Dual-Clustering-Based Hyperspectral Band Selection by Contextual Analysis

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

***Abstract*—Hyperspectral image (HSI) involves vast quantities of information that can help with the image analysis. However, this information has sometimes been proved to be redundant, considering specific applications such as HSI classification and anomaly detection. To address this problem, hyperspectral band selection is viewed as an effective dimensionality reduction method that can remove the redundant components of HSI. Various HSI band selection methods have been proposed recently, and the clustering-based method is a traditional one. This agglomerative method has been considered simple and straightforward, while the performance is generally inferior to the state of the art. To tackle the inherent drawbacks of the clustering-based band selection method, a new framework concerning on dual clustering is proposed in this paper. The main contribution can be concluded as follows: 1) a novel descriptor that reveals the context of HSI efficiently; 2) a dual clustering method that includes the contextual information in the clustering process; 3) a new strategy that selects the cluster representatives jointly considering the mutual effects of each cluster. Experimental results on three real-world HSIs verify the noticeable accuracy of the proposed method, with regard to the HSI classification application. The main comparison has been conducted among several recent clustering-based band selection methods and constraint-based band selection methods, demonstrating the superiority of the technique that we present.**

***Index Terms*—Band selection, context, dual clustering, hyperspectral angle, hyperspectral image (HSI).**

I. Introduction

HYPERSPECTRAL images (HSI) contain rich discriminative physical clues that come from the narrow continuous spectral bands. Each of these bands reflects some specific characteristics that are closely related to the property of the target. Therefore, a wide range of real-world applications is benefited, based on the discriminative attribute of HSI, such as biological analysis [1], product quality inspection [2], and medical imaging [3].

However, a problem also exists in that the huge volume of image information contained in the HSI is not easy to tackle with, particularly for the case that little labeled information is included. This drawback brings a heavy burden to HSI classification or segmentation [4]. Moreover, the neighboring bands of HSI are of high correlation, and they are not as discriminative as we expected, which means that redundant information is contained. This will introduce problems concerning the computational complexity and, at the same time, affect the following classification or segmentation process owing to the "curse of dimensionality" [5]. From this point of view, dimensionality reduction is necessary.

Existing dimension reduction methods can be divided into two main branches. The first branch is *feature extraction* [6], [7]. This typical method projects the initial HSI information to a lower dimensional space, leading to a more abstract representation [8]. The representatives of this branch include wavelet transform (WT) [9], principal component analysis [10], linear discriminant analysis [11], and independent component analysis [12]. Although these methods can produce satisfying results, they are not always the most appropriate choice for two inherent drawbacks. The first is that these feature extraction methods have to consider and deal with the transformation of the whole data volume to extract new features, which is with high time complexity. The second is that some crucial information is distorted due to the destruction of band correlation in the HSI data transforming process [13], causing loss of physical meaning and interpretation of HSI.

Compared with feature extraction, the other branch is well known as *feature selection*, which is with apparent advantage [14], [15]. The aim is to cannibalize the most informative and distinctive HSI bands to construct a subset. These desired candidates should be the ones with the most critical factors. The ultimate selected fewer decisive bands should represent the whole image with no loss of effectiveness [16].

As an important topic in HSI analysis, hyperspectral band selection has recently attracted the attention of researchers. A reliable band selection process not only facilitates the HSI identification [17], transmission [18], and detection [19] but also increases the efficiency of HSI analysis [20]. This paper mainly concentrates on the classification application facilitated by the band selection [20], [21].

Various methods have been proposed to support the process of HSI band selection, and three main groups of methods can be summarized: constraint-based [22]–[24], clustering-based [13], [25]–[27], and sparse-based methods [4], [28]. The first type

Manuscript received November 13, 2014; revised April 9, 2015 and July 13, 2015; accepted September 17, 2015. Date of publication October 9, 2015; date of current version February 24, 2016. This work was supported by the National Basic Research Program of China (Youth 973 Program) under Grant 2013CB336500; by the State Key Program of National Natural Science of China under Grant 61232010; by the National Natural Science Foundation of China under Grant 61172143, Grant 61105012, and Grant 61379094; by the Natural Science Foundation Research Project of Shaanxi Province under Grant 2015JM6264; by the Fundamental Research Funds for the Central Universities under Grant 3102014JC02020G07; and by the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences. *(Corresponding author: Qi Wang.)*

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

Q. Wang is with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@nwpu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2015.2480866

is by imposing a constraint such as dependence minimization to achieve the selected bands. The representative work includes *constrained energy minimization* (CEM) and *linearly constrained minimum variance* (LCMV) [22], [29], [30]. Both methods linearly constrain the target bands by minimizing the interfering effect brought by the other bands. However, this kind of method is only based on the consideration of band correlation, regardless of the representation of the initial HSI cube. The clustering-based method is only complementary to the former one. It can be generally summarized into two steps [13], [31]. First, group the bands into clusters, in which the intracluster variance is minimized and the intercluster variance is maximized. Second, choose the bands with the highest average correlations from their corresponding clusters as the final output. However, this rough traditional strategy, which only focuses on the raw spectral features of the HSI cube, barely digs for the contextual clue of the HSI pixel, and constraints among bands are not taken into serious consideration. A popular kind of method that has appeared in recent works is the sparsity-based method [4], [28]. It takes the desired bands as the dictionary. Through adjusting the expressive coefficient, the initial HSI cube is reconstructed by the dictionary. This joint selection process improves the correlation of the selected bands. However, the representativeness of these selected bands still needs to be enhanced, and the efficiency is also a vital problem, as compared with the former clustering-based methods.

We believe that the clustering-based method with high efficiency can be further exploited, if the mutual affection among the representatives of the clusters is taken into consideration. To be more specific, the selected bands should be treated as a whole rather than as independent ones [25]. This assumption gives a hint to the potential improvement of the traditional clustering framework. In this paper, we propose a new framework named *dual-clustering-based band selection by context analysis* (DCCA). The main contributions are as follows:

- Consider the context information of HSI bands in the process of dual clustering. The context of a specific element in HSI contains both the neighboring bands and its neighboring pixels. We include the context information of HSI bands into the band clustering process and convert the band selection to a *dual clustering* problem [32]. In our clustering process, the contexts of HSI and the raw HSI are grouped simultaneously. Then, the two results will influence each other through the dual clustering principle. As far as we are concerned, utilizing context to enhance the unsupervised raw band clustering has never been proposed before.
- Design a new *pairwise hyperspectral angle* (PHA) descriptor for HSI. Hyperspectral angle is one of the means that measure the similarity between two neighboring pixels [33], [34]. The two sides of this angle are constructed by vectors representing the spectra of the two pixels. In this paper, we introduce two new kinds of hyperspectral angles named PHA, to exploit the context information of a specific pixel. Each pairwise angle appears simultaneously, i.e., one for neighboring bands and the other for neighboring pixels. The two descriptors explore the different aspects of HSI and act as a complementary to each other.

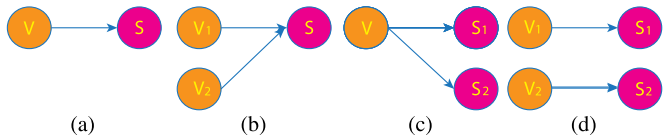

Fig. 1. Illustration for the relation between observable views and labeling solutions in the clustering process. (a) One view against one solution. (b) Two views against one solution. (c) One view against two solutions. (d) Two views against two solutions.

- Propose a *groupwise strategy for representation* (GSR) of clusters. In traditional clustering-based band selection methods, every cluster is treated as an independent one, and the selected representatives (bands) of each cluster have no relation to each other. We hold the view that the chosen representatives should be viewed as a whole to realize a better representation of the original data, and a *joint Euclidean distance* framework is introduced to solve this problem.

The remainder of this paper is organized as follows: Section II reviews the previous works for hyperspectral band selection. Section III presents the detailed component descriptions of the proposed framework. Section IV verifies the superiority of the proposed framework by experiments and comparisons. Finally, we conclude the work in Section V.

II. Related Work

Band selection is an effective dimensionality reduction method [35] for HSI. The main purpose of this technique is to summarize the most critical information of HSI, which can, at the same time, relieve the computational burden of the following data analysis and processing. As for the process of band selection, existing methods can be roughly divided into the supervised and the unsupervised. The existence of training phase makes these two differ from each other, discriminatively. Due to this difference, unsupervised methods tend to be more practical in real applications, for the reason that training samples do not always exist.

This paper mainly focuses on the research of unsupervised method. Existing unsupervised methods are mainly based on basic techniques, such as PCA [36], discrete WT [37], and ICA [38]. The ultimate goal of these methods is to find out the most discriminative bands to construct the best representative subset of the original HSI. However, the results of these methods are still far from satisfying.

The clustering-based method is also one of the main branches for unsupervised band selection [13]. Existing clustering methods can be sorted into four categories, according to the different number of observable views and latent categories [32], as shown in Fig. 1. The most popular one is the ordinary clustering illustrated in Fig. 1(a), such as k-means and ISOdata. These methods only consider one view of the data set, and the solution is single. Recent advance has come from two perspectives. The first is the additional observable views of the data set, as shown in Fig. 1(b), leading to the multiview-based clustering [39], [40] method employing multiple views of the original data, instead of one view, to get better performance. There is only

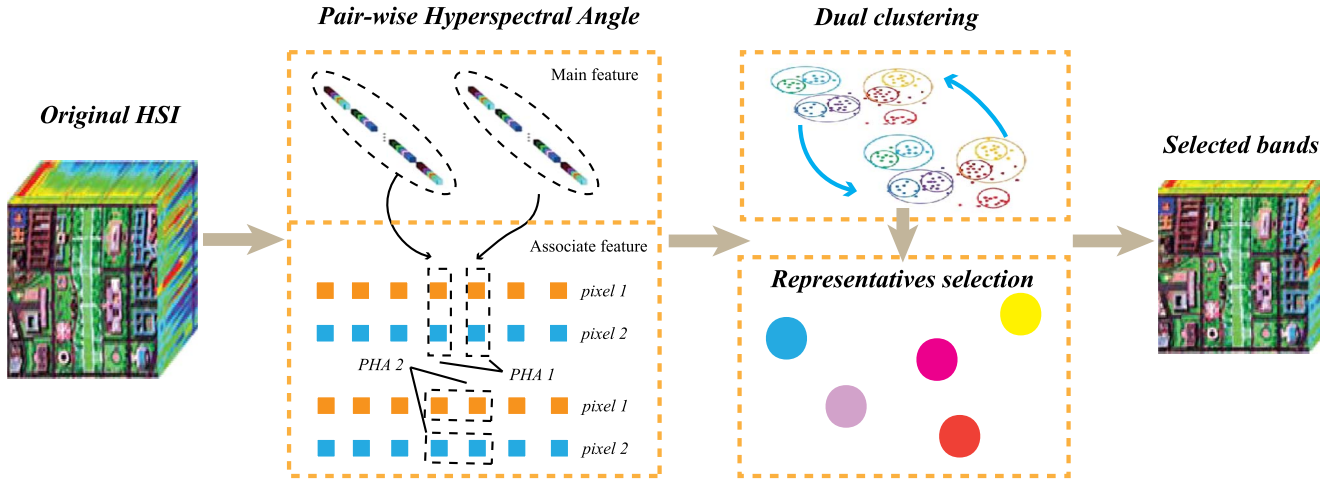

Fig. 2. HSI band selection pipeline. For an input HSI, the first step is to extract the main raw feature and associate PHA feature, and these two features include both the spatial relation and the spectral attribute of HSI. Then, a dual clustering framework is constructed based on these two features, and the final clustering result is obtained through the mutual effect of the two features. After that, a groupwise strategy is utilized for representation of clusters. The chosen representatives are taken as the selected bands, finally.

one true clustering result obtained by the effort of both views and the mutual information (MI) of them. Another well-known method results in two clustering solutions based on only one view, as shown in Fig. 1(c). This alternative clustering method [41] gets multiple clustering results, in the process of which both the individual clusterings benefit each other. The last one in Fig. 1(d) is dual assignment clustering [32], [42]. In this method, each clustering comes from one view, and the two processes act interactively on each other, to find the optimal solution of both clusterings.

As far as HSI is concerned, clustering-based band selection can be conducted, according to various views. These views can also be regarded as the features of HSI. Different features of HSI, such as spectral gradient feature [43], gabor texture feature [44], and shape feature [45], will result in different clustering solutions. The result is not only closely related with the feature we choose but also depends on the interaction among the selected features.

One significant feature of HSI is the spectral angle, which reflects the spatial relation of neighboring pixels in HSI. It can be viewed as an effective supplement to the raw feature. The two sides of this angle is constructed by vectors representing the spectra of pixels. Therefore, this angle can effectively measure the similarity of two neighboring pixels. A smaller angle means higher correlation of the two pixels. However, the construction of the spectral angle is limited to the utilization of the whole bands for a specific pixel. The global spectral information is not necessarily the best. Motivated by this point, we aim to explore the local contextual information of an HSI pixel and put it in the framework of dual clustering.

III. Proposed Framework

Here, we detail our DCCA framework for unsupervised band selection. The flowchart is shown in Fig. 2, in which the three procedures each has its own effect, but they should be taken as an integrated one for the reason that each step is necessary for the next. The basic idea is the dual clustering based on context exploration, which can result in a more accurate clustering of bands. Representatives are then chosen by the groupwise strategy from every cluster as the selected bands. Finally, to verify the effectiveness of the former process, these bands are used to complete the classification of HSI. In the following, we will describe each functional aspect of the proposed framework with more details.

A. PHA Descriptor

The context clue of an HSI cube includes two parts. One is the neighboring spectral bands, and the other is the neighboring pixels. This 3-D neighborhood system reflects the environment of a specific pixel. To describe this environment, a PHA descriptor is proposed.

1) Neighborhood System: Suppose the size of the L-band HSI is $w \times h$, with w indicating the width and h the height. The PHA is developed in the neighborhood subset of the HSI. As shown in Fig. 3, we randomly select a $3 \times 3 \times 3$ image cube from the original HSI, which includes three bands of nine pixels. Suppose that we denote the kth band in HSI by B_k and a specific pixel by $P_{i,j}$. In this cube, the bands range from B_{k-1} to B_{k+1}, and the pixels range from $P_{i-1,j-1}$ to $P_{i+1,j+1}$. Then, the center of this cube is defined as $b_{i,j,k}$, which denotes the B_k for $P_{i,j}$ of the original HSI. Similarly, the elements in this cube range from $b_{i-1,j-1,k-1}$ to $b_{i+1,j+1,k+1}$, and the cube can be viewed as the context of the central $b_{i,j,k}$. With these definition, the PHA is constructed on this neighborhood system.

2) Spectral Angle and Spatial Angle: Based on the theory of Spectral Angle Mapper [34], the PHA in the neighborhood subset is proposed. Traditional spectral angle is used for directly comparing the spectra of two pixels. Denote this angle by SA, the first vector by v_1, and the second one by v_2. Then, we have the following equation:

$$SA = \arccos\left(\frac{v_1 \times v_2}{\|v_1\|_2 \times \|v_2\|_2}\right). \quad (1)$$

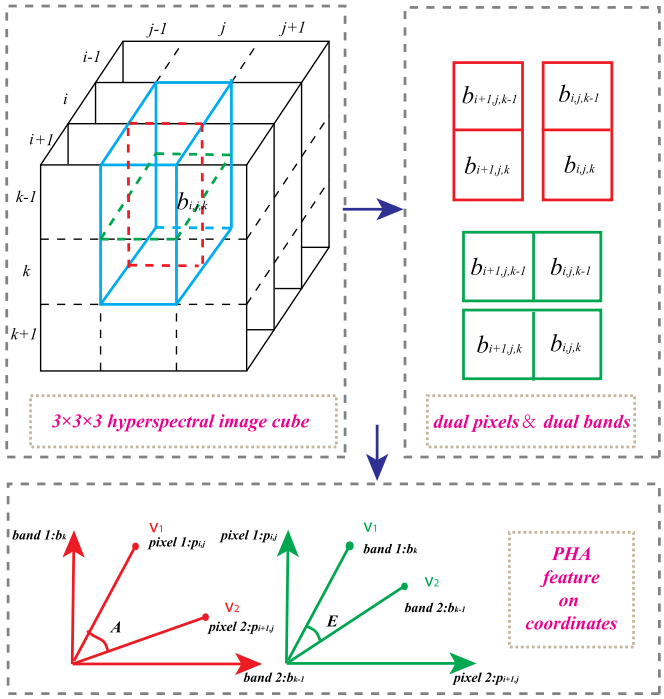

Fig. 3. Illustration of the PHA. We take the blue subcube as an example. This subcube includes four elements: $b_{i,j,k}$, $b_{i,j,k-1}$, $b_{i+1,j,k-1}$, $b_{i+1,j,k}$. The red line divides this subcube into two parts, and each part represents two bands for an HSI pixel. Therefore, we can get two 2-D vectors, and casting them to the vertical and horizontal axes can form a spatial angle of the dual pixels. Similarly, the green line divides the subcube into two parts, and each part denotes the same band of the two neighboring pixels. Thus, the spectral angle of the dual bands can also be obtained like the spatial angle.

However, this spectral angle reflects the global spectral property of two specific pixels. No local spectral and positional clues are considered. Nevertheless, these neglected information is also critical for characterizing the statistical property of the HSI. To properly calculate these clues, we change the meanings of v_1 and v_2 to dual bands and dual pixels in the context of the examined neighborhood cube, which consequently correspond to two angles, i.e., spectral angle and spatial angle. The main difference between these two angles exists in the construction of coordinates. For the spatial angle, the two sides represent the spectral clues of two neighboring pixels (e.g., $P_{i,j}$ and $P_{i+1,j}$). However, different from a traditional representation that utilizes the whole available spectra, only two consecutive bands are considered (e.g., B_{k-1} and B_k). The obtained two vectors (e.g., $[b_{i,j,k-1}\ b_{i,j,k-1}]$ and $[b_{i+1,j,k-1}\ b_{i+1,j,k}]$) span a particular angle in the coordinate space of vertical B_k and horizontal B_{k-1}. Since there are eight such combinations associated with the examined $b_{i,j,k}$, we will have eight local spatial angles.

As for the spectral angle, a similar definition is also followed, only for that the two sides of the angle are spanned by the local neighboring pixels of consecutive bands (e.g., $[b_{i+1,j,k-1}\ b_{i,j,k-1}]$ and $[b_{i+1,j,k}\ b_{i,j,k}]$). The corresponding coordinates are vertical $P_{i,j}$ and horizontal $P_{i+1,j}$. At the same time, there also exist eight spectral angles for the examined $b_{i,j,k}$.

3) PHA Descriptor: As described earlier, there are, in total, 16 angular values for the examined $b_{i,j,k}$, i.e., 8 for spatial angles and 8 for spectral angles. To get a reasonable dimensionality, we only employ the averaged values, consequently leading to the spatial angle $f^1_{k,n}$ and the spectral angle $f^2_{k,n}$, where $n = (i-1) \times w + j$ is the pixel index. With all these, we can formally define the PHA descriptor for the kth band as follows:

$$y_k = \left[f^1_{k,1}, \ldots, f^1_{k,N}, f^2_{k,1}, \ldots, f^2_{k,N}\right] \tag{2}$$

where $N = w \times h$ is the total number of pixels.

B. DCCA

Obtaining the PHA feature together with the raw feature, we can get two clustering results simultaneously. Interaction between the two clustering processes is the primary technique of the dual clustering in our work. Detailed procedures of DCCA are described in the following.

1) Definition: We define the band selection as a dual clustering problem. Two clustering processes are conducted individually in parallel. On one hand, we cluster the bands, according to their raw features. This is the major clustering process that reflects the intrinsic property of bands. On the other hand, we can get another clustering result according to the context, which can be viewed as the associate clustering. These two processes are never independent, for the reason that there should be correlations and the two results should be consistent. Therefore, the ultimate goal of this dual clustering is to establish the relationship between them to enhance the performance of both clusterings. Finally, the enhanced result of the major clustering process is taken as the final result, whereas the associate clustering only acts as an assistance.

In order to model this relationship, the correlation matrix [32] is employed

$$R = \frac{p(M, A)}{P(M) \times P(A)} \tag{3}$$

in which R is the correlation matrix, and M and A are the main feature and associate feature, respectively. Three cases can be summarized from this equation. If $R_{m,a} > 1$, m and a are positively correlated, which means that there is a high probability that the main feature m coexists with the associate feature a. While, if $R_{m,a} = 1$, m and a are mutually independent. The last case is that m and a are negatively correlated if $R_{m,a} < 1$.

2) Dual Clustering: Suppose that the raw description of the HSI is denoted by $X = \{x_1, x_2, \ldots, x_L\}$, in which x_k corresponds to the vectorized representation of the kth band. The PHA associated description is similarly defined as $Y = \{y_1, y_2, \ldots, y_L\}$. There are two phases for the dual clustering, i.e., initial clustering and reclustering.

Initial Clustering: Traditional k-means clustering tries to minimize

$$J = \sum_{l=1}^{L} \sum_{p=1}^{P} r_{l,p} \|x_l - \mu_p\|^2 \tag{4}$$

using the raw feature, in which P represents the total number of clusters, μ_p is the center of the pth cluster, and $r_{l,p}$ represents a

binary value indicating whether x_l belongs to the pth cluster or not. Note that every x_l only belongs to one single cluster.

Analogous to the aforementioned process, Q clusters can also be obtained for Y, according to the principle in (4). However, these initial clustering results are acquired independently. No interaction and correlation are explored, which might be otherwise helpful for the clustering purpose. In order to jointly integrate the two features, we apply the same principle to X and Y, simultaneously, and develop a similar objective function

$$J' = \sum_{l=1}^{L}\sum_{p=1}^{P}\sum_{q=1}^{Q} r_{l,p}^{x} r_{l,q}^{y} \left\|x_l - \mu_p^x\right\|^2 \left\|y_l - \mu_q^y\right\|^2. \quad (5)$$

The minimization of (5), with respect to $r_{l,p}$ and $r_{l,q}$, is to reduce the sum distances both to the centroid of X and Y at the same time.

From the other aspect, considering the relationship of these two clusters, we should also minimize the following equation, which is viewed as the second objective function:

$$H = -\sum_{p=1}^{P}\sum_{q=1}^{Q} R_{p,q}\log(R_{p,q}+1). \quad (6)$$

The $R_{p,q}$ follows the definition of (3) as follows:

$$R_{p,q} = \sum_{l=1}^{L} \frac{p\left(C_{l,p}^{x}, C_{l,q}^{y}\right)}{p\left(C_{l,p}^{x}\right) p\left(C_{l,q}^{y}\right)} \quad (7)$$

in which $p(C_{l,p}^{x})$ represents the probability of band l belonging to the pth cluster for feature x. This probability is approximated by the percentage of bands that are allocated to the pth cluster. The same definition is with $p(C_{l,q}^{y})$. $p(C_{l,p}^{x}, C_{l,q}^{y})$ is the joint probability. We approximate the joint probability by the relative contingency table, in which the element $p(C_{l,p}^{x}, C_{l,q}^{y})$ is the percentage of bands that are allocated to the pth cluster and the qth cluster in the two clustering processes, simultaneously.

The value of H reflects the consistency between the two clusterings. Smaller H with low entropy is more acceptable, while larger H should be avoided. To be more clear, H represents the entropy of R that is utilized to measure the correlation between the dual clustering processes. Unfortunately, the initial results from these two clusterings are mostly different, which implies that the clustering mechanism and results should be adjusted. To restrict H and J' to an appropriate balanced one, we have to optimize the both objective functions simultaneously.

Reclustering: From the aforementioned analysis, we can find the contradiction between the two objective functions. On one hand, we hope the results to be with low entropy, while on the other hand, the results in accord with the criterion of k-means clustering are not always with the lowest entropy. To strike a balance, a reclustering process is proposed. This process is accomplished through introducing the correlation matrix into the original k-means algorithm, which is defined as follows:

$$\arg\min_{p,q} \frac{\left\|x_l - \mu_p^x\right\|^2 + \left\|y_l - \mu_q^y\right\|^2}{R_{p,q}^1} \quad (8)$$

where $R_{p,q}^1$ is the normalized representation of $R_{p,q}$ as follows:

$$R_{p,q}^1 = \frac{\log(1+R_{p,q})}{\sum_{p,q}\log(1+R_{p,q})}. \quad (9)$$

In (8), the objective function of the two clusterings is divided by the corresponding $R_{p,q}^1$. The minimization is to get better clustering results C^x and C^y with higher correlation between the dual clusterings.

After this procedure, each band will be reclustered into the new optimal dual clusters C^x and C^y. We recalculate the newly constructed clusters' centroid $\{\mu_1^x, \mu_2^x, \ldots, \mu_P^x\}$ and $\{\mu_1^y, \mu_2^y, \ldots, \mu_Q^y\}$, and then return to the first step.

The iteration of these two phases does not terminate until both the clustering C^x and C^y are stable. Then, the associate cluster is deserted, and the enhanced major cluster C^x enters into the next step of DCCA. A more detailed pseudocode is shown in Algorithm 1.

Algorithm 1 Dual Clustering by Context Analysis

Input: $X = \{x_1, x_2, \ldots, x_L\}, Y = \{y_1, y_2, \ldots, y_L\}$
Output: C^x, C^y

μ_x and $\mu_y \leftarrow$ INITIALCLUSTERING(X, Y)
function RECLUSTERING (μ^x, μ^y)
  while C^x or C^y changes **do**
    calculate $\arg\min_{p,q} \frac{\|x_l-\mu_p^x\|^2+\|y_l-\mu_q^y\|^2}{R_{p,q}^1}$, update C^x and C^y
    $R_{p,q} \leftarrow \sum_{l=1}^{L} \frac{p(C_{l,p}^x, C_{l,q}^y)}{p(C_{l,p}^x)p(C_{l,q}^y)}$
    $R_{p,q}^1 \leftarrow \frac{\log(1+R_{p,q})}{\sum_{p,q}\log(1+R_{p,q})}$
    calculate the μ_x and μ_y in C^x and C^y
  end while
  return C^x and C^y
end function

function INITIALCLUSTERING(X, Y)
  randomly choose both μ_x and μ_y
  while μ_x or μ_y changes **do**
    $J' \leftarrow \sum_{l=1}^{L}\sum_{p=1}^{P}\sum_{q=1}^{Q} r_{l,p}^x r_{l,q}^y \|x_l - \mu_p^x\|^2 \|y_l - \mu_q^y\|^2$
    minimize J', calculate $r_{l,p}$ and $r_{l,q}$
    $\mu_p = \mu_x \leftarrow \frac{\sum_{l=1}^{L} r_{l,p} x_l}{\sum_{l=1}^{L} r_{l,p}}$
    $\mu_p = \mu_y \leftarrow \frac{\sum_{l=1}^{L} r_{l,q} y_l}{\sum_{l=1}^{L} r_{l,q}}$
  end while
  return μ_x and μ_y
end function

TABLE I
COMPARISON OF TIME COMPLEXITY BETWEEN DUAL CLUSTERING AND k-MEANS METHOD

method	initial clustering	re-clustering	Overall	T(n)
k-means	kL(2n-1)+1	nL	2nkL+1	$O(n)$
dual clustering	4nkL+2+$n(4kL+3\mathrm{kL}^2)-k$	2kL	$n(8kL+3\mathrm{kL}^2)+k(2L-1)+2$	$O(n)$

3) Computational Complexity Analysis: We discuss the computational cost of the proposed dual clustering algorithm for HSI band selection in comparison to the standard k-means clustering method in this subsection. To differentiate among the complexities of dual clustering and k-means method with higher precision, both the arithmetic operations and the big O notation are used to calculate the computational cost.

First, it is supposed that the iteration times of k-means and dual clustering are the same and represented by k, the other parameters of are similar with Algorithm 1. We also suppose that the k-means process is the initialization of X of dual clustering in Algorithm 1. The number of k-means clusters is n, and for dual clustering, this is $2n$ (suppose that the two clustering processes have the same number of clusters). The number of bands is L. It is also supposed that all the data that we need are prepared in advance (including X, Y and the all possible relation matrix $R_{p,q}$). The iteration process of k-means and dual clustering is both divided into two parts, i.e., initial clustering and reclustering. Time complexities are calculated, respectively for these two parts. The initial clustering consists of the initial center selection, as well as clustering, and the reclustering is to update new clustering centers. These two operations and the total time complexity are compared in Table I.

It is not difficult to find that the dual clustering process is more time consuming than the k-means method, obviously, which means that this dual clustering is not efficient enough. However, we can further approximate the time complexity of both the k-means and dual clustering by $O(n)$. From this point, the time complexity of these two processes are still with the same order of magnitude. The extra time of dual clustering mainly spends on the interaction between the two clustering processes. However, due to the existence of this interaction, the clustering accuracy of dual clustering, most of the time, exceeds the k-means by 5%–10% [32].

C. Groupwise Representative Selection

In this step, we choose representatives from every cluster as the final selected bands. Traditional methods choose the representatives only by comparing the bands inside its corresponding cluster [13]. To a certain extent, these choices can best represent the clusters. However, the selected bands might not be necessarily discriminative enough, which can lead to a poor classification performance for the further task. We have to consider both the representativeness and the distinctiveness of these representatives, simultaneously. In order to explain this point, we refer to Fig. 4 as an illustration, from which the triangular points indicate the mean center of each group and the circled ones are with maximum interclass distances. Traditional methods may choose the former as the representatives, but in our view, which of these two sets will perform better in the following classification process is undeterminable. It is possible that we may achieve a better classification result in the second case, for the reason that the triangular points are so close and similar that they are not distinctive enough. From this aspect, the problem can be translated into finding the optimal points in this figure that can, on one hand, represent the all data properly and, on the other hand, demonstrate considerable disparity.

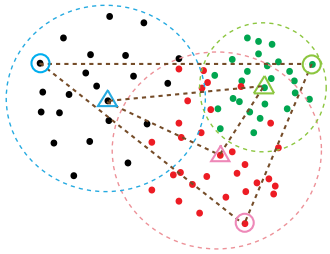

Fig. 4. Choosing process of cluster representatives. Most traditional methods choose the cluster centers as the representatives, as the triangles show. However, this setting is not always the best because the discriminative ability is not necessarily powerful. The between cluster distance should also be considered, as the circles show.

In order to solve this dilemma, we have to consider the relationship between the two principles. We not only consider the individual representative ability for each cluster but also explore their distinctive ability by treating the selected bands together. In the following, detailed procedures will be introduced.

We cast the problem as a joint Euclidean distance maximization problem, which includes the intracluster (the representative and the other points in a cluster) and intercluster (the representatives of each cluster) distances. Suppose that the representatives of the clusters are denoted by $E=\{e_1,e_2,\ldots,e_P\}$, and the mean centers are denoted by $\{\mu_1,\mu_2,\ldots,\mu_P\}$. Let $D_{i,j}$ represent the Euclidean distance between arbitrary two representatives e_i and e_j, $d_{i,j}$ the distance between each corresponding two cluster centers μ_i and μ_j, and $c_{i,j}$ the bias between e_i and μ_j. Then, the problem of searching for the most appropriate representatives can be formulated as

$$\underset{e_1,e_2,\ldots,e_P}{\arg\max}\ S(E) \tag{10}$$

where

$$S(E)=\frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\left(\frac{D_{i,j}}{d_{i,j}}+\lambda\frac{1}{c_{i,i}+c_{i,j}}\right). \tag{11}$$

For this joint Euclidean distance, we first encourage the discrimination among representatives by the first term. For this term, arbitrary $D_{i,j}$ is restricted to the same scale by $d_{i,j}$ to avoid bias. This term is named the real disparity among representatives, which is abbreviated by RD. Higher RD implies better distinction. Second, we hope that the representatives are not far from the mean centers and constrained by the second

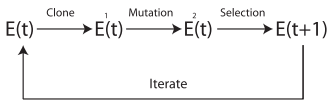

Fig. 5. Illustration of ICS.

term. The nearer the representative is to the center, the higher representativeness it demonstrates. For each pair of representatives i and j, this term includes the sum of $c_{i,i}$ and $c_{i,j}$, which is named as to-center distance TD. The reciprocal of TD is the expression of this term. The two terms are balanced by λ. A higher value for this objective function means that the chosen representatives are with higher discrimination that can better represent the whole HSI data, and at the same time, the expressive ability inside each cluster is more powerful. Moreover, we find that the TD changes much faster than RD. To balance their effects, we heuristically set lambda as 0.1.

However, a problem exists in that its an NP problem to optimize this equation. Every settlement of an individual representative will influence the choice of other representatives. Therefore, we have to jointly consider all the distances to find the global optimal solution. Suppose that there are n clusters and the average cluster volume is k. The computation complexity of exhaustive search can be approximated by $O(n^2 \times k^n)$. To reduce this to polynomial complexity, we introduce the Immune Clonal Strategy (ICS) [24] to solve this problem. We take (10) as the affinity function in the ICS. The representatives are regarded as the antibodies, and the original HSI data is viewed as the antigen. Various antibodies have different effectiveness for the antigen. By maximizing the affinity function, the best antibodies are chosen. The detailed optimization procedure is modeled as an iterative process, as shown in Fig. 5. It includes three phases: clone, mutation, and selection. Through this operation, the general time complexity reduces to $O(n^2)$.

Clone Phase: At the very first, we randomly pick out n sets of candidate representatives as the antibodies $\mathscr{E}(t) = \{E_1(t), E_2(t), \ldots, E_n(t)\}$. For notational simplicity, we omit the iteration number t in the following explanation with no loss of understandability. Take the ith antibody as an example. The number of clones $n_c(E_i)$ depends on the affinity of E_i. The antibody with higher affinity is encouraged to generate more clones. To be specific, suppose that N_c is the maximum number of clones predefined as a threshold. Then, $n_c(E_i)$ is defined as

$$n_c(E_i) = \mathrm{Int}\left(\frac{S(E_i)}{\max\limits_{j=1,2,\ldots,n} S(E_j)} \times N_c\right) \tag{12}$$

where Int(.) is the rounding up function. After the clone phase, we denote the clones of the antibody E_i by $\mathscr{E}_i^1 = \{E_{i,1}^1, E_{i,2}^1, \ldots, E_{i,n_c(E_i)}^1\}$.

Mutation Phase: The randomly chosen antibodies are not the best. Therefore, a mutation phase after the clone phase is necessary. For example, for the antibody E_i, we randomly replace N_m representatives in its clone $E_{i,j}^1$ by the same number of elements, each from the corresponding cluster. There is no doubt that these newly introduced elements should differ from the former representatives, which enrich the diversity of the original antibodies. After this, mutated antibodies $\mathscr{E}_i^2 = \{E_{i,1}^2, E_{i,2}^2, \ldots, E_{i,n_c(E_i)}^2\}$ are obtained.

TABLE II
NUMBER OF SAMPLES FOR EACH CLASS OF THE INDIAN PINES IMAGE

Class	U	V	T	Total
Corn-notill	80	20	1328	1428
Corn-mintill	80	20	730	830
Grass-pasture	80	20	383	483
Grass/Trees	80	20	630	730
Hay-windrowed	80	20	378	478
soybeans-notill	80	20	872	972
Soybeans-min	80	20	2355	2455
Soybean-clean	80	20	493	593
Woods	80	20	1165	1265

Selection Phase: With the obtained antibodies, which are manifestly more various than the original set, we will select the most promising ones for the next round of processing. The principle is also defined with the affinity values. Higher ones indicate more fitness. Therefore, we have

$$E_i(t+1) = \arg\max_{E} \left\{ S\left(E_{i,1}^1\right), S\left(E_{i,2}^1\right), \ldots, \times S\left(E_{i,n_c(E_i)}^1\right), S\left(E_{i,1}^2\right), S\left(E_{i,2}^2\right), \ldots, S\left(E_{i,n_c(E_i)}^2\right) \right\} \tag{13}$$

which means that the antibody with the largest affinity value is taken as $E_i(t+1)$ to enter the next iteration.

The iteration does not terminate until the change between $S(E_i)$ and $S(E_i(t+1))$ is smaller than M_{th} or the maximum number of iteration M_{it} is reached.

IV. EXPERIMENTS AND ANALYSES

This section will show the experimental results of our method compared with the existing band selection methods. To verify the performance, we apply the selected bands to hyperspectral classification. The number of selected bands has a wide range, and three data sets are chosen to demonstrate the superiority of our method.

A. Data Sets

Three publicly available HSIs are applied to verify the superiority of our method. They are Indian Pines, Salinas, and Pavia University. The description are as follows.

The *Indian Pines* image was gathered over a vegetation area in Northwestern Indiana by the AVIRIS sensor. The image whose spatial resolution is 20 m/pixel consists of 145 × 145 pixels and 224 spectral reflectance bands. Sixteen classes of interest are contained in this image, of which the nine major categories are selected in our experiment. One hundred training samples for each class of interest are randomly chosen for training. These 100 training samples are further divided into training and validation sets in our cross-validation process. The detailed allocation of samples is listed in Table II, in which U represents the training set for cross-validation, V represents the validation set, and T represents the testing samples. The same acronyms are used for Tables III and IV.

The *Salinas scene* was also collected by the AVIRIS sensor over Salinas Valley, California. The image is characterized by a

TABLE III
NUMBER OF SAMPLES FOR EACH CLASS OF THE SALINAS SCENE IMAGE

Class	U	V	T	Total
Brocoli-green-weeds-1	80	20	1999	2009
Brocoli-green-weeds-2	80	20	3626	3726
Fallow	80	20	1876	1976
Fallow-rough-plow	80	20	1294	1394
Fallow-smooth	80	20	2578	2678
Stubble	80	20	3859	3959
Celery	80	20	3479	3579
Grapes-untrained	80	20	11171	11271
Soil-vinyard-develop	80	20	6103	6203
Corn-senesced-green-weeds	80	20	3178	3278
Lettuce-romaine-4wk	80	20	968	1068
Lettuce-romaine-5wk	80	20	1827	1927
Lettuce-romaine-6wk	80	20	816	916
Lettuce-romaine-7wk	80	20	970	1070
Vinyard-untrained	80	20	7168	7268
Vinyard-vertical-trellis	80	20	1707	1807

TABLE IV
NUMBER OF SAMPLES FOR EACH CLASS OF THE PAVIA UNIVERSITY IMAGE

Class	U	V	T	Total
Asphalt	240	60	6004	6304
Meadows	240	60	17846	18146
Gravel	240	60	1515	1815
Trees	240	60	2612	2912
Metal sheets	240	60	813	1113
Bare soil	240	60	4272	4572
Bitumen	240	60	681	981
Bricks	240	60	3064	3364
Shadows	240	60	495	795

spatial resolution of 3.7 m/pixel with 224 spectral bands, which comprises 512 $\times$ 217 samples. The image contains 16 classes of interest, including vegetables, bare soils, and so on. The 100 training samples for each class of interest are randomly selected to accomplish the experiment. The detailed allocation of samples is listed in Table III.

The *Pavia University* was captured by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over Pavia, Northern Italy. The spatial resolution of the image is 1.3 m/pixel, and 103 spectral bands are included. This image comprises 610 $\times$ 340 samples and nine classes of interest. For this image, we complete the trial with 300 training samples for each class of interest. The detailed allocation of samples is listed in Table IV.

B. Competitors

To verify the superiority of the proposed method, the comparison experiment mainly includes three parts.

- The first part comes from the comparison with the existing *clustering-based band selection* (CBBS) methods [4], [13]. These methods tend to join similar bands together to derive clusters that preserve low variance among bands inside one cluster, and at the same time, the variance is high among different clusters. From every cluster, the best representative, which is the finally selected band, is chosen. This agglomerative clustering strategy can also preserve the hierarchal property of hyperspectral data. To measure the similarity of different bands, two main methods, i.e., the MI and the Kullback–Leibler divergence (KLD), are adopted, which we denote as CBBS-MI and CBBS-KLD. The first one is to measure to what extent can a specific band explain another one, and based on this criterion, the similarity of bands is measured. The second one is a method that measures the information loss of substituting one band with another, which is also a popular CBBS method.
- The second part is the comparison with another kind of band selection method, namely, *constrained band selection* (CBS) [22], [29], [30]. This method uses different strategies to minimize the correlation and dependence to select bands. The main strategy includes the CEM and LCMV. The first one converts a band of image to a vector, and the later one takes a band image as a matrix. Four specific criteria, i.e., band correlation minimization (BCM), band dependence minimization (BDM), band correlation constraint (BCC), and band dependence constraint (BDC), divide these competitors to four parts: CEM-BCM/BDM, CEM-BCC/BDC, LCMV-BCM/BDM, and LCMV-BCC/BDC.
- In the third part, we verify the effectiveness of different components in DCCA. First, we replace the dual clustering with the traditional k-means method and keep the other steps unchanged. The obtained method is named as *ordinary-clustering-based band selection* (OCBBS). The comparison between DCCA and OCBBS can prove the effectiveness of dual clustering, and the comparison between OCBBS and CBBS can show the usefulness of the groupwise representative selection strategy. Second, for DCCA, we compare it with *dual-clustering-based band selection without context analysis* (DCWCA) to justify the effectiveness of contextual clues. We take the spectral value feature and spectral gradient feature as two observable views of DCWCA, instead of the proposed contextual clue.

C. Classifier Description

In order to evaluate the performance of the proposed band selection method, the selected bands are used to accomplish the process of HSI classification. Higher classification accuracy means that the selected bands are more discriminative and can represent the original HSI better. The classification methods that we choose include the naive Bayes method [46], k-nearest neighborhood (kNN) [47], classification and regression trees (CART) [48], and support vector machine (SVM) [49], [50]. For the naive Bayes method, we suppose that the data are in accord with a normal Gaussian distribution. Density function is estimated by the maximum-likelihood method, and the prior probabilities for each class are estimated from the relative occurrence frequencies of the classes in training. As a nonparametric, the CART method whose parameter is not that important only includes two regular steps: tree building and predict. For the parameter of kNN, the k represents the scale of neighboring pixels, and the distance specifying the distance metric is Euclidean. As for SVM, we select the radial basis function (RBF) kernel and multiclass support based on a one-against-all scheme to accomplish our experiment, taking advantage of LibSVM (C-support vector classification

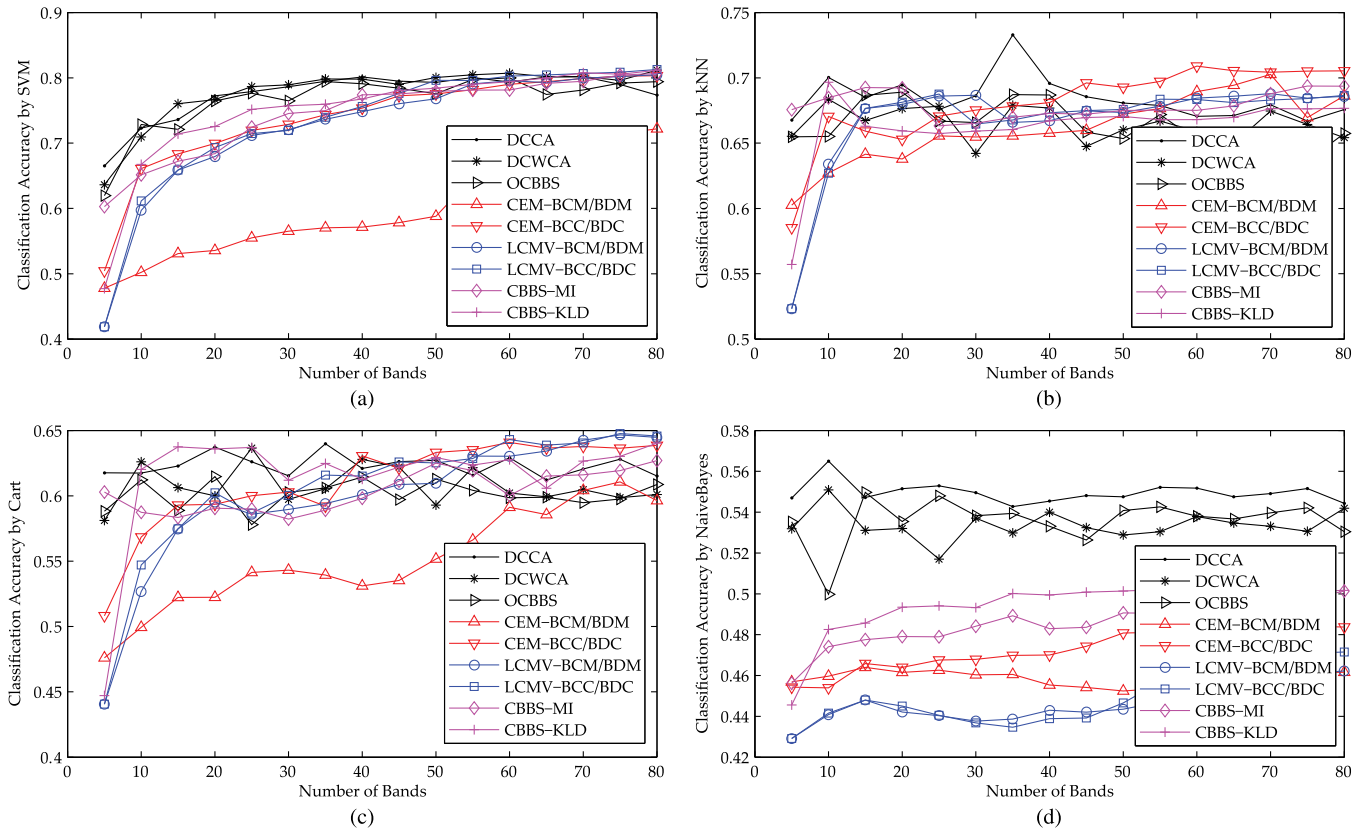

Fig. 6. Classification result on Indian Pines image. (a)–(d) Classification results of SVM, kNN, CART, and Naive Bayes, respectively.

(C-SVC), RBF). The remaining insignificant parameters of these four methods are the same as the default setting of MATLAB 2014a without any changes. We use the cross-validation to train the parameters in SVM and kNN. The corresponding allocations of samples are given in Tables II–IV.

1) Naive Bayes: This traditional method is widely used in image classification. The prior possibility and distribution is first calculated on the training samples, and based on this information, the possibility of each testing sample belonging to each class is obtained. This method is with high efficiency, although the accuracy is always unsatisfying.

2) kNN: This HSI classification method is based on a majority-voting scheme. The testing sample is classified by the neighboring training samples. The main categories of the training samples may determine the central one. When the training samples are large, the result will turn out to be with high accuracy. Moreover, the only parameter k representing the scale of neighboring pixels is determined in fivefold cross-validation to avoid bias. This parameter ranges from 1 to 100, with a step size increment of 1. By the experiment, we find with a neighboring size of 2 in Indian Pines, 4 in Salinas scene, and 4 in Pavia University, The cross-validation is with the best performance.

3) CART: As a famous nonparametric method, this method is based on the binary-decision-tree technique. The main idea of this algorithm is to split the initial data. The initial data are first divided to two subdata, and then, the same process continues in the subdata. The data splitting do not stop until the ultimate classification result is achieved. The category of testing sample is predicted effectively by this method.

4) SVM: SVM is one of the most accurate classification technique. This method can roughly be divided into two types, i.e., one-against-one SVM and one-against-all SVM, and we choose the latter. This method first trains the decision boundary on the training set. Then, for the testing samples, the categories are obtained by maximizing the margin of the decision boundary. We use SVM classification with RBF kernel. Fivefold cross-validation is conducted to avoid parameter bias, with the penalty parameter C of the SVM being tested between 1×10^3 and 1×10^6, with a step size increment of 1×10^3, and the kernel length scale σ of the RBF kernel being tested between 10 and 1000, with a step size increment of 10. On different images, the best parameters are almost the same, and we only set C as 1×10^5 and σ as 100, after cross validation. These obtained best parameter values are used in the classification.

D. Experimental Results

To show a convictive result of the band selection method, the selected band number ranges from 5 to 80, for the reason that, when the number of bands reaches 80, the accuracy is almost stable. However, for a clear presentation, the results are sampled every five bands. Before detailed introduction and comparison, some preliminaries should be introduced first.

- In our method, the clustering step is the fundamental process. However, both the dual clustering process and the k-means method can only get similar but never the same clustering results for every time of experiment. The main reason lies in that the initial centers of clusters

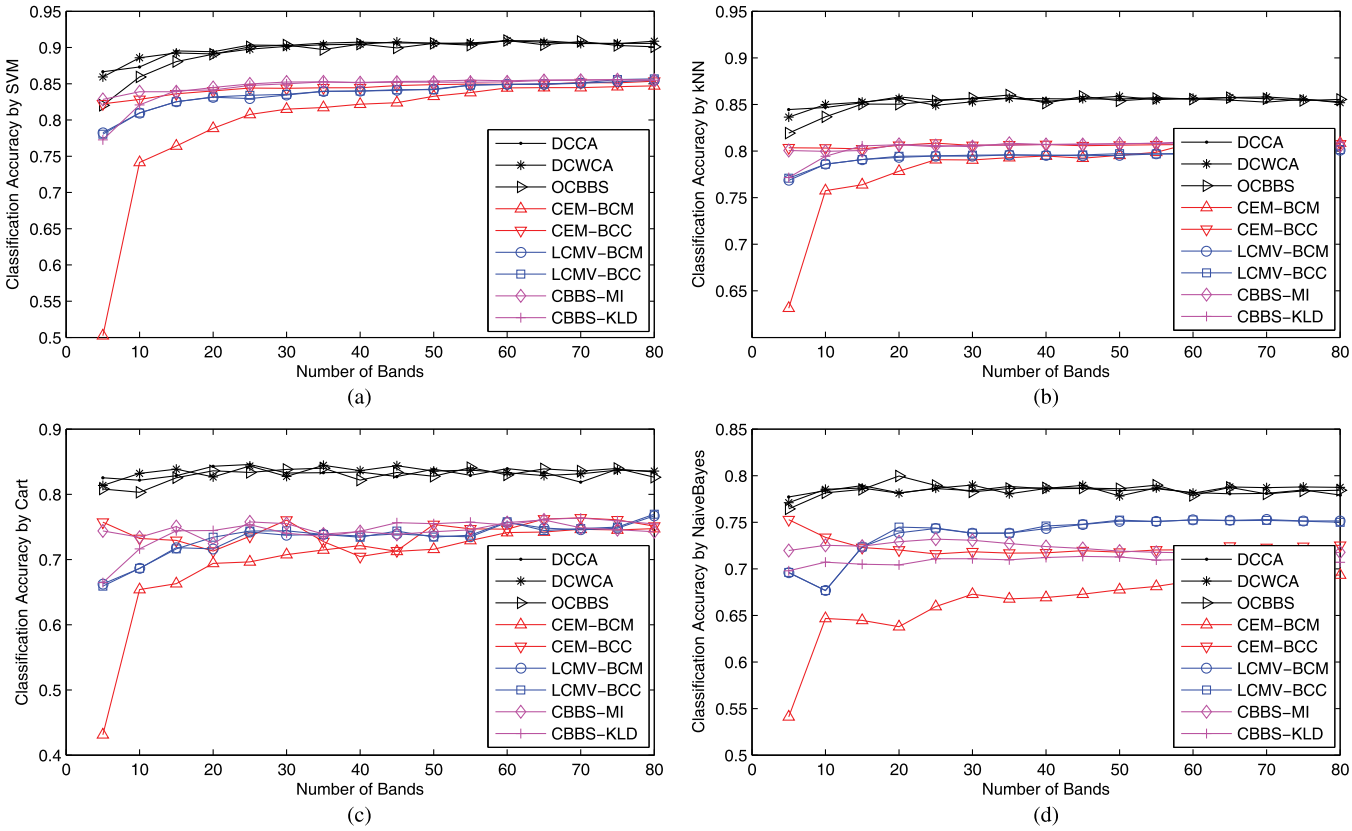

Fig. 7. Classification result on Salinas scene image. (a)–(d) Classification results of SVM, kNN, CART, and Naive Bayes, respectively.

are chosen randomly that causes the inherent instability. Therefore, the accuracy results involving the two clustering procedures are all averaged results for several times of clustering tests.

- In our view, the classification results with fewer bands would reflect the effectiveness of band selection process more obviously. The reason exists in that the main purpose of band selection is to find out the least bands with the most abundant information, which can relieve the workload of the following HSI processing such as image classification or segmentation. As a result, the redundant bands should be as few as possible. If the selected band number is not large but the performance is satisfying, we think that the band selection method is very effective. To make a distinct comparison of the bands that we selected, we first demonstrate the results in Figs. 6–8. Another more detailed list with mean values and variances is in Tables VI–VIII, where results focusing on the bands ranging from 5 to 20 are further enhanced. Furthermore, we also give out the specific chosen bands in Table V as representatives that correspond to the listed result shown in Tables VI–VIII.
- There are several parameters to be set in this experiment. In the step of representative band selection, the two parameters N_c and N_m need to be determined. However, we find that N_c and N_m mainly influence the number of iterations. Larger values will lead to more clones and mutations in a single step. However, the change rate may be too quick to miss the right answer. Smaller values will give a steady change of the antibodies but more iterations. Therefore, we empirically set the two parameters as 5 in all the experiments. There are also two thresholds in the ICS step, i.e., $M_{\rm th}$ and $M_{\rm it}$. From the experimental observation, 10^{-4} and 60 are adequate for the two variables.

In the following, we will give a detailed analysis and comparison of the experimental results. In Figs. 6–8 and Tables VI–VIII, we can get a general impression that the proposed DCCA can outperform the others most of the time. However, the performances on different images and classifiers slightly vary with each other.

1) Comparison With Clustering-Based Methods: The proposed DCCA, DCWCA, and OCBBS are based on the clustering prototype. Thus, we compare them with traditional clustering-based methods, including CBBS-MI and CBBS-KLD. From the results, we can say the best performance comes from the dual-clustering-based methods DCCA and DCWCA. Moreover, OCBBS also performs well. The CBBS-MI and CBBS-KLD methods perform slightly worse compared with the former ones. The main reason for this difference comes from the clustering process.

These five methods are all based on clustering. However, as [13] shows, the clustering process plays an important role in this kind of technique. The better the cluster process, the more adequate the selected bands become. The former two methods DCCA and DCWCA introduce the dual clustering process, which consider the clustering process from two views. The two views benefit each other and bring a better cluster result. Nevertheless, the latter ones only take one perspective into

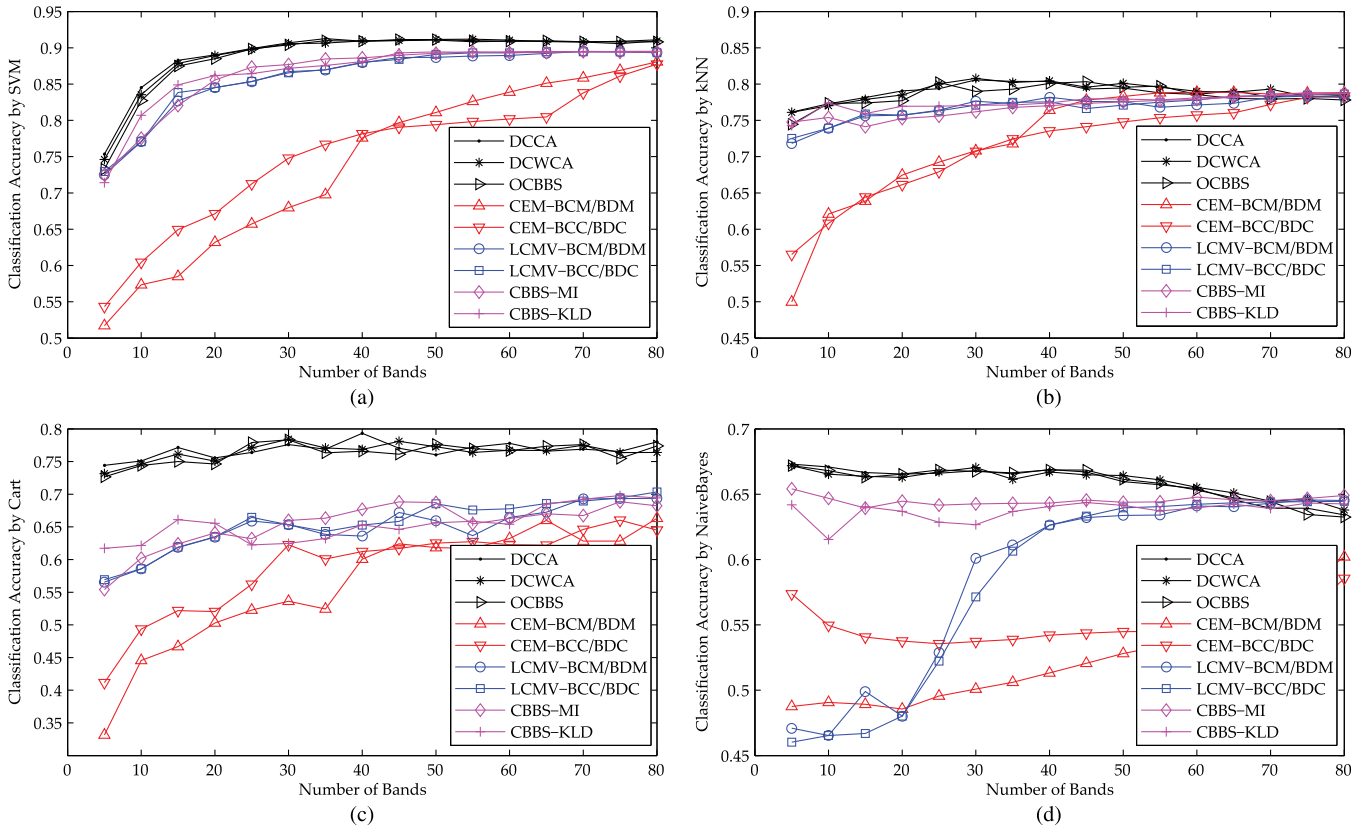

Fig. 8. Classification result on Pavia University image. (a)–(d) Classification results of SVM, kNN, CART, and Naive Bayes, respectively.

TABLE V
CHOSEN BANDS OF DCCA FOR DIFFERENT IMAGES

Chosen bands	K = 5	K = 10	K = 15	K = 20
Indian Pines image	21, 42, 69, 132, 175	9, 18, 48, 55, 76, 98, 143, 154, 189,196	13, 22, 28, 37, 51, 62, 70, 81, 86, 98, 126, 141, 163, 195, 201	8, 13, 27, 34, 39, 41, 48, 68, 74, 76, 77, 100, 133, 141, 143, 149, 162, 168, 188, 200
Salinas scene image	32, 45, 91, 142, 168	26, 40, 42, 46, 59, 95, 117, 123, 182, 217	7, 17, 20, 33, 38, 40, 42, 52, 64, 66, 80, 116, 137, 194, 220	6, 17, 35, 39, 40, 53, 61, 76, 80, 81, 91, 103, 114, 136, 156, 186, 194, 205, 213, 215
Pavia University image	10, 59 , 70, 76, 88	2, 11, 23, 28, 44, 50, 72, 76, 83, 91	1, 4, 8, 11, 17, 26, 38, 42, 48, 63, 68, 72, 78, 82, 92	1, 4, 8, 12, 17, 22, 29, 32, 43, 48, 55, 62, 68, 72 ,74, 78, 83, 85, 92, 99

consideration. As for CBBS-MI and CBBS-KLD, these two methods achieve almost identical results, for the reason that KLD and MI are only two kinds of criterions for similarity measurement, and the basic technique for this kind of methods are the same.

For DCCA and DCWCA, the former is relatively more superior than the latter, particularly as shown in Tables VI–VIII. The difference between these two verifies the importance of context clue. Most of the time, DCCA, which makes use of the PHA feature of the band context, outperforms DCWCA. Nevertheless, the two views of DCWCA come from the raw feature and the gradient feature [51]. Compared with PHA, the gradient feature, which merely illustrates the changing process of the neighboring bands, fails to consider the spatial relation of HSI. The advantage of PHA helps DCCA select more competitive bands. In particular, on Indian Pines and Pavia University images, the superiority of DCCA is manifest.

For OCBBS and DCCA, the comparison can more convincingly tell the fact that dual clustering is more useful than traditional clustering. Thus, DCCA performs better than OCBBS. Another comparison between OCBBS- and CBBS-based methods demonstrates the significance of groupwise band selection strategy. The fundamental difference of them is that the groupwise process selects representatives of clusters. The proposed DCCA takes the intercluster distance into consideration, which will enhance the discrimination among bands. While the CBBS-based methods only consider the intracluster distance, which leads to a limited discriminative ability.

TABLE VI
CLASSIFICATION ACCURACY FOR INDIAN PINES IMAGE (THE BEST PERFORMANCE IS EMPHASIZED IN BOLDFACE)

SVM	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	**0.6383**	0.6333	0.6190	0.6028	0.4773	0.4779	0.5042	0.4188	0.4188
K = 10	0.7301	0.7290	**0.7309**	0.6515	0.6675	0.5021	0.6611	0.5970	0.6116
K = 15	0.7383	0.7548	**0.7644**	0.6723	0.7140	0.5308	0.6835	0.6587	0.6587
K = 20	0.7875	**0.7857**	0.7721	0.6840	0.7255	0.5356	0.6994	0.6796	0.6933
kNN	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	0.6625	0.6599	0.6604	**0.6756**	0.5571	0.6025	0.5852	0.5232	0.5232
K = 10	**0.7001**	0.6776	0.6766	0.6848	0.6966	0.6272	0.6704	0.6340	0.6270
K = 15	0.6788	0.6711	0.6871	**0.6926**	0.6629	0.6415	0.6592	0.6766	0.6766
K = 20	0.6733	0.6831	0.6833	**0.6923**	0.6594	0.6379	0.6529	0.6797	0.6814
Cart	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	**0.6250**	0.6194	0.5698	0.6028	0.4471	0.4760	0.5082	0.4406	0.4406
K = 10	0.6172	**0.6232**	0.6079	0.5874	0.5874	0.4994	0.5685	0.5268	0.5471
K = 15	**0.6216**	0.6168	0.6036	0.5835	0.5835	0.5223	0.5929	0.5747	0.5747
K = 20	**0.6301**	0.5985	0.6103	0.5905	0.5905	0.5224	0.5935	0.5953	0.6024
NaiveBayes	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	**0.5112**	0.5042	0.4874	0.4562	0.4456	0.4569	0.4543	0.4291	0.4291
K = 10	0.5144	**0.5247**	0.5096	0.4741	0.4825	0.4597	0.4539	0.4408	0.4416
K = 15	0.5080	0.5122	**0.5206**	0.4775	0.4857	0.4640	0.4658	0.4480	0.4480
K = 20	**0.5292**	0.5208	0.5263	0.4791	0.4935	0.4615	0.4640	0.4421	0.4451

TABLE VII
CLASSIFICATION ACCURACY FOR SALINAS SCENE IMAGE (THE BEST PERFORMANCE IS EMPHASIZED IN BOLDFACE)

SVM	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	0.8681	**0.8747**	0.8683	0.8284	0.7726	0.5024	0.8223	0.7825	0.7804
K = 10	0.8836	0.8845	**0.8939**	0.8390	0.8209	0.7414	0.8283	0.8096	0.8096
K = 15	**0.9002**	0.8989	0.8881	0.8390	0.8407	0.7641	0.8362	0.8251	0.8251
K = 20	**0.8994**	0.8947	0.8977	0.8449	0.8415	0.7886	0.8401	0.8314	0.8318
kNN	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	**0.8405**	0.8396	0.8361	0.8006	0.7716	0.6313	0.8035	0.7685	0.7709
K = 10	**0.8502**	0.8452	0.8471	0.7997	0.7944	0.7575	0.8032	0.7859	0.7859
K = 15	0.8461	0.8491	**0.8495**	0.8006	0.8056	0.7639	0.8023	0.7908	0.7908
K = 20	**0.8502**	0.8466	0.8461	0.8072	0.8068	0.7784	0.8061	0.7933	0.7944
Cart	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	**0.8213**	0.8212	0.8174	0.7437	0.6645	0.4311	0.7573	0.6625	0.6591
K = 10	0.8284	0.8271	**0.8310**	0.7341	0.7163	0.6539	0.7323	0.6862	0.6865
K = 15	0.8282	0.8310	**0.8311**	0.7506	0.7440	0.6630	0.7292	0.7181	0.7164
K = 20	**0.8389**	0.8290	0.8299	0.7243	0.7444	0.6939	0.7129	0.7164	0.7342
NaiveBayes	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	0.7667	0.7695	**0.7908**	0.7196	0.6978	0.5412	0.7528	0.6957	0.6902
K = 10	0.7850	**0.7911**	0.7879	0.7252	0.7071	0.6467	0.7338	0.6767	0.6767
K = 15	0.7922	**0.7985**	0.7835	0.7243	0.7049	0.6447	0.7229	0.7235	0.7235
K = 20	0.7873	**0.7928**	0.7843	0.7291	0.7042	0.6379	0.7202	0.7387	0.7447

2) Comparison With Constraint-Based Methods: We also compare our method DCCA with the CBS methods, including CEM-BCM/BDM, CEM-BCC/BDC, LCMV-BCM/BDM, and LCMV-BCC/BDC. In general, DCCA outperforms the constraint-based ones, although the LCMV-based method performs slightly better on the Pavia University image. The CEM-based methods are always the poorest. Another principle that can also be concluded for constraint-based method is that the accuracy always rises with the increasing number of selected bands. In addition, the better performance of these kinds of methods always comes with a higher number of selected bands. This phenomenon diverts from the desired property of band selection. We hope that the fewer bands can achieve satisfying results.

3) Analysis for Different Images: We also want to make a detailed analysis on each image. Generally, the Indian Pines image is with higher difficulty for classification, and the other two ones are easier to classify. On Indian Pines, the dual clustering-based methods take the leading position for SVM and naive Bayes classifiers, as shown in Fig. 6 and Table VI. For the other two classifiers CART and kNN, their performance is about the same with CBBS-MI and CBBS-KLD. In addition, we have to emphasize that, for DCCA, when the selected bands are few, the results of classification are still satisfying, which is not always the case for the other comparative methods. For example, the accuracy of kNN reaches 70.01% with ten bands, which outperforms the classification result of 69.46% obtained by all bands. The same case is true for the SVM classifier.

TABLE VIII
CLASSIFICATION ACCURACY FOR PAVIA UNIVERSITY IMAGE (THE BEST PERFORMANCE IS EMPHASIZED IN BOLDFACE)

SVM	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	0.748	**0.7610**	0.7602	0.7252	0.7142	0.5171	0.5434	0.7249	0.7295
K = 10	0.8416	**0.8579**	0.8545	0.7754	0.8070	0.5735	0.6043	0.7706	0.7706
K = 15	0.8737	0.8711	**0.8766**	0.8205	0.8489	0.5849	0.6493	0.8270	0.8383
K = 20	**0.8992**	0.8868	0.8933	0.8556	0.8618	0.6321	0.6713	0.8452	0.8452
kNN	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	**0.7838**	0.7589	0.7741	0.7476	0.7427	0.4997	0.5653	0.7182	0.7251
K = 10	0.7860	0.7852	**0.7890**	0.7539	0.7741	0.6209	0.6081	0.7393	0.7393
K = 15	0.7929	0.7891	**0.7936**	0.7413	0.7594	0.6387	0.6443	0.7562	0.7589
K = 20	0.7942	0.7936	**0.7971**	0.7527	0.7696	0.6745	0.6614	0.7570	0.7570
Cart	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	**0.7227**	0.7015	0.7035	0.5532	0.6173	0.3316	0.4115	0.5654	0.5693
K = 10	**0.7385**	0.7272	0.7292	0.6020	0.6216	0.4456	0.4935	0.5856	0.5858
K = 15	**0.7477**	0.7383	0.7464	0.6238	0.6610	0.4668	0.5218	0.6184	0.6189
K = 20	**0.7406**	0.7324	0.7350	0.6407	0.6555	0.5029	0.5204	0.6343	0.6346
NaiveBayes	DCCA	DCWCA	OCBBS	CBBS-MI	CBBS-KLD	CEM-BCM/BDM	CEM-BCC/BDC	LCMV-BCM/BDM	LCMV-BCC/BDC
K = 5	**0.6809**	0.6663	0.6715	0.6541	0.6419	0.4876	0.5737	0.4709	0.4603
K = 10	**0.6753**	0.6644	0.6696	0.6469	0.6154	0.4907	0.5496	0.4653	0.4653
K = 15	**0.6840**	0.6745	0.6680	0.6393	0.6404	0.4892	0.5406	0.4990	0.4668
K = 20	**0.6776**	0.6715	0.6773	0.6446	0.6369	0.4856	0.5377	0.4881	0.4801

The accuracy reaches 78.75%, with 20 bands, which is not far from the 79.69%, with all bands. Moreover, when the number of bands reaches 40, the accuracy is 80.14% and sometimes higher, with more bands.

On the Salinas image, as shown in Fig. 8 and Table VII, DCCA is generally with the best performance, and DCWCA and OCBBS sometimes also get the comparable results as far as the classification accuracy is concerned. On this image, for different classifiers, the advantage of the proposed band selection methods is obvious. Most of the time, the classification performance is already acceptable with ten bands selected by our methods. When the number of bands increases, the accuracy generally rises, but it does not change too much. However, for the other comparative band selection methods, the performance is not fine, particularly for the constraint-based methods. CEM-BCM/BDM is always the poorest, which we think is not suitable for this image.

For the Pavia University image, the superiority of the proposed methods is also obvious. When the bands are few, the dual clustering based methods are with the best performance, as shown in Table VIII. When the number of bands increases, CBBS-based methods sometimes slightly surpass DCCA, DCWCA, and OCBBS in Fig. 8(b) and (d). However, this result does not cast doubt to the advantage of our methods, for the reason that the performance of classification with fewer bands is more convincing in the band selection. Another phenomenon particularly distinct on this image is that, with the increase of band number, the accuracy of Naive Bayes reduces. This is because, for Naive Bayes, the labels of pixels mainly are inferred by the comparison between the training samples and the testing samples. However, the number of bands does not influence the training pixel number. While for the other classification methods, more bands bring more information that benefits the classification process. Moreover, we can conclude that LCMV-based methods lacked consistency. For Indian Pine and Salinas scenes, the performance of this method is poor. However, for Pavia University, the result is barely satisfying. To our regret, the bad performance of CEM-based methods still does not ameliorate on this image.

In summary, although the performances of the proposed methods on different images and classifiers differ from each other, in general the advantage of DCCA can still be concluded. As shown in Figs. 6–8 and Tables VI–VIII, the accuracy values of classifications with DCCA selected bands, most of the time, are the best or nearly the best. We also have to emphasize that, although the result on Indian Pines is generally inferior to the other two images, DCCA is still appropriate for Indian Pines. It can be found that DCCA, most of the time, performs well compared with the competitors, particularly when the number of the selected bands is small.

E. Discussion

There are also some abnormal phenomena in the experiments that we would analyze one by one. On Indian Pines, the advantage of dual-clustering-based methods is apparent, when the SVM and the naive Bayes method are used for classification. However, for the other two classification methods CART and kNN, the dual-clustering-based methods are only neck and neck with the others. Even for the kNN method, the generally inferior CEM-based methods sometimes achieve the best accuracy. This abnormal phenomenon for kNN classification also appears on the Pavia University image. The main reason is that the kNN classification is also a kind of clustering method, which is conducted on spatial neighbors. The clustering-based band selection method conducts the cluster process on spectral neighbors that are perpendicular to the spatial plane. The advantage of clustering-based band selection methods cannot be reflected.

Another abnormal phenomenon is that, when the number of bands is large, the dual-clustering-based methods sometimes cannot take the leading position. The reason is that, for these methods, when the bands are 40 or more, the correlation among

bands increases faster than discrimination. From this point, we can conclude that the best number of selected bands for dual-clustering-based methods is 25 to 40. For these 25 to 40 selected bands, the discrimination is high, and the correlation is the lowest.

The third abnormal phenomenon is that, among DCCA, DCWCA, and OCBBS, the DCCA sometimes is not the best. The reason lies in that, in very rare cases, the context may include too much unnecessary information that may affect the performance of DCCA. At the same time, the two views of the dual clustering sometimes cannot bring nice mutual effects to each other. In this situation, the OCBBS may outperform the other two ones. However, this phenomenon is very rare.

A general surprising phenomenon for all methods is that the best performances do not always exist in the result with the most bands. Moreover, the accuracy of classification, most of the time, does not rise with the increase of band number. The curve only waves as the band number changes. The correlation and distinction among bands bring this phenomenon. We cannot always strike a balance between these two limitations by the varying of the number of bands. Therefore, more bands do not mean higher accuracy, while a reasonable number of bands will result in the best accuracy.

V. Conclusion

In this paper, a novel technique named dual-clustering-based HSI classification by context analysis (DCCA) is proposed. The main work is to select the most discriminative bands, which can effectively represent the original HSI. By this means, we can reduce the redundant information of HSI and still enable a high classification accuracy.

The proposed method introduces dual clustering to HSI band selection, based on the context information. The context information is mainly included in the novel PHA feature descriptor, which can be viewed as the first contribution. The descriptor can take the spatial and spectral information into consideration, simultaneously, and effectively embody the local discriminative statistics of HSI. The second contribution lies in the dual clustering process, which clusters the context and the raw feature together. The two clustering processes enhance their respective performances by mutual effects and output the final augmented result. The last contribution is the representative selection from each cluster. Different from the existing clustering-based band selection methods, our method chooses the representative of each cluster groupwisely. We notice that the selected bands should not only be the bands that can represent their corresponding clusters best but also the bands with the lowest correlation between each other if we take the selected representatives as a whole. This groupwise technique is particularly suitable for our dual clustering band selection framework.

As an unsupervised band selection method, DCCA is robust and effective, which has been verified in the experiments. Extensive comparisons also demonstrate the superiority over the traditional competitors.

However, problems also exist. As for time cost and computation complexity, the dual clustering process is not as efficient as k-means clustering. To be more specific, this process may take tens of seconds. The solution to this drawback is our objective for the future work.

References

[1] B. Luo, C. Yang, J. Chanussot, and L. Zhang, "Crop yield estimation based on unsupervised linear unmixing of multidate hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 162–173, Jan. 2013.

[2] V. E. Brando and A. G. Dekker, "Satellite hyperspectral remote sensing for estimating estuarine and coastal water quality," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1378–1387, Jun. 2003.

[3] L. Zhi, D. Zhang, J.-q. Yan, Q.-L. Li, and Q.-L. Tang, "Classification of hyperspectral medical tongue images for tongue diagnosis," *Comput. Med. Imag. Graph.*, vol. 31, no. 8, pp. 672–678, Dec. 2007.

[4] Y. Yuan, G. Zhu, and Q. Wang, "Hyperspectral band selection by multitask sparsity pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 631–644, Feb. 2015.

[5] M. Ghamary Asl, M. R. Mobasheri, and B. Mojaradi, "Unsupervised feature selection using geometrical measures in prototype space for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 3774–3787, Jul. 2014.

[6] Q. Zhang, L. Zhang, Y. Yang, Y. Tian, and L. Weng, "Local patch discriminative metric learning for hyperspectral image feature extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 3, pp. 612–616, Mar. 2014.

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

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

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

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

[11] M. Imani and H. Ghassemian, "Boundary based discriminant analysis for feature extraction in classification of hyperspectral images," in *Proc. Int. Symp. Telecommun.*, 2014, pp. 424–429.

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

[13] A. M. Uso, F. Pla, J. M. Sotoca, and P. Garcia-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, Dec. 2007.

[14] C.-I. Chang, Q. Du, T.-L. Sun, and M. L. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2631–2641, Nov. 1999.

[15] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, Jul. 2001.

[16] P. Toivanen, O. Kubasova, and J. Mielikainen, "Correlation-based band-ordering heuristic for lossless compression of hyperspectral sounder data," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 1, pp. 50–54, Jan. 2005.

[17] N. Keshava, "Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1552–1565, Jul. 2004.

[18] J.-H. Kim, C. N. Georghiades, and G. M. Huang, "Adaptive data transmission based on band-selection for MC-CDMA systems," in *Proc. IEEE Global Telecommun. Conf.*, 2001, vol. 5, pp. 3125–3129.

[19] Q. Du, "Band selection and its impact on target detection and classification in hyperspectral image analysis," in *Proc. IEEE Workshop Adv. Tech. Anal. Remotely Sensed Data*, 2003, pp. 374–377.

[20] H. Yang, Q. Du, H. Su, and Y. Sheng, "An efficient method for supervised hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 138–142, Jan. 2011.

[21] S. Li, H. Wu, D. Wan, and J. Zhu, "An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine," *Knowl.-Based Syst.*, vol. 24, no. 1, pp. 40–48, Feb. 2011.

[22] C. I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006.
[23] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang, "Mutual-information-based semi-supervised hyperspectral band selection with high discrimination, high information, and low redundancy," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2956–2969, May 2015.
[24] J. Feng, L. Jiao, X. Zhang, and T. A. Sun, "Hyperspectral band selection based on trivariate mutual information and clonal selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4092–4105, Jul. 2014.
[25] K. Sun, X. Geng, and L. Ji, "Exemplar component analysis: A fast band selection method for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 998–1002, May 2015.
[26] J. C. Wu and G. C. Tsuei, "Unsupervised cluster-based band selection for hyperspectral image classification," in *Proc. Int. Conf. Adv. Comput. Sci. Electron. Inf.*, 2013, pp. 562–565.
[27] H. Su, H. Yang, Q. Du, and Y. Sheng, "Semisupervised band clustering for dimensionality reduction of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 6, pp. 1135–1139, Nov. 2011.
[28] K. Sun, X. Geng, and L. Ji, "A new sparsity-based band selection method for target detection of hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 12, no. 2, pp. 329–333, Feb. 2015.
[29] C. I. Chang, "Target signature-constrained mixed pixel classification for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 5, pp. 1065–1081, May 2002.
[30] C. I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. New York, NY, USA: Plenum, 2003.
[31] A. M. Uso, F. Pla, J. M. Sotoca, and P. Garcia-Sevilla, "Clustering-based multispectral band selection using mutual information," in *Proc. IEEE Int. Conf. Pattern Recog.*, 2006, vol. 2, pp. 760–763.
[32] S. Jones and L. Shao, "Unsupervised spectral dual assignment clustering of human actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 604–611.
[33] H. Li, W. S. Lee, K. Wang, R. Ehsani, and C. Yang, "Extended Spectral Angle Mapping (ESAM)for citrus greening disease detection using airborne hyperspectral imaging," *Precis. Agric.*, vol. 15, no. 2, pp. 162–183, Apr. 2014.
[34] M. A. Cho, R. Mathieu, and P. Debba, "Multiple endmember spectral-angle-mapper (SAM) analysis improves discrimination of savanna tree species," in *Proc. Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, 2009, pp. 1–4.
[35] S. Kumar, J. Ghosh, and M. M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1368–1379, Jul. 2001.
[36] S. Parasad and L. Mann Bruce, "Limitations of principal components analysis for hyperspectral target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 625–629, Oct. 2008.
[37] A. Karami, M. Yazdi, and G. Mercier, "Compression of hyperspectral images using discrete wavelet transform and Tucker decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 444–450, Apr. 2012.
[38] M. D. Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2010.
[39] L. Zhang and L. Zhang, "Improvement of remote sensing classification method by multiway support tensor machine," in *Proc. Int. Conf. Multimedia Technol.*, 2011, pp. 387–390.
[40] L. Zhang, L. Zhang, D. Tao, and X. Huang, "A multifeature tensor for remote-sensing target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 374–378, Mar. 2011.
[41] X. H. Dang and J. Bailey, "Generation of alternative clusterings using the CAMI approach," in *Proc. SIAM Int. Conf. Data Mining*, 2010, pp. 118–129.
[42] J. Cheng, L. Wang, and C. Leckie, "Dual clustering for categorization of action sequences," in *Proc. IEEE Int. Conf. Pattern Recog.*, 2008, pp. 1–4.
[43] A. Erturk, M. K. Gullu, and S. Erturk, "Hyperspectral image classification using empirical mode decomposition with spectral gradient enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2787–2798, May 2013.
[44] J. Li, H. Zhang, Y. Huang, and L. Zhang, "Hyperspectral image classification by nonlocal joint collaborative representation with a locally adaptive dictionary," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3707–3719, Jun. 2014.
[45] J. Chen, C. Wang, and R. Wang, "Using stacked generalization to combine SVMs in magnitude and shape feature spaces for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2193–2205, Jul. 2009.
[46] L. Jiang, H. Zhang, and Z. Cai, "A novel Bayes model: Hidden naive Bayes," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 10, pp. 1361–1371, Oct. 2009.
[47] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[48] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. New York, NY, USA: Chapman & Hall, 1984.
[49] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[50] C. C. Chang and C. J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[51] 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.

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

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

His current research interests include computer vision and machine learning.

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

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