by removing bands covering the regions of water absorption: [104–108], [150–163], and 220 (Gualtieri and Cromp, 1999). The spatial resolution for this image is about 20 m. In this image, we randomly sample 60 pixels for each class as the training samples and the rest as the test pixels. This image contains 10 ground-truth classes which is visually shown in Fig. 1(b), and the numbers of the training and test sets are shown in Table 2.

The next experimental image, which is named the Pavia University scene, was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor during flight campaigns in 2003 over the Pavia area, northern Italy. The ROSIS sensor generates 115 spectral bands ranging from 0.43–0.86 μm with a geometric resolution of 1.3 m per pixel. With several of the noisy bands removed, this image contains 103 available bands. For the Pavia University image sized as 610 × 340 pixels, it contains nine ground-truth classes, as shown in Table 3. From the ground truth, we randomly select 40 pixels for each class as the training samples, and the rest as the test samples to validate the performances of the aforementioned classifiers. The false color composite of the Pavia University image is shown in Fig. 2(a), and the label of each ground truth pixel can be shown in Fig. 2(b).

4.2. Experimental results

The visual classification results for the two datasets are shown in Figs. 1 and 2, respectively. Tables 5 and 7 summarize the classification accuracies of the methods under comparison. In these tables, the OA is the ratio between the correctly classified test pixels and the total number of test samples. The quantity disagreement Q and the allocation disagreement A (Pontius and Millones, 2011) are two robust and informative measures of the degree of disagreement. In Tables 5 and 7, the best result for each quality index is labeled in bold, while the sub-optimal one is underlined. The classification accuracies using the different classifiers with the test set for each class can be found in the corresponding columns.

4.2.1. Indian Pines image

For the Indian Pines dataset, the regularization parameter λ for the CR-based classification algorithms (it is noted that SRC with the ℓ_1-norm regularization is a special instance of the generalized CR-based algorithm) ranges from $1e-8$ to 10. For the NJCRC-KE, NJCRC-KC, and linear NJCRC algorithms, the number of the joint sparse atoms K is chosen between $K = 40$ and $K = 100$. The neighborhood window size T for the spatially extended classifiers ranges from 9 to 169, and the kernel parameter σ for the KE-related algorithms ranges from $1e-3$ to $1e-1$. The optimal sets of parameters for the nine ℓ_2-norm based algorithms are shown in Table 4. In addition, the parameters for SVM, SVM-NS, SRC, and JSRC are set as the corresponding optimal, and the optimal sizes of the spatial neighborhood for JSRC and SVM-NS are 25 and 169, respectively.

Table 1
The kernel NJCRC algorithm for HSI classification.

```
Input: (1) An HSI containing training samples and a test set, in which the test pixel located at pcan be represented as $\mathbf{s}_p \in \mathbb{R}^B$
  (2) An entire dictionary $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_M] \in \mathbb{R}^{B \times N}$ for $M$ classes
  (3) Parameters: regulation parameter $\lambda$, spatial neighborhood size $T$ , the number of the joint signal $K$, and the kernel parameter $\sigma$ for the Euclidean distance
    (this parameter is omitted if the chi-squared distance is utilized)
Initialization: Construct the entire dictionary $A$ with all the training set in this HSI; normalize the columns of $A$ to have a unit $\ell_2$-norm and map the dictionary into the
    kernel feature space $\kappa(\mathbf{A})$
Main iteration:
  For each test pixel in the HSI:
  1. Construct the initial joint signal matrix $\mathbf{S} = [\mathbf{s}_1 \mathbf{s}_2 \ldots \mathbf{s}_T] \in \mathbb{R}^{B \times T}$, where $\boldsymbol{s}_1$ locates at the center of the neighborhood window, and map this matrix to the kernel feature
     space
  2. Construct the nonlocal joint collaborative matrix $\mathbf{S}'_K$ in the kernel feature space
  3. Code $\mathbf{S}'_K$ over $\kappa(\mathbf{A})$with Eq. (13)
  4. Compute the regularized residuals and label the test pixel with Eq. (15)
  5. Turn to the next pixel
  End for
Output: A 2-D matrix which records the labels of the HSI
```

The classification maps of the various classification methods are visually shown in Fig. 1(c–o), respectively. The quantitative accuracy results, which include the classification accuracy for every class, the overall accuracy, quantity disagreement, and the allocation disagreement, are shown in Table 5. For the Indian Pines image with a medium spatial resolution, the improvements are mainly caused by the spatial smoothness of the HSI, as the number of nonlocal neighboring pixels is large. With the help from the nonlocal neighboring pixels, the "salt and pepper" phenomenon can be significantly alleviated, especially for the pixels located in the inner part of a block, as shown in Fig. 1. For the classification accuracy, the improvement of the nonlocal spatial information based classifiers over the spatial contextual prior based classifiers suggests that the nonlocal joint signal selection can further improve the performance, as can be seen in Table 5. Details of the improvements are further shown in Section 4.3.2.

4.2.2. Pavia University image

For the Pavia University image, the regularization parameter λ for the CR-based classification algorithms ranges from $1e-8$ to 10. For the NJCRC-KE, NJCRC-KC, and linear NJCRC algorithms, the number of the joint sparse atoms K is chosen between $K = 30$

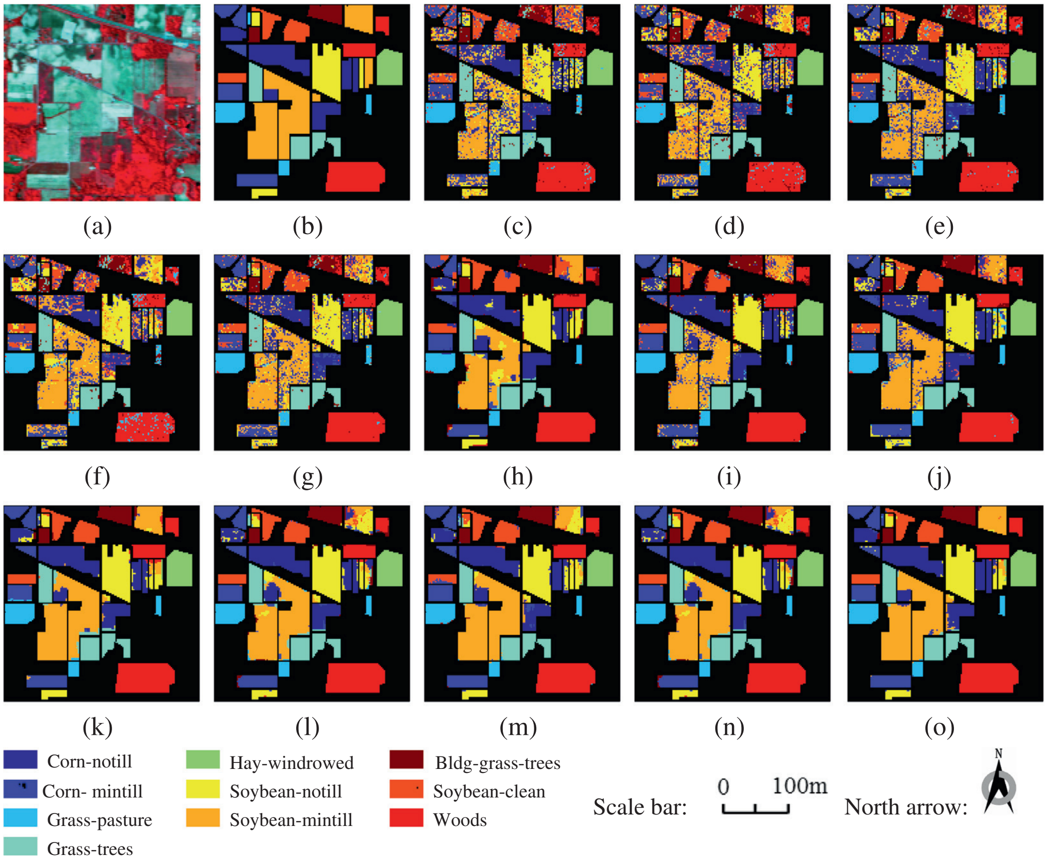


Fig. 1. Classification results with the Indian Pines image: (a) false color image (R:57, G:27, B:17), (b) ground truth, (c) CRC, (d) SRC, (e) SVM, (f) CRC-KE, (g) CRC-KC, (h) JCRC, (i) JSRC, (j) SVM-NS, (k) NJCRC, (l) JCRC-KC, (m) NJCRC-KC, (n) JCRC-KE, and (o) NJCRC-KE. In addition, the legend, scale bar, and north arrow of this image are shown in the last row. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
The ten ground-truth classes in the AVIRIS Indian Pines image dataset, and the training and test sample sets for each class.

No.	Class name	Training samples	Test samples
1	Corn-notill	60	1368
2	Corn-mintill	60	770
3	Grass-pasture	60	423
4	Grass-trees	60	670
5	Hay-windrowed	60	418
6	Soybean-notill	60	912
7	Soybean-mintill	60	2395
8	Soybean-clean	60	533
9	Woods	60	1205
10	Buildings-grass-trees	60	326
	Total	600	9020

Table 3
The nine ground-truth classes in the ROSIS Pavia University dataset, and the training and test sample sets for each class.

No.	Class name	Training samples	Test samples
1	Asphalt	40	6591
2	Meadows	40	18,609
3	Gravel	40	2059
4	Trees	40	3024
5	Metal sheet	40	1305
6	Bare Soil	40	4989
7	Bitumen	40	1290
8	Brick	40	3642
9	Shadows	40	907
	Total	360	42,416

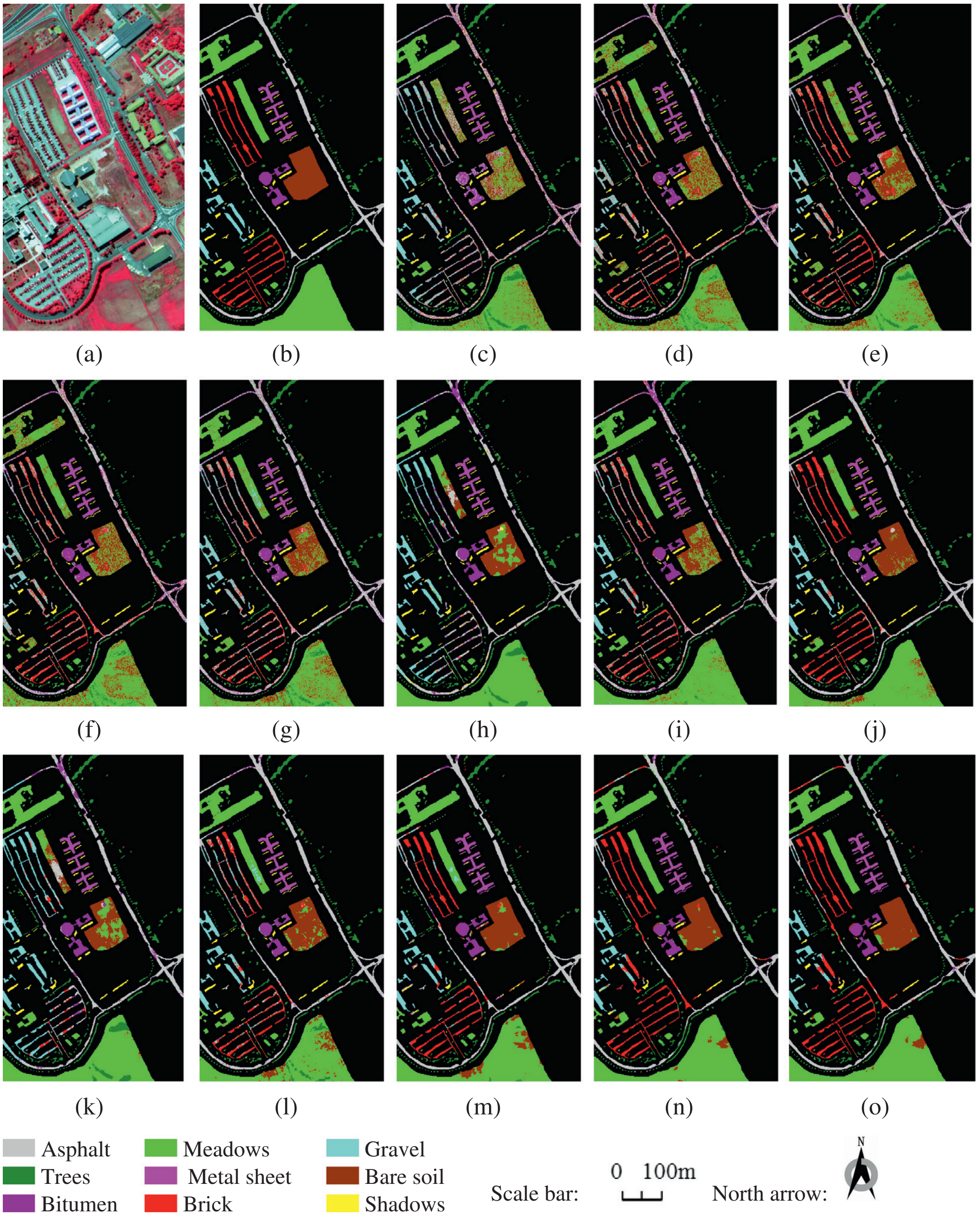


Fig. 2. Classification results with the Pavia University image: (a) false color image, (R:102, G:56, B:31), (b) ground truth, (c) CRC, (d) SRC, (e) SVM, (f) CRC-KE, (g) CRC-KC, (h) JCRC, (i) JSRC, (j) SVM-NS, (k) NJCRC, (l) JCRC-KC, (m) NJCRC-KC, (n) JCRC-KE, and (o) NJCRC-KE. In addition, the legend, scale bar, and north arrow of this image are shown in the last row. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
The optimal combination of parameters for the nine ℓ_2-norm based classifiers with the Indian Pines dataset.

	Linear version	Chi-squared distance (KC)	Euclidean distance (KE)
CRC	$\lambda = 1e-5$	$\lambda = 1e-5$	$\lambda = 1$ and $\sigma = 0.0330$
JCRC	$\lambda = 1e-7$, $T = 81$	$\lambda = 1e-7$ and $T = 49$	$\lambda = 1$, $T = 81$, and $\sigma = 0.0133$
NJCRC	$\lambda = 1e-7$, $T = 121$, and $K = 80$	$\lambda = 1e-7$, $T = 121$, and $K = 70$	$\lambda = 1e-1$, $T = 121$, $K = 85$, and $\sigma = 0.0133$

Table 5
Classification accuracy (%) for the Indian Pines image with the test set.

C	SRC	SVM	CRC	CRC-KC	CRC-KE	SVM-NS	JCRC	JCRC-KC	JCRC-KE	NJCRC	NJCRC-KC	NJCRC-KE
1	0.4942	0.6923	0.6689	0.6557	0.4466	0.8085	0.8852	0.9101	0.9145	0.9393	0.8882	**0.9444**
2	0.5468	0.7792	0.5481	0.7455	0.5974	0.9104	0.8727	0.9792	0.9649	0.9416	**0.9948**	0.9805
3	0.8487	0.9125	0.8747	0.8983	0.8865	0.9456	0.9149	0.9598	**0.9698**	0.9173	0.9456	0.9551
4	0.9463	0.9418	0.9358	0.9612	0.9612	0.9910	0.9925	0.9985	0.9970	0.9970	0.9910	**1**
5	0.9952	0.9952	0.9952	0.9952	0.9952	**1**	**1**	**1**	**1**	**1**	**1**	**1**
6	0.6371	0.716	0.6689	0.7467	0.7215	0.9046	0.9441	0.9452	0.9497	**0.9748**	0.9276	0.9561
7	0.5194	0.5908	0.4472	0.6259	0.6171	0.7265	0.7061	0.7311	0.8939	0.7779	0.8342	**0.9357**
8	0.6323	0.7992	0.7148	0.8330	0.5760	0.8780	0.9250	0.9887	0.9887	0.9681	0.9306	**0.9944**
9	0.8722	0.9245	0.8863	0.9535	0.8365	0.961	0.9336	0.9710	**1**	0.9743	0.9876	**1**
10	0.7209	0.7178	0.681	0.7117	0.6227	0.9141	0.954	0.9939	**1**	**1**	0.9969	**1**
OA	0.6601	0.7563	0.6765	0.7667	0.6829	0.8623	0.8685	0.9009	0.9478	0.9149	0.9222	**0.9660**
A	0.27	0.17	0.26	0.17	0.24	0.07	0.05	0.03	**0.02**	**0.02**	**0.02**	**0.02**
Q	0.05	0.05	0.05	0.05	0.05	0.05	0.09	0.07	0.03	0.05	0.05	**0.02**

Table 6
The optimal combination of parameters for the nine ℓ_2-norm based classifiers with the Pavia University image.

	Linear version	Chi-squared distance (KC)	Euclidean distance (KE)
CRC	$\lambda = 1e-4$	$\lambda = 1e-5$	$\lambda = 1$ and $\sigma = 0.0191$
JCRC	$\lambda = 1e-4$, $T = 49$	$\lambda = 1e-6$ and $T = 9$	$\lambda = 1e-1$, $T = 121$, and $\sigma = 0.0476$
NJCRC	$\lambda = 1e-5$, $T = 81$, and $K = 55$	$\lambda = 1e-7$, $T = 81$, and $K = 50$	$\lambda = 1e-1$, $T = 225$, $K = 180$, and $\sigma = 0.0476$

Table 7
Classification accuracy for the Pavia University image with the test set.

C	SRC	SVM	CRC	CRC-KC	CRC-KE	SVM-NS	JCRC	JCRC-KC	JCRC-KE	NJCRC	NJCRC-KC	NJCRC-KE
1	0.5743	0.7161	0.3995	0.7151	0.5759	0.9147	0.6973	0.8877	0.8012	0.7527	**0.9436**	0.8304
2	0.7256	0.7968	0.7724	0.7779	0.7279	0.9355	0.8917	0.8768	**0.9726**	0.8747	0.9236	0.9715
3	0.6595	0.7411	0.8324	0.8222	0.6824	0.9296	0.9213	0.9087	0.7776	**0.9903**	0.9354	0.7712
4	0.9137	0.9312	0.9319	0.9382	0.9160	**0.9841**	0.9246	0.9659	0.9130	0.9362	0.9190	0.9196
5	0.9946	0.9946	0.9992	0.9977	0.9946	0.9969	0.9724	**1**	**1**	0.9992	**1**	**1**
6	0.6274	0.7198	0.5256	0.7482	0.6312	0.9373	0.7462	0.8721	0.9455	0.6853	0.9431	**0.9561**
7	0.8566	0.8829	0.7426	0.8767	0.8736	0.9124	0.9349	0.9233	0.9240	**0.9806**	**0.9806**	0.9349
8	0.6068	0.7751	0.1571	0.5681	0.6669	0.8861	0.0203	0.7133	0.9333	0.1483	0.9097	**0.9454**
9	0.9735	0.9945	0.8820	0.9151	0.9735	0.9912	**1**	0.9338	0.4609	0.9713	0.2249	0.4068
OA	0.7081	0.7932	0.6553	0.7729	0.7168	**0.9341**	0.7795	0.8782	0.9141	0.7902	0.9172	0.9198
A	0.17	0.12	0.25	0.13	0.16	0.04	0.11	0.07	0.04	0.10	0.04	**0.03**
Q	0.06	0.05	0.05	0.06	0.07	**0.02**	0.07	0.04	0.04	0.07	0.03	0.04

and $K = 200$. The neighborhood window size T for the spatially extended classifiers ranges from 9 to 289, and the kernel parameter σ for the KE-related algorithms ranges from 0.015 to 0.1. The optimal combinations of parameters for the nine ℓ_2-norm based classifiers are shown in Table 6. In addition, the parameters for SVM, SRC, and JSRC are set as the corresponding optimal, and the optimal size of the spatial neighborhood in JSRC is 9.

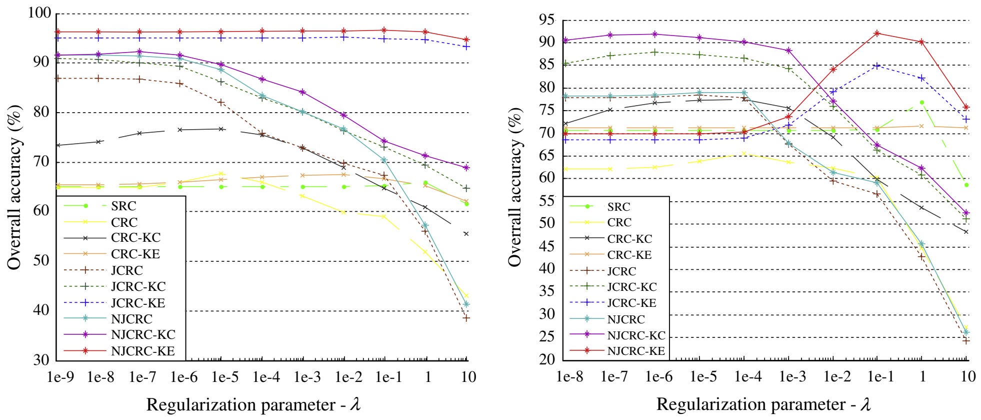
The classification results for the various different classifiers are visually displayed in Fig. 2(c–o), respectively. The quantitative evaluation results, which include the classification accuracy for every class, the overall accuracy, quantity disagreement, and the allocation disagreement, are shown in Table 7. To allow a comparison, we also list the detailed improvements for the nonlinear CR-based classifiers over the linear CRC in Table 9, which is further analyzed in the following subsection. For the Pavia dataset with a high spatial resolution, the ranges of the numbers of the nonlocal joint hyperspectral signals suggest an obstacle caused by increased internal spectral/feature variability of each land-cover type on the joint collaboration model. Comparing the SVM-related algorithms with the ℓ_2-norm related ones, it can be observed that SVM-NS

Table 8
Summary of the classification comparisons undertaken with the Indian Pines dataset. A resampling method was used to conduct the McNemar's test to compare the proportions of the correctly allocated pixels. All tests shown were one-sided, and a 5% level of significance was selected.

Classifier1	Classifier2	Comparison of the proportions and disagreement				
		ΔQ	ΔA	ΔOA	$\lvert z\rvert$	Significant?
NJCRC-KE	CRC	−0.24	−0.03	0.2895	7.57	Yes
NJCRC-KC	CRC	−0.24	0	0.2456	7.23	Yes
NJCRC	CRC	−0.24	0	0.2384	6.68	Yes
JCRC-KE	CRC	−0.24	−0.02	0.2713	7.57	Yes
JCRC-KC	CRC	−0.23	0.02	0.2244	6.90	Yes
JCRC	CRC	−0.21	0.04	0.1920	6.25	Yes
CRC-KE	CRC	−0.02	0	0.0064	1.26	No
CRC-KC	CRC	−0.09	0	0.0902	3.32	No
SVM	CRC	−0.09	0	0.0798	2.78	No
SRC	CRC	0.01	0	−0.0164	0.66	No

Table 9
Summary of the classification comparisons undertaken with the Pavia University image. A resampling method was used to conduct the McNemar's test to compare the proportions of the correctly allocated pixels. All tests shown were one-sided, and a 5% level of significance was selected.

Classifier1	Classifier2	Comparison of the proportions and disagreement				
		ΔQ	ΔA	ΔOA	$\lvert z\rvert$	Significant?
NJCRC-KE	CRC	−0.22	−0.01	0.2645	5.74	Yes
NJCRC-KC	CRC	−0.21	−0.02	0.2619	5.74	Yes
NJCRC	CRC	−0.15	0.02	0.1349	2.95	No
JCRC-KE	CRC	−0.21	−0.01	0.2588	6.23	Yes
JCRC-KC	CRC	−0.18	−0.01	0.2229	6.23	Yes
JCRC	CRC	−0.14	0.02	0.1242	2.24	No
CRC-KE	CRC	−0.09	0.02	0.0615	0.98	No
CRC-KC	CRC	−0.12	0.01	0.1176	2.77	No
SVM	CRC	−0.13	0	0.1379	1.99	No
SRC	CRC	−0.08	0.01	0.0528	0.83	No



(a)

(b)

Fig. 3. The classification accuracy versus the regularization parameter λ for the various classification algorithms: (a) the Indian Pines image, and (b) the Pavia University image.

yields the best overall performance. Note that in the training stage of SVM-NS, in order to extract the spatial feature for each training sample, SVM-NS requires knowledge of the neighboring pixels, which may not be available in the training set. Therefore, we could say that SVM-NS uses more training samples than the other methods, especially in our experiment setting, where the training sets are randomly selected.

We next demonstrate the impact of the kernel approach on the classification result of the Pavia University dataset. For the single-signal algorithms, CRC-KC shows the best performance, while NJCRC-KE shows the best performance for the multiple-signal simultaneous representation category. Although the performance in several classes shows that the original spectral feature can reach a more accurate recognition rate (such as the shadows class in the Pavia University image), it can still be concluded that the kernel strategy can improve the classification accuracy in most classes.

4.3. Parameter analysis and performance discussion

4.3.1. Parameter analysis

In this subsection, we examine the effect of the parameters on the classification performance of the various algorithms with the

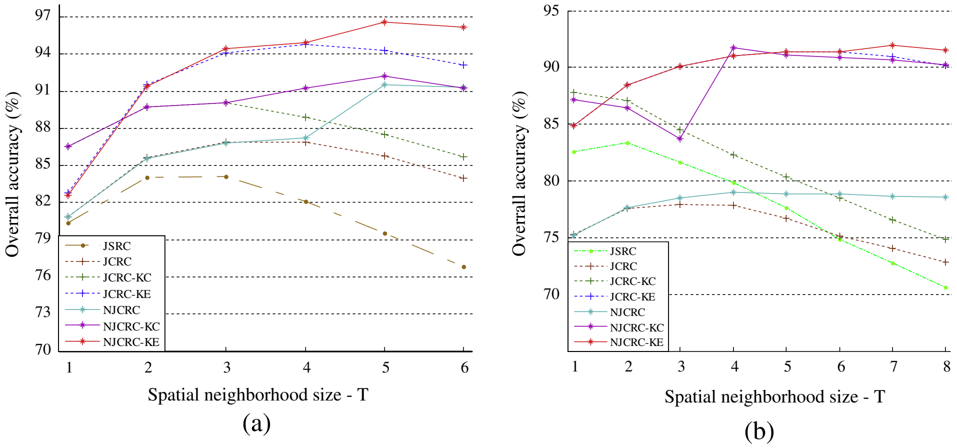


Fig. 4. The classification accuracy versus the spatial neighborhood size *T* for the various classification algorithms: (a) the Indian Pines image, and (b) the Pavia University image.

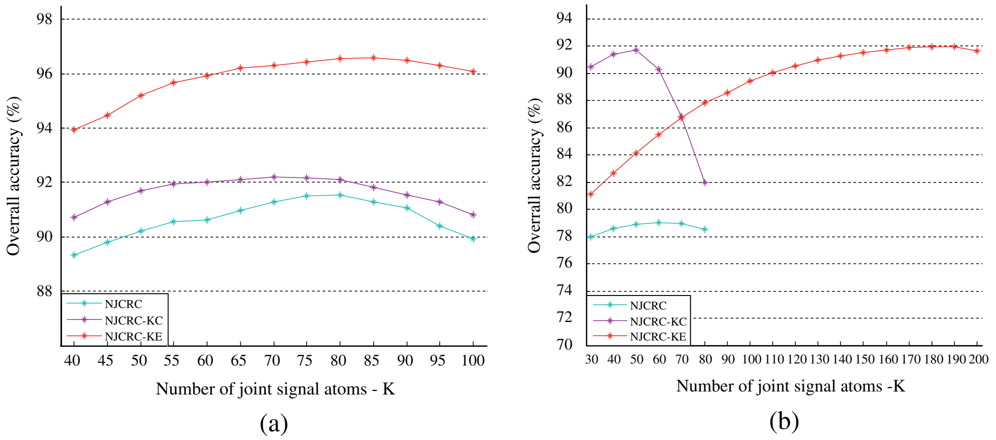


Fig. 5. The classification accuracy versus the number of joint signal atoms *K* for the various classification algorithms: (a) the Indian Pines image, and (b) the Pavia University image.

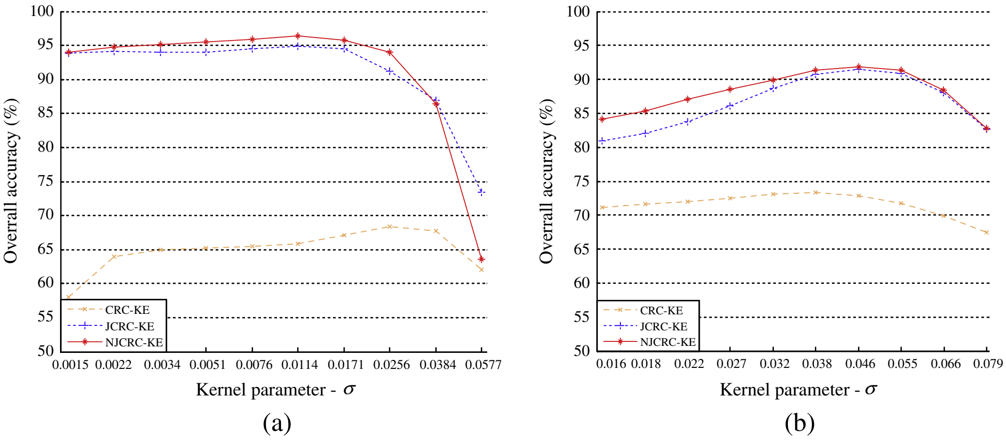


Fig. 6. The classification accuracy versus the kernel width parameter σ for the various classification algorithms: (a) the Indian Pines image; (b) the Pavia University image.

Indian Pines image and the Pavia University image, respectively. When analyzing one specific parameter, we fix the other parameters as the corresponding optimal.

4.3.1.1. Effect of the regularization parameter. Fig. 3(a and b) record the effect of the regularization parameter λ on the two datasets, respectively. For the Indian Pines image, SRC, CRC-KE, JCRC-KE, and NJCRC-KE show a robust performance, and NJCRC-KE is the best one among all the algorithms. With the increase in λ, the performances of the KC-related algorithms reaches the optimal at first and then declines when λ exceeds a certain threshold. It can also be seen that the linear versions show the worst performances. For the Pavia University image, the kernel algorithms with the Euclidean distance show a weaker capability with a low value,

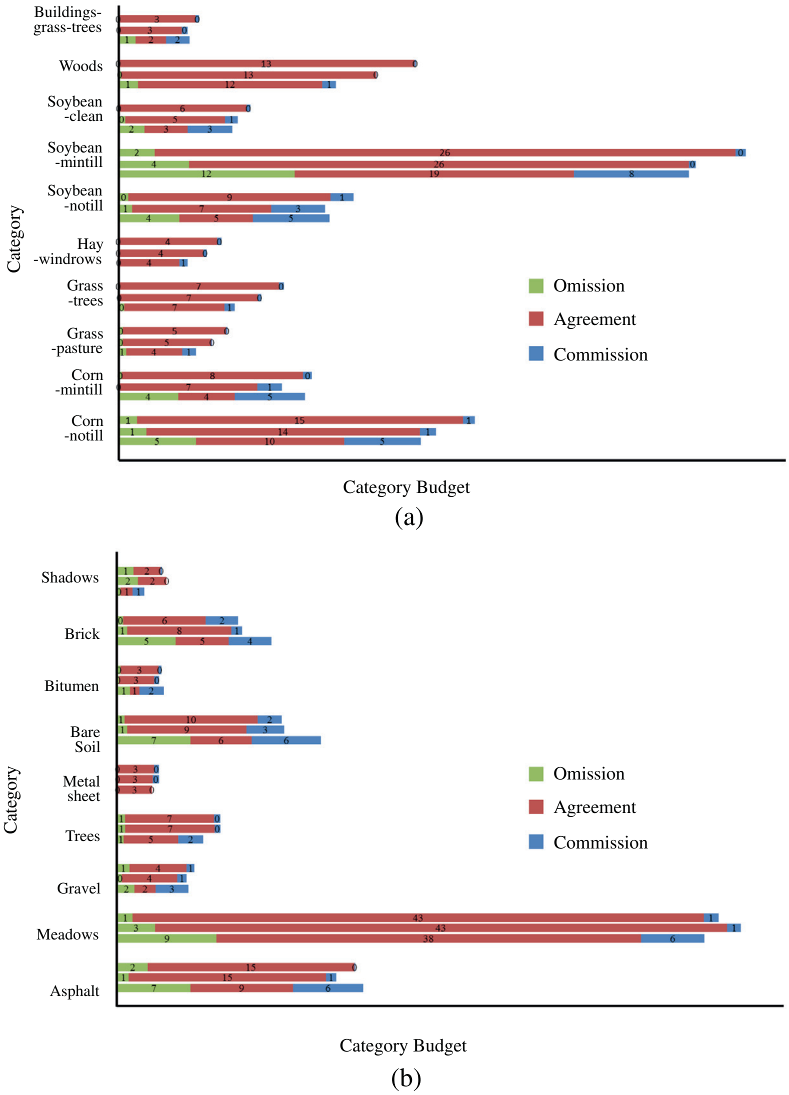


Fig. 7. Classification omission, agreement, and commission for each category over the three classifiers. For each class, the upper bar in the set is for NJCRC-KE, the middle bar is associated with NJCRC-KC, and the bottom one refers to CRC. Furthermore, the number in each bar is the associated proportion of the whole test dataset. (a) The Indian Pines image, and (b) the Pavia University image.

and then increase rapidly with a suitable λ. In addition, the KC-related algorithms and linear versions also show similar performances for the Indian Pines image.

4.3.1.2. Effect of the spatial neighborhood size and the number of joint signal atoms. The joint spatial aspects of the final proposed algorithms contain two parameters: the neighborhood size T and the number of the nonlocal signals K. For the Indian Pines image, we fix the other parameters as the optimal, and we show the plots of the classification results versus the two parameters in Figs. 4(a) and 5(a), respectively. Both plots rise quickly and reach a maximum point, and then remain relatively stable with only a tiny decline, which shows the robustness of the proposed algorithms with respect to the two parameters. For the Pavia University image, the optimal parameter settings for the various algorithms can be seen in Table 6. The plot for JCRC-KE is quite similar to that for NJCRC-KE in Fig. 4(b) by setting $K = 180$. Without the nonlocal signal selection approach, the rest of the spatial information involved algorithms show a rapid decrease as the spatial neighborhood size increases. In Fig. 5(b), we fix $T = 81$ for NJCRC and NJCRC-KC, and $T = 225$ for NJCRC-KE, to demonstrate the effects of the number of nonlocal signals on these three algorithms. The plots in Fig. 5(b) suggest that the preferred K for NJCRC and NJCRC-KC are close to each other, while the performance of NJCRC-KC is much better than that of the algorithms working in the original spectral space. Under a large spatial window, the plot for NJCRC-KE rises quickly and reaches a maximum point, and then stays quite stable. It can be concluded that the suitable parameter settings for the KC-related algorithms are close to those for the corresponding linear versions, while the settings for the KE-related algorithms are quite different.

4.3.1.3. Effect of the kernel width parameter. We next investigate the effect of the kernel width parameter σ on the KE-related classifiers. Fig. 6(a and b) show the performances with the Indian Pines image and the Pavia University image, respectively. It can be observed that the performance of the proposed NJCRC-KE method is quite robust and stable with regard to the kernel parameter σ.

4.3.2. Performance discussion

In this subsection, we analyze the classification performances of some of the related classifiers. To allow a comparison, we first list the detailed improvements for the nonlinear CR-based classifiers over classical CRC for the Indian Pines image in Table 8, and those for the Pavia University image in Table 9, respectively. The ΔQ values in these tables are the difference in the quantity disagreement between the practical Classifer1 and Classifier2, which indicates the superiority of the classifier when $\Delta Q < 0$. The next two indexes ΔA and ΔOA also tell a similar story, and are associated with the specific score area. The McNemar' test (Foody, 2004), which is a non-parametric statistical significance test of the difference between two classifications, is also utilized in this paper. Compared with the linear CRC, most of the kernel-based algorithms with spatial information can acquire superior classification performances, and show a significant superiority with both datasets. For the vector-oriented representation algorithms, CRC-KC is superior to the linear classifiers, and shows comparable performance to the SVM with RBF kernel.

We finally analyze the classification performance with each category of the two hyperspectral datasets. The two proposed kernel-based algorithms and classical CRC are utilized to make a comparison, as shown in Fig. 7. The classification omission, agreement, and commission are shown as the sub-bars for each category in a group, and the detailed number in each sub-bar denotes the associated proportion. In Fig. 7(a), the performance for each category is improved with the kernel and nonlocal spatial constraint based algorithms, and the three soybean-related categories show a significant improvement, considering all three of the evaluation criteria. For the experiments with the Pavia University dataset, a similar observation can be made. Pixels belonging to the asphalt, meadow, bare soil, and brick classes dominate the improvement. To sum up, it is demonstrated in the experiments that the kernel and nonlocal spatial constraint based algorithms can significantly improve the classification result, for both the classical agricultural AVIRIS image and the urban ROSIS image.

5. Conclusions

In this paper, we propose a new HSI classification technique based on collaborative representation in a nonlinear feature space induced by a column-generation kernel method. For the proposed algorithms, we first map the spectral signal into the high-dimensional feature space, and we then utilize a nonlocal joint collaborative regression model for the kernel signal reconstruction and the subsequent pixel classification. After the explicit kernel dictionary and explicit kernel signal are obtained, a standard linear regression model can be directly extended to a kernel version. The nonlocal contextual information is incorporated to constrain the dominant representation through the joint collaborative representation. The kernel technique in this paper differs from the conventional kernel mapping in the RKHS feature space. The column generation directly treats the similarity measures between spectral pixels as a feature, while the conventional kernel method replaces the original feature vector as an implicit kernel feature by the inner product operation. We also focus on the absolute/relative difference between pixel pairs, and we apply two specific RBF kernel functions to investigate the efficient performance of this kernel technique. The proposed algorithms were tested on AVIRIS and ROSIS hyperspectral datasets, and the extensive experimental results confirm the effectiveness of the nonlinear strategy.

We should mention that the proposed algorithms still have room for improvement, such as better contextual information extraction, which could automatically obtain the joint signal matrix, adaptively fix the kernel width parameter for the KE-related algorithms, or utilize a dictionary learning method to reduce the size of dictionary while keeping its representative and discriminative ability. We will focus on these issues in our future work.

Acknowledgements

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB707105, and by the National Natural Science Foundation of China under Grant 61201342. The authors would like to thank Prof. D. Landgrebe from Purdue University for providing the AVIRIS image of Indian Pines and Prof. Gamba from University of Pavia for providing the ROSIS dataset. Thanks also to the handling editor and anonymous reviewers for their careful reading and helpful remarks.

References

Bi, J., Zhang, T., Bennett, K.P., 2004. Column-generation boosting methods for mixture of kernels. In: Proc. 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 22–25 August, Seattle, WA, pp. 521–526.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proc. 5th Annual Workshop on Computational Learning Theory, 27–29 July, Pittsburgh, PA, pp. 144–152.

Chen, Y., Nasrabadi, N.M., Tran, T.D., 2011. Hyperspectral image classification using dictionary-based sparse representation. IEEE Trans. Geosci. Remote Sens. 49 (10), 3973–3985.

Chen, Y., Nasrabadi, N.M., Tran, T.D., 2013. Hyperspectral image classification via kernel sparse representation. IEEE Trans. Geosci. Remote Sens. 51 (1), 217–231.

Demir, B., Erturk, S., 2010. Empirical mode decomposition of hyperspectral images for support vector machine classification. IEEE Trans. Geosci. Remote Sens. 48 (11), 4071–4084.

Dópido, I., Villa, A., Plaza, A., Gamba, P., 2012. A quantitative and comparative assessment of unmixing-based feature extraction techniques for hyperspectral image classification. IEEE J. Selected Topics Appl. Earth Observations Remote Sens. 5 (2), 421–435.

Duarte, M.F., Sarvotham, S., Baron, D., Wakin, M.B., Baraniuk, R.G., 2005. Distributed compressed sensing of jointly sparse signals. In: Proc. 39th Asilomar Conference on Signals, Systems and Computers, 30 October–2 November, Pacific Grove, CA, pp. 1537–1541.

Fauvel, M., Benediktsson, J.A., Chanussot, J., Sveinsson, J.R., 2008. Spectral and spatial classification of hyperspectral data using SVMS and morphological profiles. IEEE Trans. Geosci. Remote Sens. 46 (11), 3804–3814.

Foody, G.M., 2004. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. Photogr. Eng. Remote Sens. 70 (5), 627–634.

Gehler, P., Nowozin, S., 2009. On feature combination for multiclass object classification. In: Proc. 12th International Conference on Computer Vision, 27 September–4 October, Kyoto, pp. 221–228.

Gualtieri, J.A., Cromp, R.F., 1999. Support vector machines for hyperspectral remote sensing classification. In: Proc. 27th AIPR Workshop on Advances in Computer-Assisted Recognition, 14–16 October, Washington, DC, pp. 221–232.

Keller, J.M., Gray, M.R., Givens, J.A., 1985. A fuzzy *k*-nearest neighbor algorithm. IEEE Trans. Syst. Man Cyber. 15 (4), 580–585.

Kwon, H., Nasrabadi, N.M., 2006. Kernel matched subspace detectors for hyperspectral target detection. IEEE Trans. Pattern Anal. Mach. Intell. 28 (2), 178–194.

Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conference on Machine Learning, 28 June–1 July, Williams College, Williamstown, pp. 282–289.

Leslie, C., Eskin, E., Noble, W.S., 2002. The spectrum kernel: A string kernel for SVM protein classification. In: Proc. 7th Pacific Symposium on Biocomputing, 3–7 January, Lihue, Hawaii, pp. 566–575.

Li, J., Bioucas-Dias, J.M., Plaza, A., 2012. Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields. IEEE Trans. Geosci. Remote Sens. 50 (3), 809–823.

Li, J., Zhang, H., Huang, Y., Zhang, L., 2013a. Hyperspectral image classification by nonlocal joint collaborative representation with locality-adaptive dictionary. IEEE Geosci. Remote Sens. Lett. 52 (6), 3707–3719.

Li, J., Zhang, H., Zhang, 2013b. Supervised segmentation of very high resolution images by the use of extended morphological attribute profiles and a sparse transform. IEEE Geosci. Remote Sens. Lett. 11 (8), 1409–1413.

Li, J., Zhang, H., Zhang, L., Huang, X., Zhang, L., 2013c. Joint collaborative representation with multi-task learning for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. http://dx.doi.org/10.1109/TGRS.2013.2293732.

Li, J., Marpu, P.R., Plaza, A., Bioucas-Dias, J.M., Benediktsson, J.A., 2013d. Generalized composite kernel framework for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 51 (9), 4816–4829.

Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sens. 42 (8), 1778–1790.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT press, Cambridge.

Nash, S., Sofer, A., 1996. Linear and Nonlinear Programming. McGraw–Hill, New York.

Nidamanuri, R.R., Zbell, B., 2011. Use of field reflectance data for crop mapping using airborne hyperspectral image. ISPRS J. Photogr. Remote Sens. 66 (5), 683–691.

Philippe, V.J., Koji, T., Bernhard, S., 2004. A primer on kernel methods. Kernel Methods Comput. Biol., 35–70.

Plaza, A., Benediktsson, J.A., Boardman, J.A., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., 2009. Recent advances in techniques for hyperspectral image processing. Remote Sens. Environ. 113, S110–S122.

Pontius Jr., R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int. J. Remote Sens. 32 (15), 4407–4429.

Qian, Y., Ye, M., Zhou, J., 2012. Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features. IEEE Trans. Geosci. Remote Sens. 51 (4), 2276–2291.

Senthil, K.A., Keerthi, V., Manjunath, A., Werff, H.V.D., Meer, F.V.D., 2010. Hyperspectral image classification by a variable interval spectral average and spectral curve matching combined algorithm. Int. J. Appl. Earth Obs. Geoinf. 12, 261–269.

Suykens, J.A., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9 (3), 293–300.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc.: Ser. B (Stat. Method.) 73 (3), 273–282.

Tropp, J.A., Gilbert, A.C., Strauss, M.J., 2006. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. Signal Process. 86 (3), 572–588.

Tsang, I.W., Kwok, J.T., Cheung, P.M., 2006. Core vector machines: fast SVM training on very large data sets. J. Mach. Learn. Res. 6 (4), 363–392.

Vapnik, V., 1999. The Nature of Statistical Learning Theory. Springer, Berlin.

Waqas, J., Yi, Z., Zhang, L., 2012. Collaborative neighbor representation based classification using l2-minimization approach. Pattern Recogn. Lett. 34 (2), 201–208.

Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31 (2), 210–227.

Yang, J., Yu, K., Gong, Y., T. Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification. In: Proc. 22nd IEEE Conference on Computer Vision and Pattern Recognition, 20–25 June, Miami, pp. 1794–1801.

Yang, M., Zhang, L., Zhang, D., Wang, S., 2012. Relaxed collaborative representation for pattern classification. In: Proc. 25th IEEE Conference on Computer Vision and Pattern Recognition, Providence, 16–21 June, Rhode Island, pp. 2224–2231.

Yuan, X., Liu, X., Yan, S., 2012. Visual classification with multi-task joint sparse representation. IEEE Trans. Image Process. 21 (10), 4349–4360.

Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C., 2007. Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vision 73 (2), 213–238.

Zhang, H., Shen, H., Zhang, L., 2012a. A super-resolution reconstruction algorithm for hyperspectral images. Signal Process. 92 (9), 2082–2096.

Zhang, L., Yang, M., Feng, X., Ma, Y., Zhang, D., 2012b. Collaborative representation based classification for face recognition. Arxiv Preprint 1204, 2358.

Zhang, H., Li, J., Huang, Y., Zhang, L., 2013. A nonlocal weighted joint sparse representation classification method for hyperspectral imagery. IEEE J. Selected Topics Appl. Earth Observations Remote Sens. http://dx.doi.org/10.1109/JSTARS.2013. 2264720.

Zheng, X., Sun, X., Fu, K., Wang, H., 2013. Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint. IEEE Geosci. Remote Sens. Lett. 10 (4), 652–656.

Zhong, P., Wang, R., 2010. Learning conditional random fields for classification of hyperspectral images. IEEE Trans. Image Process. 19 (7), 1890–1907.