Semisupervised Local Discriminant Analysis for Feature Extraction in Hyperspectral Images

Wenzhi Liao, *Student Member, IEEE*, Aleksandra Pižurica, *Member, IEEE*, Paul Scheunders, *Member, IEEE*, Wilfried Philips, *Senior Member, IEEE*, and Youguo Pi

***Abstract*—We propose a novel semisupervised local discriminant analysis method for feature extraction in hyperspectral remote sensing imagery, with improved performance in both ill-posed and poor-posed conditions. The proposed method combines unsupervised methods (local linear feature extraction methods and supervised method (linear discriminant analysis) in a novel framework without any free parameters. The underlying idea is to design an optimal projection matrix, which preserves the local neighborhood information inferred from unlabeled samples, while simultaneously maximizing the class discrimination of the data inferred from the labeled samples. Experimental results on four real hyperspectral images demonstrate that the proposed method compares favorably with conventional feature extraction methods.**

***Index Terms*—Classification, feature extraction, hyperspectral remote sensing, semisupervised.**

I. Introduction

HYPERSPECTRAL sensors collect information as a set of images represented by hundreds of spectral bands. While offering much richer spectral information than regular RGB and multispectral images, this large number of spectral bands creates also a challenge for traditional spectral data processing techniques. Conventional classification methods perform poorly on hyperspectral data due to the curse of dimensionality (i.e., the Hughes phenomenon [1]: for a limited number of training samples, the classification accuracy decreases as the dimension increases). Feature extraction aims at reducing the dimensionality of hyperspectral data while keeping as much intrinsic information as possible. Relatively few bands can represent most information of the hyperspectral images [2], making feature extraction very useful for classification, detection, and visualization of remote sensing data [2]–[5].

A number of approaches exist for feature extraction of hyperspectral images [2]–[4], [6], [7], ranging from unsupervised methods to supervised ones. One of the best known unsupervised methods is principle component analysis (PCA) [8], which is widely used for hyperspectral images [2], [9], [10]. Recently, some local methods, which preserve the properties of local neighborhoods were proposed to reduce the dimensionality of hyperspectral images [2], [11]–[13], such as locally linear embedding [12], Laplacian eigenmap [14], and local tangent space alignment [15]. Their linear approximations, such as neighborhood preserving embedding (NPE) [16], locality preserving projection (LPP) [17], and linear local tangent space alignment (LLTSA) [18] were recently applied to feature extraction in hyperspectral images [2], [19]. By considering neighborhood information around the data points, these local methods can preserve local neighborhood information and detect the manifold embedded in the high-dimensional feature space.

Supervised methods rely on the existence of labeled samples to infer class separability. Two widely used supervised feature extraction methods for hyperspectral images are the linear discriminant analysis (LDA) [20] and nonparametric weighted feature extraction (NWFE) [7]. Many extensions to these two methods have been proposed in recent years, such as modified Fisher's LDA [21], regularized LDA [6], modified nonparametric weight feature extraction using spatial and spectral information [22], and kernel NWFE [23].

In real-world applications, labeled data are usually very limited, and labeling large amounts of data may sometimes require considerable human resources or expertise. On the other hand, unlabeled data are available in large quantities at very low cost. For this reason, semisupervised methods [5], [24]–[29], which aim at improved classification by utilizing both unlabeled and limited labeled data gained popularity in the machine learning community. Some of the representative semisupervised learning methods include cotraining [26] and transductive support vector machine (SVM) [27], [28], and Graph-based semisupervised learning methods [25], [29]. Some semisupervised feature extraction methods add a regularization term to preserve certain potential properties of the data. For example, semisupervised discriminant analysis (SDA) [30] adds a regularizer into the objective function of LDA. The resulting method makes use of a limited number of labeled samples to maximize the class discrimination and employs both labeled and unlabeled samples to preserve the local properties of the data. The approach of [31] proposed a general semisupervised dimensionality reduction framework based on pairwise

Manuscript received May 1, 2011; revised September 16, 2011 and February 17, 2012; accepted April 8, 2012. This work was supported by the China Scholarship Council, FWO project G.0123.08N "Spatial extension for classification of multispectral images" and SBO-IWT project Chameleon: Domain-specific Hyperspectral Imaging Systems for Relevant Industrial Applications. A. Pižurica is a postdoctoral researcher of the Fund for Scientific Research in Flanders (FWO).

W. Liao is with the Department of Telecommunications and Information Processing, Ghent University, 9000 Ghent, Belgium, and also with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: wenzhi.liao@telin.ugent.be).

A. Pižurica and W. Philips are with the Department of Telecommunications and Information Processing, Ghent University, 9000 Ghent, Belgium (e-mail: sanja@telin.ugent.be; Wilfried.Philips@telin.ugent.be).

P. Scheunders is with Department of Physics, University of Antwerp, B-2610 Antwerp, Belgium (e-mail: paul.scheunders@ua.ac.be).

Y. Pi is with School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: auygpi@scut.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.2012.2200106

TABLE I
SOME NOTATIONS USED IN THIS PAPER

Notation	Description	Notation	Description
N	number of total training samples	r	number of features in the mapped space
d	dimension of the input space (bands)	C	number of classes
n	number of labeled training samples	$\mathbf{S}_t \in \Re^{d\times d}$	total scatter matrix
n_k	number of labeled samples from class k	u	number of unlabeled samples
$\mathbf{x} \in \Re^d$	vector of d features (bands)	$\mathbf{X} \in \Re^{d\times N}$	data matrix of all training samples
$\mathbf{X}^{(k)} \in \Re^{d\times n_k}$	data matrix of samples in class k	e	number of nearest neighbors
$\mathbf{u}_k$	mean vector of class k	$\mathbf{u}$	mean vector of all training samples
$\mathbf{w} \in \Re^r$	mapped vector of r features	$W \in \Re^{d\times r}$	transformation matrix
$\mathbf{S}_b \in \Re^{d\times d}$	between scatter matrix	$\mathbf{S}_w \in \Re^{d\times d}$	winthin scatter matrix
$\mathbf{x}_i^{(k)} \in \Re^d$	ith sample in the kth class	y_i	label of the sample $\mathbf{x}_i$

constraints, which employs regularization with sparse representation. Other semisupervised feature extraction methods combine supervised methods with unsupervised ones using a tradeoff parameter, such as semisupervised local Fisher discriminant analysis (SELF) [32]. However, it may not be easy to specify the optimal parameter values in these and similar semisupervised techniques, as mentioned in [31] and [32].

In this paper, we propose a novel semisupervised local discriminant analysis (SELD) method to reduce the dimensionality of the hyperspectral images. The proposed SELD method aims to find a projection which can preserve local neighborhood information and maximize the class discrimination of the data. We combine an unsupervised method (from the class of local linear feature extraction methods (LLFE), such as NPE, LPP and LLTSA) and a supervised method LDA in a novel framework without any tuning parameters. Contrasting to related semisupervised methods, such as SELF [32], we do not combine supervised and unsupervised methods linearly. Instead of using both labeled and unlabeled samples together, we first divide the samples into two sets: labeled and unlabeled. Then, we employ the labeled samples through the supervised method (LDA) only and the unlabeled ones through an unsupervised, locality preserving method (LLFE) only. Preliminary results for one specific instance of the present approach were presented in [33]. In this paper, we give much more extended theoretical motivation, we show a more general framework where different LLFE methods can be combined with the supervised one, and we present much more extended experimental evaluation.

We propose a natural way to combine unsupervised and supervised methods without any free parameters, making fully the use of strengths of both approaches in different scenarios. The supervised component maximizes class discrimination (for the available number of labeled samples), and the local unsupervised component ensures neighborhood information preservation. While we employ the LLFE [16]–[18] and LDA [20] methods, this novel framework can be applied in combination with other supervised and unsupervised methods too. Another advantage is that our method can extract as many features as the number of spectral bands. This also increases classification accuracy with respect to methods where the number of extracted features is limited by the number of classes (LDA and SDA). The results demonstrate improved classification accuracy when compared to related semisupervised methods.

The organization of the paper is as follows. Section II provides a brief review of existing feature extraction methods that are most relevant for this work. In Section III, we present the proposed SELD method. The experimental results on four hyperspectral images are presented and discussed in Section IV. Finally, the conclusions of the paper are drawn in Section V.

II. BACKGROUND AND RELATED WORK

Let $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \Re^d$ denote high-dimensional data, $\{\mathbf{z}_i\}_{i=1}^N$, and $\mathbf{z}_i \in \Re^r$ its low-dimensional representations with $r \leq d$. In our application, d is the number of spectral bands of hyperspectral images, and r is the dimensionality of the projected subspace. The assumption is that there exists a mapping function $f : \Re^d \rightarrow \Re^r$, which can map every original data point $\mathbf{x}_i$ to $\mathbf{z}_i = f(\mathbf{x}_i)$ such that most information of the high-dimensional data is kept in a much lower dimensional projected subspace. This mapping is usually represented by a $d \times r$ projection matrix $\mathbf{W}$

$$\mathbf{z}_i = f(\mathbf{x}_i) = \mathbf{W}^T\mathbf{x}_i. \tag{1}$$

In many feature extraction methods, the projection matrix $\mathbf{W}$ can be obtained by solving the following optimization problem, where $\mathbf{W}$ denotes one of the vectors in the projection matrix $\mathbf{W}$:

$$\mathbf{w}_{opt} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\overline{\mathbf{S}}\mathbf{w}}{\mathbf{w}^T\underline{\mathbf{S}}\mathbf{w}}. \tag{2}$$

The matrices $\overline{\mathbf{S}}$ and $\underline{\mathbf{S}}$ have specific meaning in different methods as we discuss later in the text. The solution to (2) is equivalent to solving the following generalized eigenvalue problem:

$$\overline{\mathbf{S}}\mathbf{w} = \lambda\underline{\mathbf{S}}\mathbf{w}. \tag{3}$$

Or equivalently

$$\underline{\mathbf{S}}^{-1}\overline{\mathbf{S}}\mathbf{w} = \lambda\mathbf{w}. \tag{4}$$

The projection matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$ is made up by the r eigenvectors of the matrix $\underline{\mathbf{S}}^{-1}\overline{\mathbf{S}}$ associated with the largest r eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$. Some notations used throughout this paper are summarized in Table I.

A. Unsupervised Methods

Unsupervised feature extraction methods deal with the cases where no labeled samples are available, aiming to find another representation of the data in the lower dimensional space by satisfying some given criterion. In particular, LLFE [15]–[17] methods reviewed in [34] seek a projection direction on which neighborhood data points in the high-dimensional feature space $\Re^d$ are kept on neighborhood in the low-dimensional projected subspace $\Re^r$ as well. By considering neighborhood information around the data points, the goal of these methods is to preserve the local properties of the original data. We will employ these methods in our approach. Although the LLFE methods in [16]–[18] have some characteristic differences [16], [18], they are all linear approximations to local nonlinear feature extraction methods and share more or less the same technique of linearization. The optimal solution of all these three methods can be computed by eigendecomposition.

We explain here in some more detail NPE [16] as one of the unsupervised LLFE methods. For details on LPP and LLTSA, the readers should consult [17], [18]. NPE first finds e nearest neighbors (eNN) for each data point $\mathbf{x}_i$. The eNN is determined first by calculating the distance (we use Euclidean distance here) between data point $\mathbf{x}_i$ and all the data points, then sorting the distance and determining nearest neighbors based on the eth minimum distance. Then, the reconstruction weights Q_{ij} are calculated by minimizing the reconstruction error, which results from approximating $\mathbf{x}_i$ by its e nearest neighbors

$$\min \textstyle\sum_i \left\| \mathbf{x}_i - \sum_{j=1}^{e} Q_{ij}\mathbf{x}_j \right\|^2 \qquad S.t. \quad \sum_{j=1}^{e} Q_{ij} = 1. \quad (5)$$

The extracted features $\mathbf{z}_i$ in the low-dimensional projected subspace that best preserve the local neighborhood information are then obtained as

$$\min \textstyle\sum_i \left\| \mathbf{z}_i - \sum_{j=1}^{e} Q_{ij}\mathbf{z}_j \right\|^2 \qquad S.t. \quad \mathbf{z}_i^T\mathbf{z}_i = \mathbf{I}. \quad (6)$$

The projection matrix $\mathbf{W}_{NPE} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$ can be optimized as follows:

$$\mathbf{w}_{NPE} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{w}} \quad (7)$$

where $\mathbf{M} = (\mathbf{I} - \mathbf{Q})^T(\mathbf{I} - \mathbf{Q})$ and $\mathbf{I}$ represents the identity matrix.

We can express the optimal projection matrix of all LLFE methods from [16]–[18] in a unified way as follows:

$$\mathbf{w}_{LLFE} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\overline{\mathbf{C}}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}\underline{\mathbf{C}}\mathbf{X}^T\mathbf{w}}. \quad (8)$$

For NPE [16], $\overline{\mathbf{C}} = \mathbf{I}$ and $\underline{\mathbf{C}} = \mathbf{M}$. For LPP [17], $\overline{\mathbf{C}} = \mathbf{D}$ and $\underline{\mathbf{C}} = \mathbf{L}$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{L}$ is the Laplacian matrix [17]. For LLTSA [18], $\overline{\mathbf{C}} = \mathbf{I}$ and $\underline{\mathbf{C}} = \mathbf{B}$, where $\mathbf{B}$ is the alignment matrix [18]. By setting $\overline{\mathbf{S}} = \mathbf{X}\overline{\mathbf{C}}\mathbf{X}^T$ and $\underline{\mathbf{S}} = \mathbf{X}\underline{\mathbf{C}}\mathbf{X}^T$, we obtain the projection matrix $\mathbf{W}_{LLFE} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$ as in (2). The reasoning behind LLFE is that neighboring points in the high-dimensional space $\Re^d$ are likely to have similar representation in the low-dimensional projected subspace $\Re^r$ as well. Therefore, LLFE methods preserve the local neighborhood information of the data in the low-dimensional representation.

B. Supervised Methods

The best known supervised method is LDA [20], which seeks projection directions on which the ratio of the *between-class* covariance to *within-class* covariance is maximized. Taking the label information into account, LDA results in a linear transformation $\mathbf{z}_i = f(\mathbf{x}_i, y_i) = \mathbf{W}^T\mathbf{x}_i$, where y_i is the label of the data point $\mathbf{x}_i$. The corresponding projection matrix $\mathbf{W}_{LDA} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$ is optimized as follows:

$$\mathbf{w}_{LDA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \quad (9)$$

where

$$\mathbf{S}_b = \sum_{k=1}^{C} n_k \left(\mathbf{u}^{(k)} - \mathbf{u}\right)\left(\mathbf{u}^{(k)} - \mathbf{u}\right)^T \quad (10)$$

$$\mathbf{S}_w = \sum_{k=1}^{C} \left(\sum_{i=1}^{n_k} \left(\mathbf{x}_i^{(k)} - \mathbf{u}^{(k)}\right)\left(\mathbf{x}_i^{(k)} - \mathbf{u}^{(k)}\right)^T\right) \quad (11)$$

where n_k is the number of samples in the kth class, $\mathbf{u}$ is the mean of the entire training set, $\mathbf{u}^{(k)}$ is the mean of the kth class, $\mathbf{x}_i^{(k)}$ is the ith sample in the kth class. $\mathbf{S}_b$ is called the *between-class* scatter matrix and $\mathbf{S}_w$ the *within-classs* scatter matrix. (9) is equivalent to

$$\mathbf{w}_{LDA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_t\mathbf{w}} \quad (12)$$

with

$$\mathbf{S}_t = \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{u})(\mathbf{x}_i - \mathbf{u})^T \quad (13)$$

form (10), (11), and (13), we have $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$.

By setting $\overline{\mathbf{S}} = \mathbf{S}_b$ and $\underline{\mathbf{S}} = \mathbf{S}_w$ or $\underline{\mathbf{S}} = \mathbf{S}_t$, we obtain the projection matrix $\mathbf{W}_{LDA} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$ as in (2). LDA seeks projection direction on which the data points within the same class are close while separating all the data points from different classes apart. However, as the rank of the *between-class* scatter matrix $\mathbf{S}_b$ is $C - 1$, LDA can extract at most $C - 1$ features, which may not be sufficient to represent essential information of the original data.

C. Semisupervised Methods

Recently, semisupervised feature extraction methods have been proposed and applied to pattern recognition [30], [32]. The idea behind these methods is to infer the class discrimination from labeled samples, as well as the local neighborhood information from both labeled and unlabeled samples. SDA [30] imposes a regularizer in LDA

$$\mathbf{w}_{SDA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_t\mathbf{w} + \alpha J(\mathbf{w})} \quad (14)$$

where $J(\mathbf{w}) = \mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w}$ is the constraint based on graph Laplacian regularization [30], and $\mathbf{L}$ is the Laplacian matrix [17]. The parameter α controls the influence of local neighborhood information; for $\alpha = 0$, SDA reduces to LDA. SDA has the same limitation as LDA in the number of extracted features, because the rank of the *between-class* matrix $\mathbf{S}_b$ is $C-1$.

SELF [32] combines linearly PCA and LFDA [35]

$$\mathbf{w}_{SELF} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\left[(1-\beta)\mathbf{S}_{lb}+\beta\mathbf{S}_t\right]\mathbf{w}}{\mathbf{w}^T\left[(1-\beta)\mathbf{S}_{lw}+\beta\mathbf{I}\right]\mathbf{w}} \tag{15}$$

where $\mathbf{S}_{lb}$ and $\mathbf{S}_{lw}$ are local *between-class* scatter matrix and local *within-class* scatter matrix [35], $\beta(\in[0,1])$ is a trade off parameter, which controls the contribution of the supervised method LFDA and unsupervised method PCA. By setting β to a value between zero and one, SELF can separate samples from different classes while maximizing the variance of the data inferred from both the labeled and unlabeled samples. SELF overcomes some limitations of LDA and SDA (it can extract as much features as the number of the dimensions).

III. Proposed Semisupervised Local Discriminant Analysis (SELD)

In this section, we propose a novel semisupervised feature extraction method for hyperspectral images, which we call SELD. As discussed above, some semisupervised methods, such as SDA and SELF, can achieve a good class discrimination and preserve the local properties of the data with properly optimized parameters. One important issue is how to optimize tuning parameters, which is common to some related semisupervised methods [31], [32]. One solution is to employ cross-validation for this purpose. However, except for the computational cost of parameter optimization, cross-validation is not reliable when the number of labeled samples is small [35] (which is sometimes the real case in hyperspectral images). Focusing on class discrimination, LDA is in general well suited to preprocessing for the task of classification, since the transformation improves class separation. However, when only a small number of labeled samples are available, LDA tends to perform poorly due to overfitting. LLFE works directly on the data without any ground truth and incorporates the local neighborhood information of data points in its feature extraction process.

Motivated by these facts, we propose a novel semisupervised approach, which combines LLFE and LDA methods in a way that adapts automatically to the fraction of the labeled samples without any parameters. The main idea of our approach is to first divide the samples into two sets: labeled and unlabeled. The labeled samples will be used only by LDA (to maximize the class discrimination), and the unlabeled ones only through LLFE (to preserve the local neighborhood information). This will yield a natural way to combine the two as we show next.

Suppose a training data set $\mathbf{X}$ is made up of the labeled set $\mathbf{X}_{labeled}=\{(\mathbf{x}_i,y_i)\}_{i=1}^{n}$, $y_i\in\{1,2,\ldots,C\}$, where C is the number of classes, and the unlabeled set $\mathbf{X}_{unlabeled}=\{\mathbf{x}_i\}_{i=n+1}^{N}$ with u unlabeled samples, $N=n+u$, $\mathbf{X}=\{\mathbf{X}_{labeled},\mathbf{X}_{unlabeled}\}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_{n+1},\ldots,\mathbf{x}_N\}$. The kth class has n_k samples with $\sum_{k=1}^{C} n_k=n$. Without loss of generality, we center the data points by subtracting the mean vector from all the sample vectors, and assume that the labeled samples in $\mathbf{X}_{labeled}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ are ordered according to their labels, with the data matrix of the kth class $\mathbf{X}^{(k)}=\{\mathbf{x}_1^{(k)},\mathbf{x}_2^{(k)},\ldots,\mathbf{x}_{n_k}^{(k)}\}$ where $\mathbf{x}_i^{(k)}$ is the ith sample in the kth class. Then, the labeled set can be expressed as $\mathbf{X}_{labeled}=\{\mathbf{X}^{(1)},\mathbf{X}^{(2)},\ldots,\mathbf{X}^{(C)}\}$. We have

$$\begin{aligned}\mathbf{S}_b' &= \sum_{k=1}^{C} n_k\left(\mathbf{u}^{(k)}\right)\left(\mathbf{u}^{(k)}\right)^T\\ &= \sum_{k=1}^{C} n_k\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_i^{(k)}\right)\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_i^{(k)}\right)^T\\ &= \sum_{k=1}^{C}\mathbf{X}^{(k)}\mathbf{P}^{(k)}\left(\mathbf{X}^{(k)}\right)^T\end{aligned}$$

where $\mathbf{P}^{(k)}$ is the $n_k\times n_k$ matrix with all the elements equal to $\frac{1}{n_k}$. If we define a $n\times n$ matrix $\mathbf{P}_{n\times n}$ as

$$\mathbf{P}_{n\times n}=\begin{bmatrix}\mathbf{P}^{(1)} & \mathbf{0} & \ldots & \mathbf{0}\\ \mathbf{0} & \mathbf{P}^{(2)} & \ldots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{P}^{(C)}\end{bmatrix}$$

the *between-class* scatter matrix $\mathbf{S}_b'$ can be written as

$$\mathbf{S}_b'=\sum_{k=1}^{C}\mathbf{X}^{(k)}\mathbf{P}^{(k)}\left(\mathbf{X}^{(k)}\right)^T=\mathbf{X}_{labeled}\mathbf{P}_{n\times n}(\mathbf{X}_{labeled})^T. \tag{16}$$

By subtracting the *between-class* scatter matrix from the total scatter matrix $\mathbf{S}_t'$, the *within-class* scatter matrix $\mathbf{S}_w'$ is obtained as

$$\begin{aligned}\mathbf{S}_w' &= \mathbf{S}_t'-\mathbf{S}_b'\\ &= \mathbf{X}_{labeled}(\mathbf{X}_{labeled})^T-\mathbf{X}_{labeled}\mathbf{P}_{n\times n}(\mathbf{X}_{labeled})^T\\ &= \mathbf{X}_{labeled}(\mathbf{I}_{n\times n}-\mathbf{P}_{n\times n})(\mathbf{X}_{labeled})^T.\end{aligned} \tag{17}$$

In our approach, the LDA component will use the labeled samples only (to maximize the class discrimination), so we reformulate (9) as

$$\begin{aligned}\mathbf{w}_{LDA}' &= \arg\max_{\mathbf{w}}\frac{\mathbf{w}^T\mathbf{S}_b'\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w'\mathbf{w}}\\ &= \arg\max_{\mathbf{w}}\frac{\mathbf{w}^T\mathbf{X}_{labeled}\mathbf{P}_{n\times n}(\mathbf{X}_{labeled})^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}_{labeled}(\mathbf{I}_{n\times n}-\mathbf{P}_{n\times n})(\mathbf{X}_{labeled})^T\mathbf{w}}.\end{aligned} \tag{18}$$

Equivalently, we reformulate the optimization problem of LLFE in (8), so that it only uses the unlabeled samples

$$\mathbf{w}_{LLFE}'=\arg\max_{\mathbf{w}}\frac{\mathbf{w}^T\mathbf{X}_{unlabeled}\overline{\mathbf{C}}_{u\times u}(\mathbf{X}_{unlabeled})^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}_{unlabeled}\underline{\mathbf{C}}_{u\times u}(\mathbf{X}_{unlabeled})^T\mathbf{w}}. \tag{19}$$

We define the following matrics:

$$\mathbf{P}=\begin{bmatrix}\mathbf{P}_{n\times n} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\quad \bar{\mathbf{I}}=\begin{bmatrix}\mathbf{I}_{n\times n} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$$

$$\underline{\mathbf{C}}=\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \underline{\mathbf{C}}_{u\times u}\end{bmatrix}\quad \overline{\mathbf{C}}=\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \overline{\mathbf{C}}_{u\times u}\end{bmatrix}$$

Now, the reformulated optimization problems of LDA and LLFE in (18) and (19) can be written as follows:

$$\mathbf{w}'_{LDA} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\mathbf{P}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}(\bar{\mathbf{I}}-\mathbf{P})\mathbf{X}^T\mathbf{w}} \tag{20}$$

$$\mathbf{w}'_{LLFE} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\overline{\mathbf{C}}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}\underline{\mathbf{C}}\mathbf{X}^T\mathbf{w}}. \tag{21}$$

Note that full data vector $\mathbf{X}$ appears in (20) and (21), but due to the structure of the matrices $\mathbf{P}$, $\bar{\mathbf{I}}$, $\underline{\mathbf{C}}$, and $\overline{\mathbf{C}}$, the LDA (20) makes use of the labeled samples only, and LLFE (21) makes use of the unlabeled samples only. In order to make full use of the strengths of both two methods without parameter optimization, we propose a natural way to combine them as

$$\mathbf{w}_{SELD} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\overline{\mathbf{S}}_{SELD}\mathbf{w}}{\mathbf{w}^T\underline{\mathbf{S}}_{SELD}\mathbf{w}} \tag{22}$$

where

$$\begin{aligned}\overline{\mathbf{S}}_{SELD} =& \mathbf{X}_{labeled}\mathbf{P}_{n\times n}(\mathbf{X}_{labeled})^T\\ &+\mathbf{X}_{unlabeled}\overline{\mathbf{C}}_{u\times u}(\mathbf{X}_{unlabeled})^T\\ =&[\mathbf{X}_{labeled},\mathbf{X}_{unlabeled}]\begin{bmatrix}\mathbf{P}_{n\times n} & \mathbf{0}\\ \mathbf{0} & \overline{\mathbf{C}}_{u\times u}\end{bmatrix}\\ &\times[\mathbf{X}_{labeled},\mathbf{X}_{unlabeled}]^T\\ =&\mathbf{X}(\mathbf{P}+\overline{\mathbf{C}})\mathbf{X}^T \end{aligned} \tag{23}$$

$$\begin{aligned}\underline{\mathbf{S}}_{SELD} =& \mathbf{X}_{labeled}(\mathbf{I}_{n\times n}-\mathbf{P}_{n\times n})(\mathbf{X}_{labeled})^T\\ &+\mathbf{X}_{unlabeled}\underline{\mathbf{C}}_{u\times u}(\mathbf{X}_{unlabeled})^T\\ =&[\mathbf{X}_{labeled},\mathbf{X}_{unlabeled}]\begin{bmatrix}\mathbf{I}_{n\times n}-\mathbf{P}_{n\times n} & \mathbf{0}\\ \mathbf{0} & \underline{\mathbf{C}}_{u\times u}\end{bmatrix}\\ &\times[\mathbf{X}_{labeled},\mathbf{X}_{unlabeled}]^T\\ =&\mathbf{X}\left((\bar{\mathbf{I}}-\mathbf{P})+\underline{\mathbf{C}}\right)\mathbf{X}^T. \end{aligned} \tag{24}$$

The resulting method combines supervised and unsupervised components in a nonlinear way, making fully the use of their strengths in different scenarios. In the case when all the samples are labeled, the proposed method reduces to LDA, and in the case when all the samples are unlabeled, it reduces to LLFE.

To obtain the projection matrix, we solve the generalized eigenvalue problem of the proposed SELD method, which is equivalent to (3)

$$\overline{\mathbf{S}}_{SELD}\mathbf{w} = \lambda\underline{\mathbf{S}}_{SELD}\mathbf{w}. \tag{25}$$

Through its nonlinear combination of supervised and unsupervised components, the proposed SELD seeks a projection direction on which the local neighborhood information of the data can be best preserved, while simultaneously the class discrimination is maximal.

It is important to note that LDA confronts sometimes with the difficulty that the matrix $\mathbf{S}_w$ is singular. The fact is that sometimes the number of labeled training samples n is much smaller than the number of dimensions d. In this situation, the rank of $\mathbf{S}_w$ is at most n as it is evident from (17), while the size of the matrix $\mathbf{X}(\bar{\mathbf{I}}-\mathbf{P})\mathbf{X}^T$ in (20) is $d \times d$. This implies that the within-class matrix $\mathbf{S}_w$ can become singular. Simultaneously, the *between-class* matrix $\mathbf{S}_b$ in the LDA method uses the labeled samples only. The rank of $\mathbf{S}_b$ is $C-1$ [as it can be seen from (16)], implying that LDA can extract at most $C-1$ features, which is not always sufficient to represent essential information of the original data.

The proposed SELD method overcomes these problems. The matrices $\overline{\mathbf{S}}_{SELD}$ and $\underline{\mathbf{S}}_{SELD}$ in our approach are both symmetric and positive semidefinite, which makes sure that SELD can extract as much features as the number of the spectral bands and the corresponding eigenvalues are not negative. Since our method can be combined with different LLFE methods, we will use a subscript to identify the particular LLFE methods employed, e.g., $SELD_{NPE}$, $SELD_{LPP}$, or $SELD_{LLTSA}$.

A. Algorithm

The algorithmic procedure of the proposed SELD is formally stated below:

1) Divide the training set $\mathbf{X}$ into two subsets: $\mathbf{X}_{labeled}$ and $\mathbf{X}_{unlabeled}$, with $\mathbf{X} = \{\mathbf{X}_{labeled}, \mathbf{X}_{unlabeled}\} = \{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_{n+1},\ldots,\mathbf{x}_N\}$. Suppose that the n labeled training samples in $\mathbf{X}_{labeled} = \{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ are ordered according to their labels, with data matrix of the kth class $\mathbf{X}^{(k)} = \{\mathbf{x}_1^{(k)},\ldots,\mathbf{x}_{n_k}^{(k)}\}$ where $\mathbf{x}_i^{(k)}$ is the ith sample in the kth class, then the labeled subset can be expressed as $\mathbf{X}_{labeled} = \{\mathbf{X}^{(1)},\mathbf{X}^{(2)},\ldots,\mathbf{X}^{(C)}\}$. $u = N-n$ unlabeled samples constitute the unlabeled subset $\mathbf{X}_{unlabeled} = \{\mathbf{x}_i\}_{i=n+1}^N$.
2) Construct the labeled weight matrices $\mathbf{P}$ and $\bar{\mathbf{I}}$ from the labeled subset $\mathbf{X}_{labeled}$.
3) Construct the "nearest neighbors" weight matrix $\overline{\mathbf{C}}$ and $\underline{\mathbf{C}}$ from the unlabeled subset $\mathbf{X}_{unlabeled}$. The particular construction depends on the chosen LLFE methods. For NPE: $\overline{\mathbf{C}} = \mathbf{I}$ and $\underline{\mathbf{C}} = \mathbf{M}$; for LPP: $\overline{\mathbf{C}} = \mathbf{D}$ and $\underline{\mathbf{C}} = \mathbf{L}$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{L}$ is the Laplacian matrix [17]; for LLTSA: $\overline{\mathbf{C}} = \mathbf{I}$ and $\underline{\mathbf{C}} = \mathbf{B}$, where $\mathbf{B}$ is the alignment matrix [18].
4) Compute the eigenvectors and eigenvalues for the generalized eigenvector problem in (25). The projection matrix $\mathbf{W}_{SELD} = (\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_r)$ is made up by the r eigenvectors of the matrix $\underline{\mathbf{S}}_{SELD}^{-1}\overline{\mathbf{S}}_{SELD}$ associated with the largest r eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$.
5) SELD embedding: project the original d dimensional data into a lower r dimensional subspace by

$$\mathbf{x} \to \mathbf{z} = \mathbf{W}_{SELD}^T\mathbf{x}.$$

In many real-world applications, it is usually difficult, expensive, and time-consuming to collect sufficient amount of labeled samples. Meanwhile, it is much easier to obtain unlabeled samples. In this case, the within-class scatter $\mathbf{S}_w$ becomes very small, the eigendecomposition becomes inaccurate. Our approach in (22) overcomes this problem: if a small number of the labeled samples is available, the projection matrix is estimated through LLFE using a large number of unlabeled samples. Simultaneously, the available labeled samples are fully used to maximize the class discrimination. In the opposite case, when more labeled samples are available, the LDA dominates. In this way, the proposed SELD magnifies the advantages of LDA

TABLE II
COMPARISON OF SDA, SELF, AND THE PROPOSED SELD METHOD

	SDA	SELF	The proposed SELD method
Main idea	Add a regularizer in the objective function of LDA as (14)	Linearly combine LFDA and PCA as (15)	Combine LLFE and LDA in a novel framework as (22)
How to preserve neighborhood information	The added regularizer uses both labeled and unlabeled samples	LFDA uses labeled samples	LLFE uses unlabeled samples
How to maximize class discrimination	LDA uses labeled samples	LFDA uses labeled samples	LDA uses labeled samples
Parameters	*one*	*one*	*none*
Maximal features	$C-1$	d	d

TABLE III
DATA SETS USED IN THE EXPERIMENTS

No	Indian Pine		KSC		DC		Botswana	
	Class Name	# Samples	Class Name	# Samples	Class Name	# Samples	Class Name	# Samples
1	Corn-notill	1434	Scrub	761	Roof	3834	Water	270
2	Corn-min	834	Willow swamp	243	Street	416	Hippo grass	101
3	Corn	234	Cabbage palm hammock	256	Path	175	Floodplain grasses1	251
4	Grass/Pasture	497	Cabbage palm/oak hammock	252	Grass	1928	Floodplain grasses2	215
5	Grass/Trees	747	Slash pine	161	Trees	405	Reeds1	269
6	Hay-windrowed	489	Oak/broadleaf hammock	229	Water	1224	Riparian	269
7	Soybeans-notill	968	Hardwood swamp	105	Shadow	97	Firescar2	259
8	Soybeans-min	2468	Graminoid marsh	431			Island interior	203
9	Soybeans-clean	614	Spartina marsh	520			Acacia woodlands	314
10	Wheat	212	Cattail marsh	404			Acacia shrublands	248
11	Woods	1294	Salt marsh	419			Acacia grasslands	305
12	Bldg-Grass-Trees	380	Mud flats	503			Short mopane	181
13	Stone-steel towers	95	Water	927			Mixed mopane	268
14							Exposed soils	95
Total		10266		5211		8079		3248

and LLFE, and compensates for disadvantages of the two at the same time. The main differences between the proposed SELD method and related semisupervised methods SDA [30] and SELF [32] are summarized in Table II.

IV. EXPERIMENTAL RESULTS

A. Hyperspectral Image Data Sets

We use four real hyperspectral data sets in our experiments: the *Indian Pine* (a mixed forest/agricultural site in Indiana [36]), *Kennedy Space Center* (KSC) [37], the *Washington DC Mall* [36] (urban site), and *Okavango Delta, Botswana* [37]. Table III shows the number of labeled samples in each class for all the data sets. Note that the color in the cell denotes different classes in the classification maps (Figs. 2–5).

Indian Pine data set was captured by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over northwestern Indiana in June 1992, with 220 spectral bands in the wavelength range from 0.4 to 2.5 μm and spatial resolution of 20 m by pixel. The calibrated data are available online (along with detailed groundtruth information) from http://cobweb.ecn.purdue.edu/~biehl/. The whole scene, consisting of the full 145 × 145 pixels, contains 16 classes, ranging in size from 20 to 2468 pixels. Thirteen classes were selected for the experiments (see Fig. 2).

KSC data set was acquired by NASA AVIRIS instrument over the KSC, Florida in 1996 and consists of 224 bands of 10-nm width with center wavelengths from 0.4–2.5 μm. The data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. Several spectral bands were removed from the data due to noise and water absorption phenomena, leaving a total of 176 bands to be used for the analysis. For classification purposes, 13 classes representing the various land cover types that occur in this environment were defined for the site, Fig. 3 shows an RGB composition with the labeled classes highlighted. For more information, see [37] and http://www.csr.utexas.edu/hyperspectral/.

DC Mall data set was collected with an airborne sensor system over the Washington DC Mall, with 1280 × 307 pixels and 210 spectral bands in the 0.4–2.4 μm region. This data set consists of 191 spectral bands after elimination of water absorption and noisy bands and is available at http://cobweb.ecn.purdue.edu/~biehl/. Seven land cover/use classes are labeled and are highlighted in the Fig. 4.

Botswana data set was acquired over the Okavango Delta, Botswana in May 31, 2001 by the NASA EO-1 satellite, with 30-m pixel resolution over a 7.7-km strip in 242 bands covering the 0.4–2.5 μm portion of the spectrum in 10-nm windows. Uncalibrated and noisy bands that cover water absorption features were removed, leaving a total of 145 radiance channels to be used in the experiments. The data consist of observations from 14 identified classes intended to reflect the impact of flooding on vegetation, Fig. 5 shows an RGB composition with the labeled classes highlighted. For more information, see [37] and http://www.csr.utexas.edu/hyperspectral/.

TABLE IV
HIGHEST OCA USING EXTRACTED FEATURES (THE NUMBER OF EXTRACTED FEATURES IS WRITTEN IN THE BACK BRACKETS) APPLIED TO FOUR DIFFERENT DATA SETS IN CASE 1

Feature Extraction	Classifier	Data Set			
		Indian Pine	KSC	DC	Botswana
Raw	QDC	0.14	0.146	0.474	0.084
	1NN	0.524	0.728	0.965	0.835
	SVM	0.475	0.846	0.948	0.876
PCA	QDC	0.568(5)	0.703(6)	0.969(3)	0.836(4)
	1NN	0.52(20)	0.726(19)	0.965(12)	0.833(20)
	SVM	0.583(6)	0.808(16)	0.946(2)	0.878(5)
LDA	QDC	0.14(4)	0.146(5)	0.474(4)	0.124(4)
	1NN	0.108(12)	0.30(12)	0.409(4)	0.151(10)
	SVM	0.129(6)	0.393(12)	0.476(5)	0.218(13)
NPE	QDC	0.521(6)	0.71(5)	0.969(3)	0.833(4)
	1NN	0.596(15)	0.84(16)	0963(6)	0.873(7)
	SVM	0.633(12)	0.839(18)	0.966(13)	0.895(9)
LPP	QDC	0.523(6)	0.731(5)	0.97(4)	0.795(4)
	1NN	0.612(10)	0.833(12)	0.966(7)	0.848(5)
	SVM	0.643(10)	0.84(20)	0.957(10)	0.867(11)
LLTSA	QDC	0.56(5)	0.666(3)	0.969(3)	0.815(5)
	1NN	0.563(20)	0.816(14)	0.965(2)	0.864(5)
	SVM	0.604(20)	0.82(19)	0.967(2)	0.898(7)
NWFE	QDC	0.574(5)	0.763(5)	0.967(3)	0.828(4)
	1NN	0.661(10)	0.833(18)	0.97(17)	0.881(17)
	SVM	0.624(7)	0.858(17)	0.957(2)	0.891(8)
SDA	QDC	0.413(5)	0.68(5)	0.889(5)	0.704(5)
	1NN	0.539(10)	0.817(12)	0.857(6)	0.77(13)
	SVM	0.483(7)	0.811(12)	0.817(6)	0.811(6)
SELF	QDC	0.568(5)	0.703(6)	0.969(3)	0.836(4)
	1NN	0.52(20)	0.726(19)	0.965(12)	0.833(20)
	SVM	0.583(6)	0.808(16)	0.946(2)	0.878(5)
$SELD_{NPE}$	QDC	0.551(7)	0.771(4)	0.965(3)	0.826(4)
	1NN	**0.698(18)**	0.863(20)	0.974(20)	0.903(20)
	SVM	0.648(12)	**0.874(19)**	0.959(18)	0.905(9)
$SELD_{LPP}$	QDC	0.541(5)	0.758(5)	0.969(4)	0.793(4)
	1NN	0.656(16)	0.844(20)	**0.976(15)**	0.873(18)
	SVM	0.645(11)	0.857(20)	0.959(3)	0.876(7)
$SELD_{LLTSA}$	QDC	0.531(5)	0.755(4)	0.953(4)	0.829(4)
	1NN	0.667(20)	0.852(20)	0.964(8)	0.899(19)
	SVM	0.642(18)	0.833(19)	0.948(12)	**0.91(9)**

B. Experimental Setup

The training set $\mathbf{X}$ is made up of labeled subset $\mathbf{X}_{labeled}$ and unlabeled subset $\mathbf{X}_{unlabeled}$ (such that $\mathbf{X} = \mathbf{X}_{labeled} \cup \mathbf{X}_{unlabeled}$, and $\mathbf{X}_{labeled} \cap \mathbf{X}_{unlabeled} = \emptyset$). A number of unlabeled samples $u = 1500$ was randomly selected from the image parts with no labels to compose $\mathbf{X}_{unlabeled}$. The training of the classifiers (estimation of the SVM parameters) was carried out using the labeled subset $\mathbf{X}_{labeled}$. In our experiments, 70% randomly chosen samples from the labeled data set was initially assigned to the training set and the remaining 30% was used as the test set. In order to investigate the influence of the training set size on the classifier performance, the initial training set (consisting of 70% of the labeled samples) was further subsampled randomly to compose the labeled subset $\mathbf{X}_{labeled}$, with sample size conforming to one of the following two distinct cases:

- *Case* 1 $(n_k = 10)$ in ill-posed condition: $n < d$ and $n_k < d$.
- *Case* 2 $(n_k = 40)$ in poor-posed condition: $n > d$ and $n_k < d$.

We used three common classifiers: 1-nearest neighbor (1NN) like in [19], [23], [31], quadratic discriminant classifier (QDC) [38], and SVM [39]. The SVM classifier with radial basis function (RBF) kernels in Matlab SVM Toolbox, LIBSVM [40], is applied in our experiments. SVM with RBF kernels has two parameters: the penalty factor C and the RBF kernel widths γ. We apply a grid search on C and γ using fivefold cross-validation to find the best C within the given set $\{10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ and the best γ within the given set $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$.

All classifiers were evaluated against the test set. We use overall classification accuracy (OCA) to evaluate the feature extraction results. The results were averaged over ten runs, we compare the resulting classification accuracies using the proposed SELD method with those resulting from the following methods: Raw data, where the classification is simply performed on the original data sets without dimensionality reduction; PCA [8]; LDA [20]; LLFE [15]–[17] (including NPE [16], LPP [17], LLTSA [18]); NWFE [7]; SDA [30], of which the parameter α is optimized with fivefold cross-validation within the given set {0.1, 0.5, 2.5, 12.5, 62.5}; and SELF [32], where the parameter β is chosen from {0, 0.1, 0.2, . . ., 0.9, 1} by fivefold cross validation.

C. Numerical Comparison

Tables IV and V display the classification accuracies of testing data in cases 1, 2, respectively. The best accuracy of each data set (in column) is highlighted in bold font.

From these tables, we have the following findings:

1) The results confirm that feature extraction can improve the classification performance on hyperspectral images. Most information can be preserved even with a few

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
HIGHEST OCA USING EXTRACTED FEATURES (THE NUMBER OF EXTRACTED FEATURES IS WRITTEN IN THE BACK BRACKETS) APPLIED TO FOUR DIFFERENT DATA SETS IN CASE 2

Feature Extraction	Classifier	Data Set			
		Indian Pine	KSC	DC	Botswana
Raw	QDC	0.14	0.146	0.474	0.084
	1NN	0.65	0.818	0.983	0.902
	SVM	0.622	0.924	0.983	0.931
PCA	QDC	0.736(10)	0.853(17)	0.997(7)	0.937(8)
	1NN	0.646(20)	0.816(20)	0.983(14)	0.90(17)
	SVM	0.717(5)	0.896(19)	0.99(12)	0.94(7)
LDA	QDC	0.601(10)	0.864(10)	0.954(6)	0.909(6)
	1NN	0.621(11)	0.881(11)	0.975(6)	0.932(12)
	SVM	0.604(9)	0.895(12)	0.98(6)	0.901(8)
NPE	QDC	0.738(11)	0.87(20)	0.997(17)	0.941(8)
	1NN	0.687(13)	0.889(20)	0.988(13)	0.941(8)
	SVM	0.757(13)	0.916(20)	0.987(15)	0.945(8)
LPP	QDC	0.727(12)	0.891(13)	0.996(19)	0.927(12)
	1NN	0.71(10)	0.886(13)	0.985(12)	0.93(7)
	SVM	0.751(10)	0.92(20)	0.989(3)	0.925(12)
LLTSA	QDC	0.749(11)	0.872(18)	0.997(15)	0.935(7)
	1NN	0.644(19)	0.884(16)	0.982(7)	0.932(6)
	SVM	0.753(20)	0.908(18)	0.984(2)	0.932(6)
NWFE	QDC	0.752(9)	0.871(16)	0.997(13)	0.943(10)
	1NN	0.767(12)	0.87(20)	0.99(15)	0.921(19)
	SVM	0.775(8)	0.924(18)	0.988(16)	0.938(8)
SDA	QDC	0.636(9)	0.885(12)	0.993(6)	0.915(11)
	1NN	0.655(12)	0.897(11)	0.969(6)	0.939(12)
	SVM	0.637(9)	0.898(12)	0.978(6)	0.905(12)
SELF	QDC	0.736(10)	0.853(17)	0997(7)	0.937(8)
	1NN	0.646(20)	0.816(20)	0.983(14)	0.90(17)
	SVM	0.717(5)	0.896(19)	0.99(12)	0.94(7)
$SELD_{NPE}$	QDC	0.74(12)	0.906(9)	0.997(13)	0.935(8)
	1NN	**0.792(20)**	0.924(20)	0.992(20)	**0.951(18)**
	SVM	0.747(13)	**0.936(19)**	**0.998(12)**	0.948(9)
$SELD_{LPP}$	QDC	0.742(12)	0.904(15)	0.997(13)	0.931(13)
	1NN	0.785(12)	0.918(19)	0.993(19)	938(12)
	SVM	0.76(12)	0.931(18)	0.99(17)	0.933(11)
$SELD_{LLTSA}$	QDC	0.734(9)	0.911(11)	0.997(12)	0.945(11)
	1NN	0.779(19)	0.913(9)	0.992(18)	0.947(20)
	SVM	0.757(20)	0.925(19)	0.98(14)	0.949(13)

extracted features. In particular, for the raw data set with QDC classifier, the results can be improved a lot by using feature extraction as a preprocessing. SVM classifier with RBF kernel function did not perform well in the raw data set of Indian Pine, this can be improved by using feature extraction.

2) When the number of labeled samples is very limited such as in *Case* 1, the supervised LDA performs much worse than other methods. By considering the local neighborhood information inferred from both labeled and unlabeled samples, SDA improves over LDA. However, one limitation of both LDA and SDA methods is that the number of extracted features depends on the number of classes.
3) By selecting $\beta = 1$ optimized with fivefold cross-validation within the given set {0, 0.1, 0.2, . . ., 0.9, 1}, SELF performs as PCA in both cases. For the Botswana data set with QDC classifier in *Case* 1, SELF and PCA give a better performance when small number of bands are used, while for the KSC data set in *Case* 2 (Fig. 1), SELF and PCA perform worse than other methods when small number of bands are used. It should be noted though that for a small number of features, the OCAs are usually very small and useless in practice.
4) The proposed SELD outperforms the other feature extraction methods in both cases. In the ill-posed classification problems (*Case* 1, $n_k = 10 < n < d$), the highest OCA in Indian Pine, KSC, DC Mall, and Botswana data sets are 0.698 ($SELD_{NPE}$ with 1NN classifier), 0.874 ($SELD_{NPE}$ with SVM classifier), 0.976 ($SELD_{LPP}$ with 1NN classifier) and 0.91 ($SELD_{LLTSA}$ with SVM classifier), respectively. In *Case* 2 ($n_k = 40 < d < n$), the highest OCA among for the same four images are 0.792 ($SELD_{NPE}$ with 1NN classifier), 0.936 ($SELD_{NPE}$ with SVM classifier), 0.998 ($SELD_{NPE}$ with SVM classifier) and 0.951 ($SELD_{NPE}$ with 1NN classifier), respectively.

In ill-posed (*Case* 1) and poor-posed (*Case* 2) classification problems, the QDC classifier cannot be developed to the raw data sets since the input dimension is higher than the number of available training samples. In these situations, 1NN and SVM classifier show better performances than QDC. The results in Tables IV and V show that the proposed method yields best OCA on all four data sets.

The experimental results in Tables IV and V also show that none of the three classifiers achieves the highest accuracy on every data set. This can also be seen in Fig. 1. The reason may be that the distributions of data sets are very different as was mentioned in [23], [41], and [42]. In the following, we take the Indian Pine and KSC images in *Case* 2 as examples to explore the performances of different methods when the number of extracted features increases, the results were shown in Fig. 1. The statistical significance of differences was computed using McNemars test, which is based upon the standardized normal test statistic [47]

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}}$$

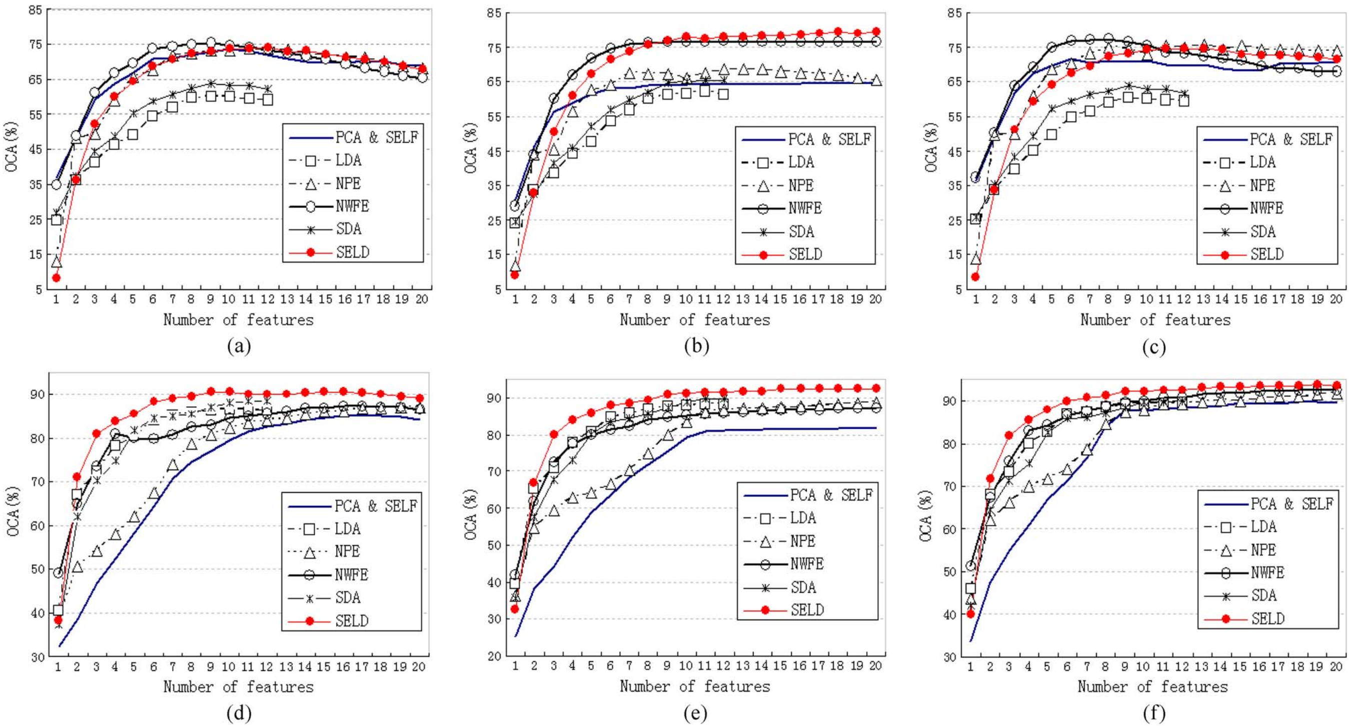


Fig. 1. Performance of each feature extraction method in Case 2 for Indian Pine and KSC data sets. Each experiment was repeated ten times, the average was acquired. By selecting $\beta = 1$ optimized with fivefold cross-validation within the given set {0, 0.1, 0.2, . . ., 0.9, 1}, SELF has the same performance as PCA. The proposed SELD method is the one which combines LDA and NPE. (a) Indian Pine with QDC classifier. (b) Indian Pine with 1NN classifier. (c) Indian Pine with SVM classifier. (d) KSC with QDC classifier. (e) KSC with 1NN classifier. (f) KSC with SVM classifier.

TABLE VI

Statistical Significance of Differences in Classification (Z) With QDC Classifier in Case 2. Each Case of the Table Represents Z_{rc} Where r Is the Row and c Is the Column. The Best Results of Each Method Over Ten Runs Are Used

Z_{rc}	Indian Pine using 9 features							KSC using 12 features						
	PCA	LDA	NPE	NWFE	SDA	SELF	SELD	PCA	LDA	NPE	NWFE	SDA	SELF	SELD
PCA	0	21.6	-1.1	-7.7	18.1	0	-2.4	0	-1.6	-4.8	-5.9	-4.7	0	-9.2
LDA	-21.6	0	-22.6	-27.8	-4.8	-21.6	-24	1.6	0	-2.5	-2.9	-4.3	1.6	-7
NPE	1.1	22.6	0	-6.1	19.1	1.1	-1.2	4.8	2.5	0	-0.5	-0.7	4.8	-5.7
NWFE	7.7	27.8	6.1	0	24.8	7.7	5	5.9	2.9	0.5	0	-0.3	5.9	-4.6
SDA	-18.1	4.8	-19.1	-24.8	0	-18.1	-20.8	4.7	4.3	0.7	0.3	0	4.7	-3.7
SELF	0	21.6	-1.1	-7.7	18.1	0	-2.4	0	-1.6	-4.8	-5.9	-4.7	0	-9.2
SELD	2.4	24	1.2	-5	20.8	2.4	0	9.2	7	5.7	4.6	3.7	9.2	0

TABLE VII

Statistical Significance of Differences in Classification (Z) With 1NN Classifier in Case 2. Each Case of the Table Represents Z_{rc} Where r Is the Row and c Is the Column. The Best Results of Each Method Over Ten Runs Are Used

Z_{rc}	Indian Pine using 9 features							KSC using 12 features						
	PCA	LDA	NPE	NWFE	SDA	SELF	SELD	PCA	LDA	NPE	NWFE	SDA	SELF	SELD
PCA	0	0.9	-9.9	-26.7	-5.8	0	-29	0	-12.2	-13.8	-13.4	-13.7	0	-19
LDA	-0.9	0	-10.3	-22.5	-9.3	-0.9	-28.3	12.2	0	-1.5	2.8	-2.7	12.2	-7.5
NPE	9.9	10.3	0	-13.7	3.5	9.9	-20.2	13.8	1.5	0	4.4	-0.6	13.8	-6
NWFE	26.7	22.5	13.7	0	16.2	26.7	-6.7	13.4	-2.8	-4.4	0	-4.7	13.4	-10.7
SDA	5.8	9.3	-3.5	-16.2	0	5.8	-22.1	13.7	2.7	0.6	4.7	0	13.7	-5.4
SELF	0	0.9	-9.9	-26.7	-5.8	0	-29	0	-12.2	-13.8	-13.4	-13.7	0	-19
SELD	29	28.3	20.2	6.7	22.1	29	0	19	7.5	6	10.7	5.4	19	0

TABLE VIII

Statistical Significance of Differences in Classification (Z) With SVM Classifier in Case 2. Each Case of the Table Represents Z_{rc} Where r Is the Row and c Is the Column. The Best Results of Each Method Over Ten Runs Are Used

Z_{rc}	Indian Pine using 9 features							KSC using 12 features						
	PCA	LDA	NPE	NWFE	SDA	SELF	SELD	PCA	LDA	NPE	NWFE	SDA	SELF	SELD
PCA	0	15.5	-13.8	-17.2	7.8	0	-12.2	0	0.1	-5.7	-7.1	-2.2	0	-11
LDA	-15.5	0	-27.6	-30.5	-10.5	-15.5	-26.6	-0.1	0	-4.8	-7.5	-3.4	-0.1	-11.9
NPE	13.8	27.6	0	-3.9	20.4	13.8	1.9	5.7	4.8	0	-2.2	2.6	5.7	-6.8
NWFE	17.2	30.5	3.9	0	23.6	17.2	5.4	7.1	7.5	2.2	0	5.2	7.1	-6.2
SDA	-7.8	10.5	-20.4	-23.6	0	-7.8	-19.3	2.2	3.4	-2.6	-5.2	0	2.2	-9.8
SELF	0	15.5	-13.8	-17.2	7.8	0	-12.2	0	0.1	-5.7	-7.1	-2.2	0	-11
SELD	12.2	26.6	-1.9	-5.4	19.3	12.2	0	11	11.9	6.8	6.2	9.8	11	0

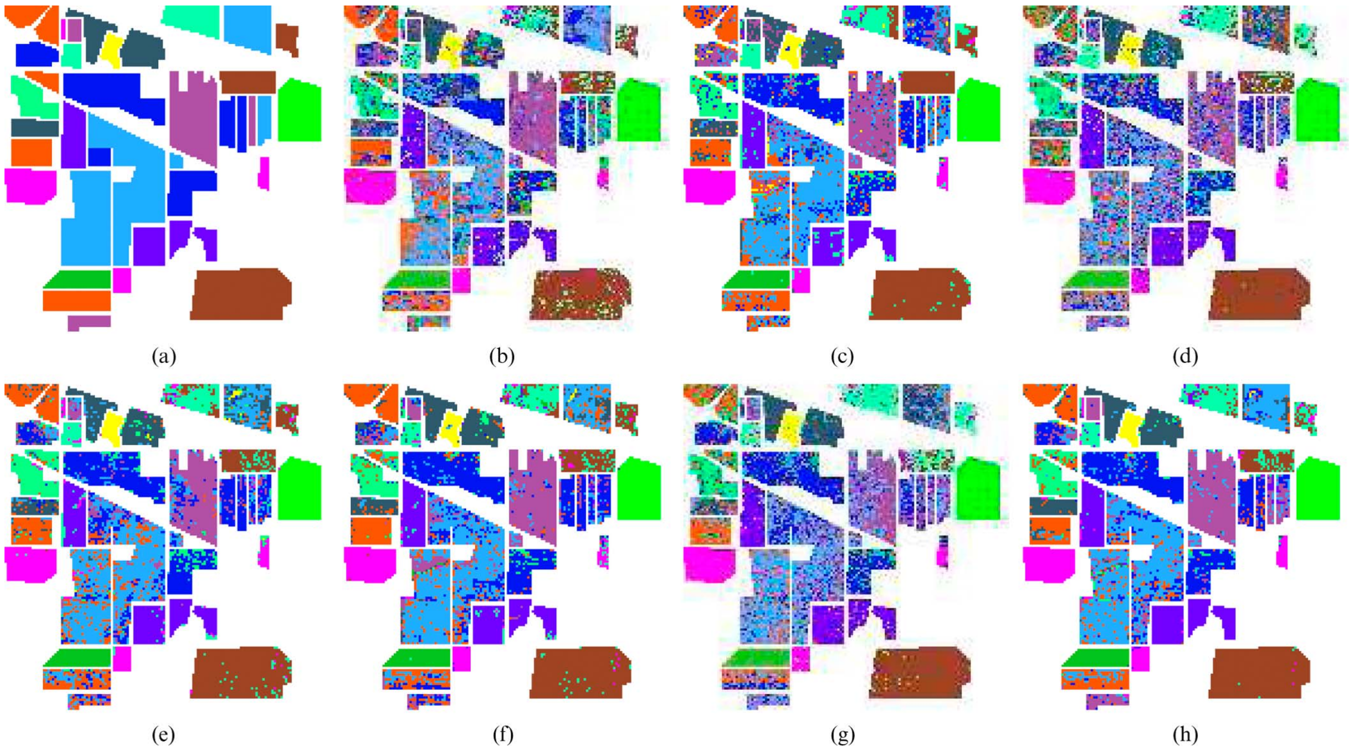


Fig. 2. Classification maps for Indian Pine with $n_k = 40$ (Case 2) (a) Ground truth of the area with 13 classes, and thematic map using (b) 1NN classifier without feature extraction ($r = 220$), (c) PCA and SELF features and QDC classifier ($r = 10$), (d) LDA features and 1NN classifier ($r = 11$), (e) NPE features and SVM classifier ($r = 13$), (f) NWFE features and SVM classifier ($r = 8$), (g) SDA features and 1NN classifier ($r = 12$), and (h) The proposed $SELD_{NPE}$ features and 1NN classifier ($r = 20$).

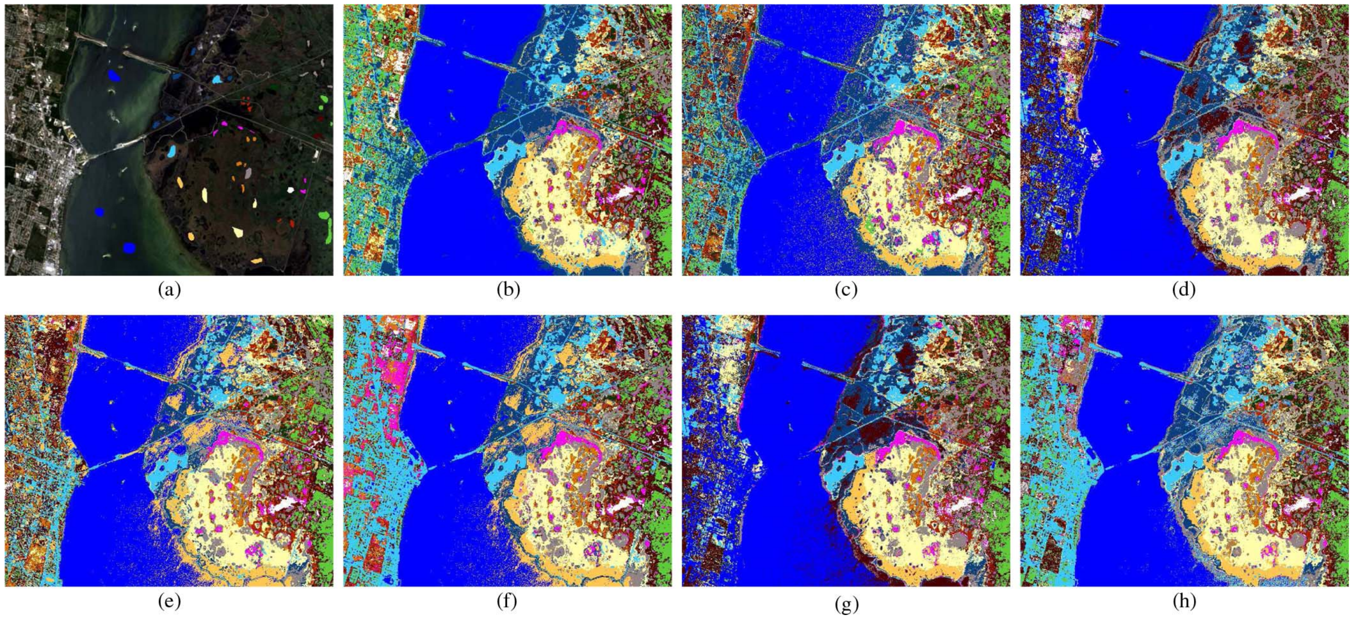


Fig. 3. Classification maps for KSC with $n_k = 40$ (Case 2) (a) RGB composition with 13 classes labeled and highlighted in the image, and thematic map using (b) SVM classifier without feature extraction ($r = 176$), (c) PCA and SELF features and SVM classifier ($r = 19$), (d) LDA features and SVM classifier ($r = 12$), (e) LPP features and SVM classifier ($r = 20$), (f) NWFE features and SVM classifier ($r = 18$), (g) SDA features and SVM classifier ($r = 12$), and (h) The proposed $SELD_{NPE}$ features and SVM classifier ($r = 19$).

where f_{12} indicates the number of samples correctly by classifier 1 and incorrectly by classifier 2. The difference in accuracy between classifiers 1 and 2 is said to be statistically significant if $|Z| > 1.96$. The sign of Z indicates whether classifier 1 is more accurate than classifier 2 ($Z > 0$) or vice versa ($Z < 0$). Tables VI–VIII show the results using the best results of each method in the same bands over ten runs.

1) On Indian Pine data set, NWFE outperforms the other methods for QDC and SVM classifiers, the difference is statistically significant, with $|Z| > 1.96$. For

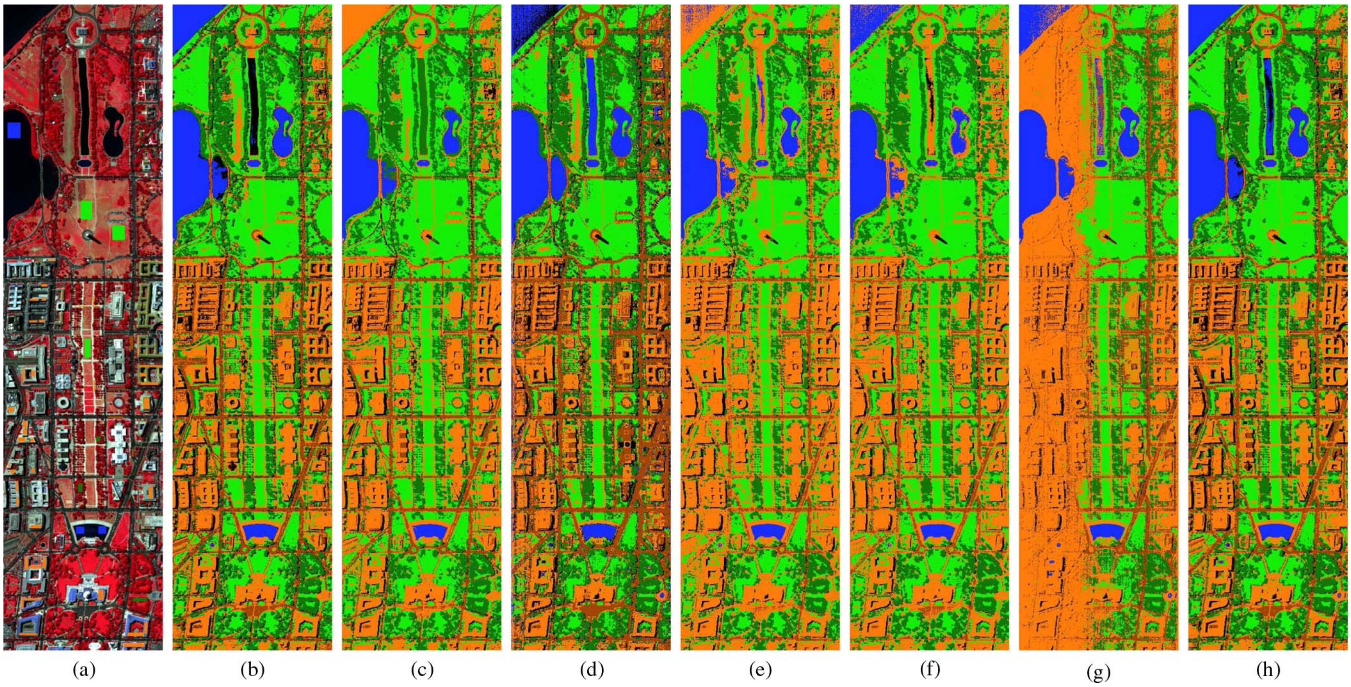


Fig. 4. Classification maps for DC Mall with $n_k = 40$ (Case 2) (a) RGB composition with seven classes labeled and highlighted in the image, and thematic map using (b) SVM classifier without feature extraction ($r = 191$), (c) PCA and SELF features and QDC classifier ($r = 7$), (d) LDA features and SVM classifier ($r = 6$), (e) NPE features and QDC classifier ($r = 17$), (f) NWFE features and QDC classifier ($r = 13$), (g) SDA features and QDC classifier ($r = 6$), and (h) The proposed $SELD_{NPE}$ features and SVM classifier ($r = 12$).

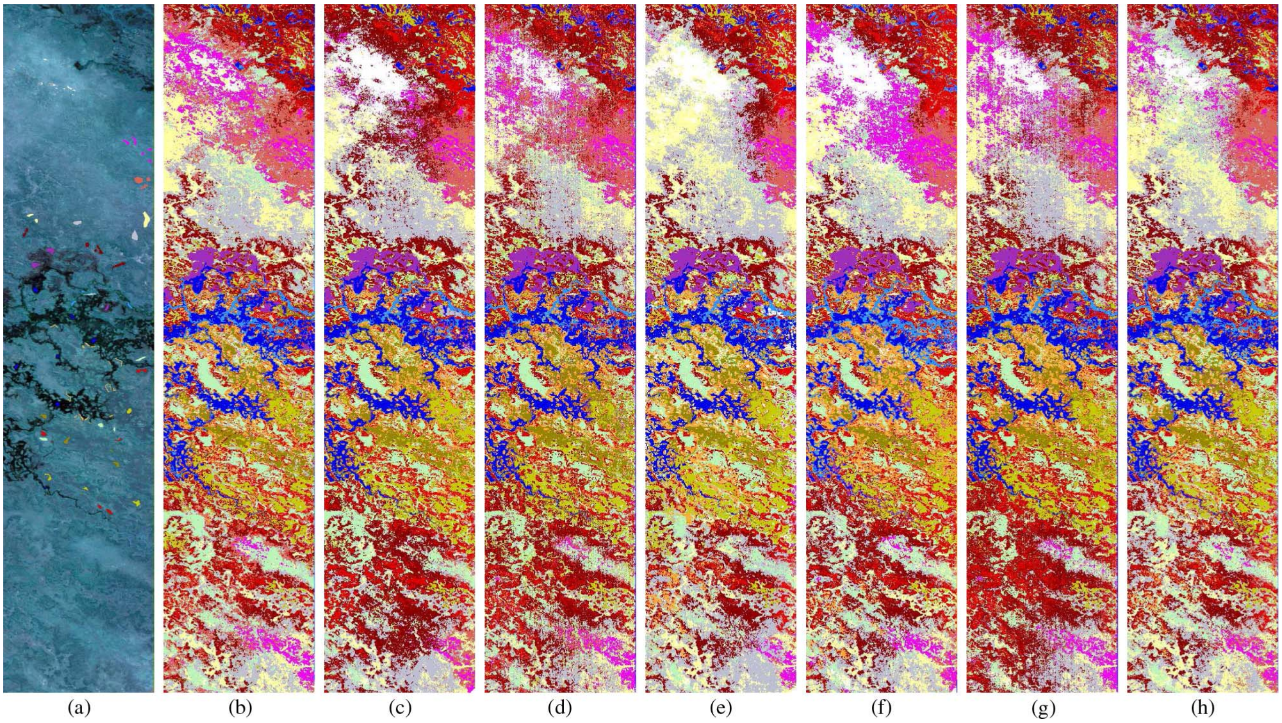


Fig. 5. Classification maps for Okavango Delta, Botswana with $n_k = 40$ (Case 2) (a) RGB composition with 14 classes labeled and highlighted in the image, and thematic map using (b) SVM classifier without feature extraction ($r = 145$), (c) PCA and SELF features and SVM classifier ($r = 7$), (d) LDA features and 1NN classifier ($r = 12$), (e) NPE features and SVM classifier ($r = 8$), (f) NWFE features and QDC classifier ($r = 10$), (g) SDA features and SVM classifier ($r = 6$), and (h) The proposed $SELD_{NPE}$ features and 1NN classifier ($r = 18$).

1NN classifier, the proposed SELD method yields the highest OCA of 79.2%, which is better than NWFE with SVM classifier 77.5%. The difference between the best results of SELD with 1NN classifier and NWFE with SVM classifier is statistically significant $(Z = 3.86)$.

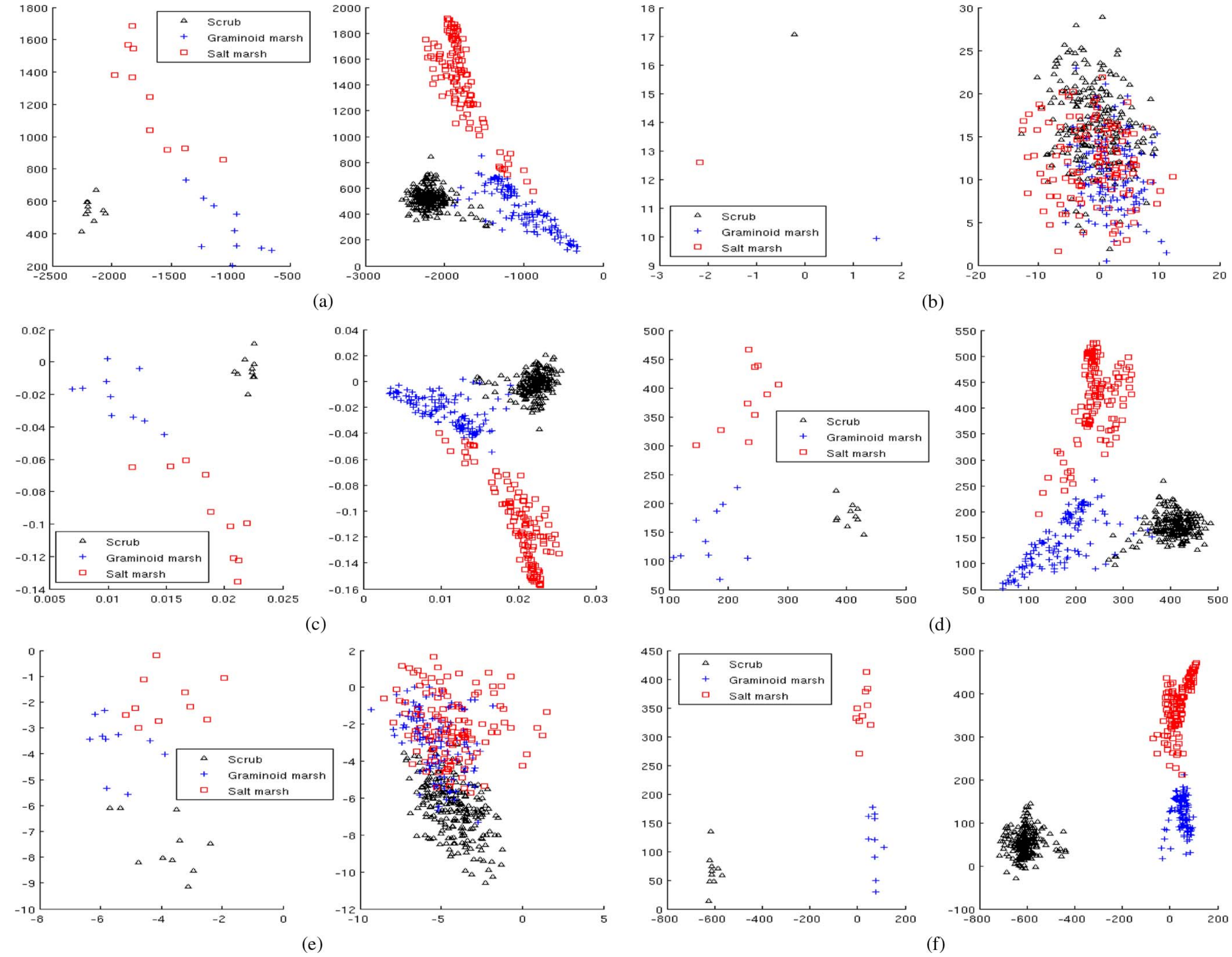


Fig. 6. Distributions of training samples and testing samples for "Scrub," "Graminoid marsh," and "Salt marsh" of KSC data set using the first two significant features obtained from different methods. In each method, the left scatter plot is for training data, and the right one is for testing data ($n_k = 10$, Case 1). (a) PCA and SELF. (b) LDA. (c) NPE. (d) NWFE. (e) SDA. (f) SELD.

2) On KSC data set, SELD performs better than the other methods with all the three classifiers. The statistical difference of accuracy $|Z| > 1.96$ clearly demonstrates the efficiency of the proposed SELD.
3) Using only $C - 1$ features may not be enough in some real situation, which is one limitation of both LDA and SDA. NPE can improve its performance by using more extracted features, as shown in Fig. 1(f). When more features are used, the OCA can be improved.

The results in Tables IV–VIII and in Fig. 1 show that SELD with 1NN classifier can have a better performance in Indian Pine image, while in the KSC image, SELD with SVM classifier will be a better choice.

In order to compare the classified maps visually, we generate classification maps with the combination of the highest OCA using different methods and classifiers in *Case* 2 ($n_k = 40$), displayed in Figs. 2–5. The results demonstrate that:

1) By incorporating the local neighborhood information of the data, SELD preserves well spatial consistency in the classification maps, for example, the "Grass" in DC Mall image (Fig. 4). SELD also produces smoother homogeneous regions in the classification maps, which is particularly significant when classifying the "Stone-steel towers" and "Grass/Trees" in the Indian Pine image (Fig. 2).
2) SELD also yields good class discrimination. For Indian Pine image, it is easy to find that SELD outperforms other feature extraction methods in "Grass/Pasture," "Grass/Trees," "Soybeans-notill," and "Soybeans-clean" parts (Fig. 2). For DC Mall image, SELD discriminates "Water" better than the other methods (Fig. 4).

The plots in Fig. 6 give more insight into class discrimination by different methods. The training and testing samples of three classes of KSC image in *Case* 1 are projected into the feature space formed by the first two eigenvectors of different feature extraction methods. The results in Fig. 6 show that LDA has overfitting problems, because in *Case* 1 ($n < d$, and $n_k < d$), both the *within-classs* scatter matrix $\mathbf{S}_w$ and the *between-class* scatter matrix $\mathbf{S}_b$ are singular, $\mathbf{S}_w$ cannot be inverted, and both $\mathbf{S}_w$ and $\mathbf{S}_b$ are not accurate. By considering the local neighborhood information inferred from both labeled and

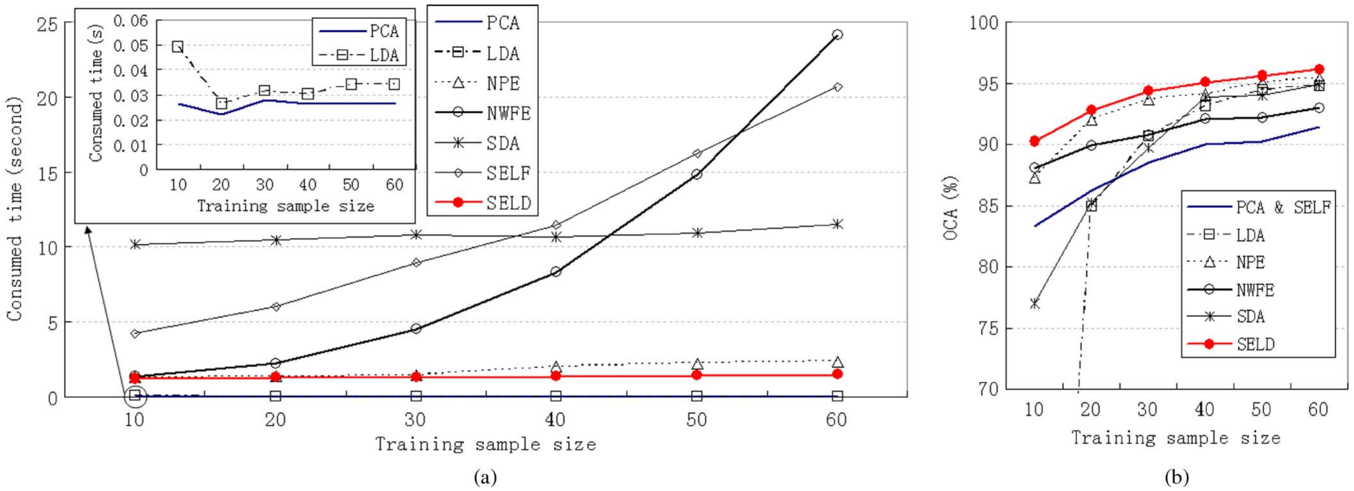


Fig. 7. Comparision of computational time (second) and OCA with different sample size, $e = 12$ and $u = 1500$. The experiments was repeated ten times, the average was acquired. The highest OCA with r changing from 1 to 20 is recorded. (a) Computational cost. (b) OCA.

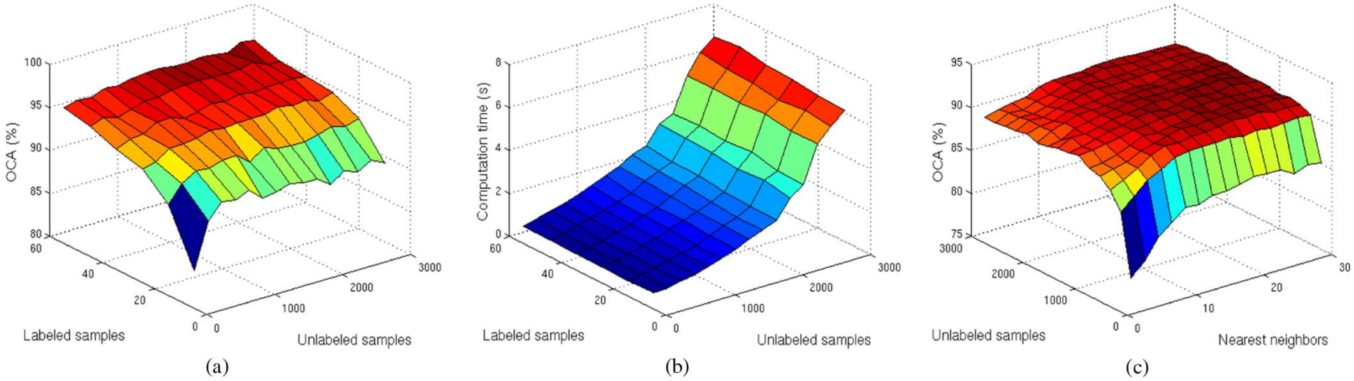


Fig. 8. Surface of (a) the OCA as a function of labeled and unlabeled samples, $r = 13$ and $e = 12$. (b) The computation time as a function of labeled and unlabeled samples, $r = 13$ and $e = 12$. (c) The OCA as a function of unlabeled samples and nearest neighbors, $n_k = 10$ and $r = 13$.

unlabeled samples, SDA improves over LDA, but the test data are projected with different classes mixed. The distributions of projected data obtained by SELD are more concentrated and more distinct as compared with those of PCA, LDA, NPE, NWFE, and SDA. This explains also classification improvement in Tables IV and V.

D. Algorithm Analysis

In this section, we analyze the proposed semisupervised algorithm in terms of the computational cost, the selection of unlabeled samples, and the selection of nearest neighbors.

1) Computational Cost: The computational complexity of the proposed SELD is mainly in finding the e nearest neighbors for all the selected unlabeled training samples. To find the e nearest neighbors for u selected unlabeled training samples in the d dimensional Euclidean space, the complexity is $O(du^2)$. However, some methods can be used to reduce the complexity of searching the e nearest neighbors, such as K-D trees [43]. $SELD_{NPE}$ and $SELD_{LLTSA}$ have additional complexities over $SELD_{LPP}$ in calculating the reconstruction weights, which is $O(due^3)$. For storing the matrix $\overline{\mathbf{C}}$ or $\underline{\mathbf{C}}$ in (21), the complexity is $O(N^2)$, where N is the total training samples including labeled and unlabeled ones. For example, if we use all the samples in the Botswana data set to train, $N = 1476 \times 256$, this will exceed the memory capacity of an ordinary PC even though the matrix is sparse. In order to reduce the computational complexity and memory consumption, some of unlabeled samples were selected in our experiments.

We compared the computational cost of different approaches. All the methods were implemented in Matlab. The experiments were carried out on 64-bit, 2.67-GHz Intel i7 920 (8 core) CPU computer with 12-GB memory, Fig. 7 shows the computational time of different approaches, and the OCA with 1NN classifier. The recorded times were only consumed in the process of feature extraction. This included the time consumed on the parameter determination of some methods (such as α in SDA, and β in SELF). We can see that PCA and LDA are the fastest, and the proposed SELD is more efficient than NWFE, SDA, and SELF as the number of training samples increases. The reason is that the parameter determination in SDA and SELF is time consuming.

2) Selection of Unlabeled Samples: The choice of unlabeled samples is very important step in the semisupervised methods. Selection of too many unlabeled samples will increase computational complexity, while a small number of unlabeled samples is not sufficient to exploit the local neighborhood information of the data sets. One easy solution is selecting unlabeled samples randomly from the whole image. Fig. 8(a) shows an example of the performances with different number of labeled and unlabeled samples. The number of unlabeled samples was evaluated from 200 to 3000 with a step of 200. Fig. 8(b) shows the corresponding computation times. The classification accuracy of SELD will be improved as more unlabeled samples are used, particularly in ill-posed (Case 1) classification problems. Generally, semisupervised methods can achieve better classification results by using more unlabeled samples than labeled ones [45], [46]. However, the usage of a large number of unlabeled samples will cause problems in computational complexity and memory consumption. This may be improved by using some spatial selection methods [44].

3) Selection of Nearest Neighbors: In graph-based feature extraction methods, the number of nearest neighbors (e) is an important parameter. We can employ cross-validation to optimize e. However, we found in our experiments that our approach produces consistently good results over a large range of e values, which suggests insensitivity to this parameter in a broad range. Fig. 8(c) shows the performance with different number of unlabeled samples and nearest neighbors when e is changed from 2 to 30 with a step of 2. Note that the maximal dimensionality of $SELD_{LLTSA}$ was set to $e - 2$ (e should be greater than r [15]).

V. Conclusion

In this paper, we present a new semisupervised feature extraction method, and we apply it to classification of hyperspectral images. The main idea of the proposed method is to divide first the samples into the labeled and the unlabeled sets. The labeled samples are employed through the supervised LDA only and the unlabeled ones through the unsupervised method only. We combine the two in a nonlinear way, which makes full use of the advantages of both approaches. Experimental results on hyperspectral images demonstrate advantages of our method and improved classification accuracy compared to some related feature extraction methods. Moreover, we do not need to optimize any tuning parameters, which makes our method more effecient. Also, the new method removes the limitation of LDA and SDA in terms of the number of extracted features. Future work will include selection of unlabeled samples and a kernelized version of this method.

Acknowledgment

The authors would like to thank Prof. Landgrebe for providing the AVIRIS Indian Pines and Washington DC Mall data sets, Prof. Crawford for providing KSC and Botswana data sets, Prof. Cai for providing SDA source code, Prof. Sugiyama for providing SELF source code, Prof. Kuo for providing NWFE source code, and the anonymous reviewers for their careful reading and valuable remarks, which have contributed in improving the quality of the paper.

References

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

[2] M. Fong, "Dimension reduction on hyperspectral images," Univ. California, Los Angeles, CA, Aug. 31, 2007, Rep..

[3] K.-S. Park, S. Hong, P. Park, and W.-D. Cho, "Spectral content characterization for efficient image detection algorithm design," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, p. 72, Jan. 2007.

[4] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Kernel principal component analysis for the classification of hyperspectral remote-sensing data over urban areas," *EURASIP J. Adv. Signal Process.*, vol. 2009, Jan. 2009, Art. 11.

[5] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 9, pp. 3180–3191, Sep. 2009.

[6] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.

[7] B. C. Kuo and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 5, pp. 1096–1105, May 2004.

[8] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 7, pp. 417–441, 1933.

[9] J. Schott, *Remote Sensing: The Image Chain Approach*. New York: Oxford Univ. Press, 1996.

[10] A. Plaza, P. Martinez, J. Plaza, and R. Perez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 466–479, Mar. 2005.

[11] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[12] G. Chen and S.-E. Qian, "Dimensionality reduction of hyperspectral imagery using improved locally linear embedding," *J. Appl. Remote Sens.*, vol. 1, pp. 1–10, Mar. 2007.

[13] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina, "Exploiting manifold geometry in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 441–454, Mar. 2005.

[14] M. Belkin and P. Niyogi, "Laplacia Eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2002, vol. 14, pp. 585–591.

[15] Z. Y. Zhang and H. Y. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, 2004.

[16] X. F. He, D. Cai, S. C. Yan, and H. J. Zhang, "Neighborhood preserving embedding," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, 2005, vol. 2, pp. 1208–1213.

[17] X. F. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, vol. 16, pp. 153–160.

[18] T. Zhang, J. Yang, D. Zhao, and X. Ge, "Linear local tangent space alignment and application to face recognition," *Neurocomput. Lett.*, vol. 70, pp. 1547–1553, Mar. 2007.

[19] H. Y. Huang and B. C. Kuo, "Double nearest proportion feature extraction for hyperspectral-image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4034–4046, Nov. 2010.

[20] C.-I. Chang and H. Ren, "An experiment-based quantitative and comparative analysis of target detection and image classification algorithms for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 2, pp. 1044–1063, Mar. 2000.

[21] Q. Du, "Modified Fisher's linear discriminant analysis for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett*, vol. 4, no. 4, pp. 503–507, Oct. 2007.

[22] B. C. Kuo, C. W. Chang, C. C. Hung, and H. P. Wang, "A modified nonparametric weight feature extraction using spatial and spectral information," in *Proc. Int. Geosci. Remote Sens. Symp.*, Jul. 31–Aug. 4 2006, pp. 172–175.

[23] B. C. Kuo, C. H. Li, and J. M. Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1139–1155, Apr. 2009.

[24] B. C. Kuo, C. H. Chang, T. W. Sheu, and C. C. Hung, "Feature extractions using labeled and unlabeled data," in *Proc. IGARSS*, Jul. 2005, pp. 1257–1260.

[25] X. Zhu, "Semi-supervised learning literature survey," Department. Comput. Sci., Univ. Wisconsin Madison, Madison, WI, Comput. Sci. TR 1530, Jul. 19, 2008.

[26] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[27] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[28] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006.

[29] G. Camps-Valls, T. Bandos, and D. Zhou, "Semisupervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.

[30] D. Cai, X. F. He, and J. W. Han, "Semi-supervised discriminant analysis," in *Proc. 11th IEEE ICCV*, Rio de Janeiro, Brazil, 2008, pp. 1–7.

[31] S. G. Chen and D. Q. Zhang, "Semisupervised dimensionality reduction with pairwise constraints for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett*, vol. 8, no. 2, pp. 369–373, Mar. 2011.

[32] M. Sugiyama, T. Ide, S. Nakajima, and J. Sese, "Semi-supervised local Fisher discriminant analysis for dimensionality reduction," *Mach. Learn.*, vol. 78, pp. 35–61, Jan. 2010.

[33] W. Z. Liao, A. Pizurica, W. Philips, and Y. G. Pi, "Feature extraction for hyperspectral image based on semi-supervised local discriminant analysis," in *Proc. IEEE JURSE*, Apr. 2011, pp. 401–404.

[34] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," Tilburg Univ., West Tilburg, The Netherlands, Tech. Rep. TiCC-TR 2009-005, 2009.

[35] M. Sugiyama and S. Roweis, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[36] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.

[37] J. Ham, Y. Chen, M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[38] R. P. W. Duin, P. Juszczak, D. de Ridder, P. Paclik, E. Pekalska, and D. M. J. Tax, PRTools, A Matlab Toolbox for Pattern Recognition, 2004. [Online]. Available: http://www.prtools.org

[39] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Disc.*, vol. 2, no. 2, pp. 121–167, Jun. 1998.

[40] C. C. Chang and C. J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[41] J. Munoz-Marf, L. Bruzzone, and G. Camps-Valls, "A support vector domain description approach to supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 8, pp. 2683–2692, Aug. 2008.

[42] B. Mojaradi, H. Abrishami-Moghaddam, M. J. V. Zoej, and R. P. W. Duin, "Dimensionality reduction of hyperspectral data via spectral feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2091–2105, Mar. 2009.

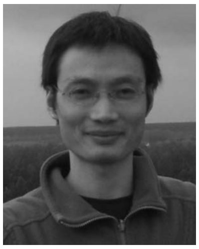
[43] M. De Berg, M. van Krefeld, M. Overmars, and O. Schwarzkopf, *Computational Geometry: Algorithms and Applications*, 3rd ed. New York: Springer-Verlag, 2008, pp. 99–101.

[44] L. Capobianco, A. Garzelli, and G. Camps-Valls, "Target detection with semisupervised kernel orthogonal subspace projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3822–3833, Jul. 2009.

[45] V. Castelli and T. Cover, "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2102–2117, Nov. 1996.

[46] K. Sinha and M. Belkin, "The value of labeled and unlabeled examples when the model is imperfect," in *Proc. NIPS*, Vancouver, BC, Canada, 2008, vol. 20.

[47] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, 2004.



Wenzhi Liao (S'10) received the B.S. degree in mathematics from Hainan Normal University, HaiKou, China, in 2006, the M.S. degree in mathematics from South China University of Technology (SCUT), GuangZhou, China, in 2008. Currently, he is working toward the Ph.D. degree both in the School of Automation Science and Enginerring, SCUT, and the Department of Telecommunications and Information Processing, Ghent University, Ghent, Belgium.

His current research interests include pattern recognition, remote sensing, and image processing.

Aleksandra Pižurica (M'03) received the Diploma degree in electical engineering from the University of Novi Sad, Novi Sad, Serbia, in 1994, the M.Sc. degree in telecommunications from The University of Belgrade, Belgrade, Serbia, in 1997, and the Ph.D. degree in engineering from Ghent University, Gent, Belgium, in 2002.

From 1994 until 1997, she was working as a Research and Teaching Assistant at the Department of Telecommunications of the University of Novi Sad, and in 1997, she joined the Department of Telecommunications and Information Systems at Ghent University. Since 2005, she has been a Postdoctoral Research Fellow of The Research Foundation—Flanders (FWO), and since 2009, a part-time Lecturer at Ghent University. She is author/coauthor of more than 30 publications in international journals and book chapters and nearly 150 publications in the proceedings of international conferences. Her research interests include statistical image modeling, multiresolution and sparse signal representations, Bayesian estimation and applications in video analysis, remote sensing, and medical imaging.

Paul Scheunders (M'98) received the B.S. degree and the Ph.D. degree in physics, with work in the field of statistical mechanics, from the University of Antwerp, Antwerp, Belgium, in 1983 and 1990, respectively.

In 1991, he became a Research Associate with the Vision Laboratory, Department of Physics, University of Antwerp, where he is currently a Professor. He has published over 120 papers in international journals and proceedings in the field of image processing and pattern recognition. His research interest includes wavelets and multispectral image processing.

Wilfried Philips (S'90–M'93–SM'10) was born in Aalst, Belgium, on October 19, 1966. He received the Diploma degree in electrical engineering and the Ph.D. degree in applied sciences from Ghent University, Ghent, Belgium, in 1989 and 1993, respectively.

From October 1989 until October 1997, he worked at the Department of Electronics and Information Systems of Ghent University for the Flemish Fund for Scientific Research (FWO-Vlaanderen), first as a Research Assistant and later as a Postdoctoral Research Fellow. Since November 1997, he is with the Department of Telecommunications and Information Processing of Ghent University, where he is currently a full-time Professor and is heading the research group "Image Processing and Interpretation," which has recently become part of the virtual Flemish ICT research institute IBBT. Some of the recent research activities in the group include image and video restoration and analysis and the modeling of image reproduction systems. Important application areas targeted by the group include remote sensing, surveillance, and industrial inspection.

Youguo Pi received the Diploma degree in automation engineering from Chongqing University, Chongqing, China, in 1982, and the Ph.D degree in mechanical engineering from South China University of Technology, Guangzhou, China, in 1998.

From July 1998 to June 2002, he was with the Information Technology Department, Automation Engineering Center, Academy Guangdong Province. Since July 2002, He has been with the College of Automation Science and Engineering, South China University of Technology, where he is currently a full-time Professor. Some of the recent research activities in the group include image processing and patter recognition and motion control. Important application areas targeted by the group include intelligent Chinese character formation and servo system.