

Research Article

Ensemble Learning Based Multiple Kernel Principal Component Analysis for Dimensionality Reduction and Classification of Hyperspectral Imagery

Hamidullah Binol 

Florida International University, Department of Electrical and Computer Engineering, Miami, FL 33174, USA

Correspondence should be addressed to Hamidullah Binol; hbinol@fiu.edu

Received 9 April 2018; Accepted 9 August 2018; Published 6 September 2018

Academic Editor: Mustansar A. Ghazanfar

Copyright © 2018 Hamidullah Binol. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Classification is one of the most challenging tasks of remotely sensed data processing, particularly for hyperspectral imaging (HSI). Dimension reduction is widely applied as a preprocessing step for classification; however the reduction of dimension using conventional methods may not always guarantee high classification rate. Principal component analysis (PCA) and its nonlinear version kernel PCA (KPCA) are known as traditional dimension reduction algorithms. In a previous work, a variant of KPCA, denoted as Adaptive KPCA (A-KPCA), is suggested to get robust unsupervised feature representation for HSI. The specified technique employs several KPCAs simultaneously to obtain better feature points from each applied KPCA which includes different candidate kernels. Nevertheless, A-KPCA neglects the influence of subkernels employing an unweighted combination. Furthermore, if there is at least one weak kernel in the set of kernels, the classification performance may be reduced significantly. To address these problems, in this paper we propose an Ensemble Learning (EL) based multiple kernel PCA (M-KPCA) strategy. M-KPCA constructs a weighted combination of kernels with high discriminative ability from a predetermined set of base kernels and then extracts features in an unsupervised fashion. The experiments on two different AVIRIS hyperspectral data sets show that the proposed algorithm can achieve a satisfactory feature extraction performance on real data.

1. Introduction

Hyperspectral imaging (HSI) provides simultaneously spatial and high resolution spectral data and helps to classify/recognize the materials that are challenging to discriminate with conventional imaging techniques [1]. However, it suffers from the curse of dimensionality. For instance, the curse of dimensionality causes increase in cost of storage, transmission, and processing of hyperspectral images. To overcome such challenges, dimensionality reduction techniques have been applied to hyperspectral data in the existing literature [2]. In general, HSI has spectral redundancy in many spectral channels. For this reason, dimension reduction or compression is possible and even necessary, especially for these bands.

Even though there are several dimension reduction approaches in the literature, including manifold learning [3, 4] and tensors [5], principal component analysis (PCA) [6] is

the one among the popular techniques [7–9]. PCA is the discrete form of the continuous Karhunen-Loève Transform and it projects the data into a subspace so that the variance retained is maximized and the least square reconstruction error is minimized [10]. Use of PCA for dimensionality reduction in HSI is a computationally suitable approach and it helps preserve the most of the variance of the raw data. Although PCA has some theoretical inadequacies [11, 12] for use on remote sensing data, particularly hyperspectral images [13], the practical applications show that the results obtained using PCA are still competitive for the purpose of classification [14, 15]. The ability of PCA is limited for high-dimensional data since it relies on only second-order statistical information. The nonlinear version of the PCA, denoted as kernel PCA (KPCA), has been proposed to overcome these limitations [16].

Since the KPCA involves the higher-order statistics, it provides more information from the original data [17] and

so it is employed in many applications including remote sensing data due to its satisfactory performance. In [18], classification performance of an artificial neural network has been demonstrated to outperform the classical approach using kernel principal components. Fauvel et al. [19] showed that the KPCA is better than the classical PCA in terms of classification accuracies. A general overview of feature reduction techniques for classification of hyperspectral images is presented in [9]. They performed comparative experiments between the unsupervised, e.g., PCA and KPCA, and supervised techniques, e.g., double nearest proportion (DNP) [20] and kernel nonparametric weighted feature extraction (KNWFE) [21]. Since the supervised learning techniques generally focus on improving class separability, these methods are expected to produce better results in terms of classification performance. The comparative results with KNWFE indicate that PCA and KPCA are still preferable to reduce dimensionality of hyperspectral images.

Fundamentally, KPCA is a version of PCA whose performance is greatly affected by the choice of the kernel and parameters. Namely, the selection of the optimal kernel and parameters is crucial for KPCA to achieve good performance. However, the application results show that no single kernel function can be best for all kinds of machine learning problems [22] and, therefore, learning of optimum kernels over a kernel set is an active research area nowadays [23–27]. Li and Yang presented an ensemble KPCA method with Bayesian inference strategy in [28]. They exploited only Gaussian radial basis function (RBF) with different scale parameters as subkernels. Zhang et al. [29] have developed a method for unsupervised kernel learning in the KPCA, dubbed as A-KPCA, and applied the new method for object recognition problems.

The A-KPCA learns the kernels via an unsupervised learning approach. The 1D input vectors, e.g., feature vectors, are transformed into 2D feature matrices by different kernels. Each column of the feature matrix comes from corresponding 1D input vector. Nonlinear feature extraction (FE) is obtained from one set of projective vectors corresponding to the column direction of the feature matrices. The set of projective vectors corresponding the row direction of the 2D feature matrices is utilized for searching optimal kernels combination simultaneously. Despite having superior performance compared to KPCA, the A-KPCA has some critical limitations. Specifically, A-KPCA works completely unsupervised, and it is thus incapable of enhancing the class separability and it has no kernel preselection process. These are the main motivations of our work.

In this paper, a novel framework is introduced for hyperspectral FE and classification based on multiple KPCA models with an Ensemble Learning (EL) strategy in a semisupervised manner. EL is a process of combining multiple models, called experts, to set up a strong model for a specific machine learning problem [30]. Strong discriminative ability of individual experts and high diversity among them are required to produce satisfactory models [31, 32]. An acceptable classification performance highly depends on the class separability of features that is directly related to the discriminative ability. Inspired by EL, we extend the A-KPCA method by employing

multiple kernels such that subkernels possessing higher discrimination ability are highlighted. The proposed approach, multiple kernel PCA (M-KPCA), learns an ensemble of multiple kernel principal components on an available labeled data set, and the final features are extracted via a weighted combination of all subkernels according to their separability performance. The early purpose of this paper is the utilization of the KPCA and A-KPCA in hyperspectral images and to determine impact of using nonlinear versions of PCA on classification performance. The further contributions and novelties in this paper can be summarized as follows: (1) a novel multikernel PCA strategy is presented by exploiting Ensemble Learning to evaluate and select the kernels; (2) M-KPCA acquires the superior classification results than PCA, KPCA, and A-KPCA by highlighting the subkernels with a class separability based weighting strategy; (3) M-KPCA produces better or competitive classification performance with other popular unsupervised FE methods like locality preserving projections (LPP) [33], random projections (RP) [34], and t-distributed stochastic neighbor embedding (t-SNE) [35]. After FE with all mentioned methods, the popular and robust support vector machines (SVMs) classifier is used for supervised classification. Since SVMs consider samples close to the class boundary, called *support vectors*, they show great performance even in high-dimensional data with small training samples [36, 37].

The paper is outlined as follows. Section 2 reviews the related work. In Section 3, the proposed framework of M-KPCA is presented. Next, a series of experiments are carried out on real data sets for verifying our method's effect in Section 4. Finally, Section 5 concludes this paper.

2. Related Work

2.1. KPCA Background. The raw data is projected into the feature space by a nonlinear mapping function and the useful information is concentrated into some principal components corresponding to the larger eigenvalues [19]. Define a learning set as \mathbf{x}_k ($k = 1, 2, \dots, N$), $\mathbf{x}_k \in \mathbb{R}^d$. Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^X$ be a nonlinear mapping from the input space to a high-dimensional feature space \mathbb{H} . The inner product in feature space is calculated by the kernel function in the original input space:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^\top \varphi(\mathbf{x}_j) \quad (1)$$

where the superscript \top represents the transpose operation. Denote $\Phi = (\varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2), \dots, \varphi(\mathbf{x}_N))$ and $\bar{\varphi} = (1/N) \sum_{k=1}^N \varphi(\mathbf{x}_k)$. Assuming $\bar{\varphi} = 0$, i.e., data are centered in \mathbb{H} , then the total scatter matrix can be defined as $\mathbf{S}_t = \Phi \Phi^\top$. To compute the projective vector \mathbf{v} for optimal solution, the KPCA employs the following norm:

$$J(\mathbf{v}) = \sum_{k=1}^N \|\mathbf{v}^\top \varphi(\mathbf{x}_k)\|^2 = \mathbf{v}^\top \mathbf{S}_t \mathbf{v}. \quad (2)$$

Computation of optimal projective vector \mathbf{v} provides solution for the eigenvalue problem: $\lambda \mathbf{v} = \mathbf{S}_t \mathbf{v}$ in which $\lambda \geq 0$

and eigenvectors $\mathbf{v} \in \mathbb{H}$. Hence, (2) can be rewritten as an equivalent problem:

$$\lambda \mathbf{z} = \mathbf{K} \mathbf{z}, \quad (3)$$

where $\mathbf{K} = \Phi^\top \Phi$ is the kernel matrix. Solutions of (3) are $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_c\}$ corresponding to the largest c eigenvalues; then $\mathbf{v}_k = \Phi \mathbf{z}_k$ is the solution vector of (2). The KPCA based FE does not include the nonlinear mapping φ as any kernel method, and it only needs a kernel function in the input space. To obtain better performance with KPCA, the parameters of the kernel are optimized. However, this optimization cannot produce adequate solutions for every application or data sets because of the nature of the kernel itself [22]. To overcome this drawback, an adaptive kernel combination technique is introduced in [29].

2.2. Adaptive KPCA (A-KPCA). As pointed out in Section 1, the performance of KPCA is notably affected by the selection of kernels and its parameters. Therefore, it needs some extensions. Let $\psi_i(\mathbf{x}) \in \mathbb{F}_i$, $i \in \{1, 2, \dots, h\}$ be a set of nonlinear mappings. As mentioned in Section 2.1, the inner products in \mathbb{F}_i are described as the kernels. Using definition of ψ_i , $\hat{\psi}_i : \mathbf{x} \rightarrow \hat{\psi}_i(\mathbf{x}) = (\underbrace{0^\top \dots 0^\top}_{1 \dots i-1}, \psi_i(\mathbf{x})^\top, \underbrace{0^\top \dots 0^\top}_{i+1 \dots h})^\top \in \mathbb{F}$ can be written. In this equation, \mathbb{F} is the Hilbert space as the direct sum of \mathbb{F}_i and the inner product in \mathbb{F} can be defined as

$$\hat{\psi}_i(\mathbf{x})^\top \hat{\psi}_j(\mathbf{y}) = 0 \quad (i \neq j), \quad (4)$$

$$\hat{\psi}_i(\mathbf{x})^\top \hat{\psi}_i(\mathbf{y}) = \psi_i(\mathbf{x})^\top \psi_i(\mathbf{y}) = K^i(\mathbf{x}, \mathbf{y}). \quad (5)$$

To construct a 2D feature matrix, a sample of learning set \mathbf{x}_k is transformed to high-dimensional feature space and then $\boldsymbol{\psi}_k = (\hat{\psi}_1(\mathbf{x}_k), \hat{\psi}_2(\mathbf{x}_k), \dots, \hat{\psi}_h(\mathbf{x}_k))$ is obtained. Here, each column of $\boldsymbol{\psi}_k$ corresponds to a nonlinear mapping generated by $\hat{\psi}_i$'s. Thus, vector-based data is converted to matrix based format. Assuming $\boldsymbol{\psi}_k$'s have zero means, i.e., $\bar{\boldsymbol{\psi}} = 0$, $\bar{\boldsymbol{\psi}} = 1/N \sum_{i=1}^N \boldsymbol{\psi}_i$ can be written. Equation (6) includes generated $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_N$ feature vectors. Appropriate \mathbf{M} and \mathbf{N} matrices must be determined to optimize

$$J(\mathbf{M}, \mathbf{N}) = \sum_{k=1}^N \|\mathbf{M}^\top \boldsymbol{\psi}_k \mathbf{N}\|_F^2, \quad (6)$$

where $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_t)$ are projective vectors corresponding to columns of $\boldsymbol{\psi}_k$ while $\mathbf{N} = (\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_p)$ corresponding to rows of $\boldsymbol{\psi}_k$. The purpose of \mathbf{M} is to extract features, while the purpose of \mathbf{N} is kernel selection. In other words, the unsupervised kernel learning and nonlinear FE are simultaneously realized according to projective vectors which are included in \mathbf{M} and \mathbf{N} . $\|\cdot\|_F$ is the *Frobenius* norm of matrix, i.e., $\|\mathbf{A}\|_F^2 := \text{tr}(\mathbf{A}\mathbf{A}^\top) = \sum_{i,j} (\mathbf{A}_{ij})^2$, where tr denotes the trace of a matrix. It can be defined as $\mathbf{M} = \boldsymbol{\psi} \mathbf{L}$, where $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_N)$. Since the size of original $\boldsymbol{\psi}$ is very large, i.e., $(h \times N)$, $\bar{\boldsymbol{\psi}} = (\sum_i \hat{\psi}_i(\mathbf{x}_1), \sum_i \hat{\psi}_i(\mathbf{x}_2), \dots, \sum_i \hat{\psi}_i(\mathbf{x}_N))$ can be

written instead of $\boldsymbol{\psi}$. Hence, $\mathbf{M} = \bar{\boldsymbol{\psi}} \mathbf{L}$, $\mathbf{L} \in \mathbb{R}^{N \times t}$ is obtained. These calculations allow us to rewrite (6) as (7):

$$J(\mathbf{L}, \mathbf{N}) = \sum_{k=1}^N \|\mathbf{L}^\top \bar{\boldsymbol{\psi}}^\top \boldsymbol{\psi}_k \mathbf{N}\|_F^2, \quad (7)$$

$$= \sum_{k=1}^N \|\mathbf{L}^\top \mathbf{K}_k \mathbf{N}\|_F^2.$$

where the constrains of (7) are $\mathbf{L} \in \mathbb{R}^{N \times t}$, $\mathbf{L}^\top \mathbf{L} = \mathbf{I}$ and $\mathbf{N} \in \mathbb{R}^{h \times p}$, $\mathbf{N}^\top \mathbf{N} = \mathbf{I}$. Here $\mathbf{K}_k = \bar{\boldsymbol{\psi}}^\top \boldsymbol{\psi}_k$ is $N \times h$ sized kernel matrix and it is constructed as follows:

$$\mathbf{K}_k(i, j) = K^j(\mathbf{x}_i, \mathbf{x}_k). \quad (8)$$

To solve this optimization problem, inspired by Ye's work [38], an iterative procedure is presented by the following theorem [29].

Theorem 1. Let \mathbf{L} and \mathbf{N} be the optimal solution to (7): then (i) t eigenvectors corresponding to the largest t eigenvalues of the matrix $\mathbf{A}_L = \sum_{k=1}^N \mathbf{K}_k \mathbf{R} \mathbf{R}^\top \mathbf{K}_k^\top$ form \mathbf{L} for a given \mathbf{N} ; (ii) p eigenvectors corresponding to the largest p eigenvalues of the matrix $\mathbf{A}_N = \sum_{k=1}^N \mathbf{K}_k \mathbf{L} \mathbf{L}^\top \mathbf{K}_k^\top$ create \mathbf{N} for a given \mathbf{L} .

After computing \mathbf{L} and \mathbf{N} , these matrices can be used to extract the nonlinear features for a test instance $\mathbf{z} \in \mathbb{R}^d$. Kernel matrix $\mathbf{K}_{test}(i, j) = K^j(\mathbf{x}_i, \mathbf{z})$ is constructed and then projected according to $\mathbf{C} = \mathbf{L}^\top \mathbf{K}_{test} \mathbf{N}$, so the nonlinear features are contained in \mathbf{C} . The A-KPCA method is given in Algorithm 1.

3. Multiple Kernel PCA (M-KPCA)

In Section 2, we have demonstrated that A-KPCA manipulates more than one subkernels. A mapping rule transforms input data samples into corresponding Reproducing Kernel Hilbert Space. Each kernel thus acquires a particular type of information from a given data set, thereby providing a partial description of view data. The value of this specific information may vary according to different machine learning tasks such as classification, clustering, dimensionality reduction, etc. For instance, in a classification problem, high discrimination ability of kernels yields the better results. Hence, we add this capability to A-KPCA with ideas of EL. Our proposed technique learns new representation for a hyperspectral image exploiting all available training data. It is thus independent of the classifier.

As seen in formulation (8) and Theorem 1, there are not any coefficients to quantify the contribution of subkernels in classification. In other words, the A-KPCA utilizes the unweighted summation. Nevertheless, the discriminative ability of kernels in FE plays significant role for the separability of the classifier. If we add a weighting coefficient on right side of (8), then it becomes

$$\mathbf{K}_k(i, j) = \alpha_j K^j(\mathbf{x}_i, \mathbf{x}_k). \quad (9)$$

The discriminative ability of a kernel can be measured by an ideal kernel in a given classification task. Cristianini et al.

Input: Given training set \mathbf{x}_k ($k = 1, 2, \dots, N$).
 (a) Create the kernel matrix \mathbf{K}_k for each \mathbf{x}_k .
 (b) Get initial \mathbf{N}_0 and $i \leftarrow 1$.
 (c) For given \mathbf{N}_{i-1} , calculate the t eigenvectors of \mathbf{A}_L corresponding to the largest t eigenvalues.
 (d) For given \mathbf{L}_i , calculate the p eigenvectors of \mathbf{A}_N corresponding to the largest p eigenvalues.
 (e) $i \leftarrow i + 1$, goto step (c) until convergence.
Output: \mathbf{L} and \mathbf{N} .

ALGORITHM 1: A-KPCA Algorithm.

[23] introduced a measure of similarity between two arbitrary kernels or between a kernel and an ideal kernel called kernel alignment (KA). The alignment between two regular kernels is given as

$$A(\mathbf{K}_1, \mathbf{K}_2) = \frac{\langle \mathbf{K}_1, \mathbf{K}_2 \rangle_F}{\sqrt{\langle \mathbf{K}_1, \mathbf{K}_1 \rangle_F \langle \mathbf{K}_2, \mathbf{K}_2 \rangle_F}} \quad (10)$$

where the *Frobenius* product $\langle \cdot, \cdot \rangle_F$ of two Gram matrices \mathbf{A} and \mathbf{B} is defined as $\langle \mathbf{A}, \mathbf{B} \rangle_F = \text{tr}(\mathbf{A}\mathbf{B})$ [23, 39]. This measure can be viewed as the cosine of the angle between \mathbf{K}_1 and \mathbf{K}_2 , so it fluctuates between $\{-1, 1\}$ for arbitrary matrices. However, since we consider only positive semidefinite Gram matrices in KA, the score is lower bounded by zero. The alignment can also be adopted to capture the degree of agreement between a kernel and the target label matrix, also considered as *ideal kernel*. A larger value of KA indicates the higher discriminative ability and it is one of the main strengths for a subclassifier such that they improve the ensemble effect in an EL strategy [40, 41]. An idealized kernel for a binary classification problem can be composed of the dot product of target labels, i.e., $\mathbf{y}\mathbf{y}^\top$, and the alignment between a kernel and the ideal kernel is written as

$$A(\mathbf{K}, \mathbf{y}\mathbf{y}^\top) = \frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^\top \rangle_F}{\|\mathbf{K}\|_F \|\mathbf{y}\mathbf{y}^\top\|_F}. \quad (11)$$

Our goal is to construct an A-KPCA based algorithm which has improved separability of multiclass patterns. Here, kernel class separability (KCS) measure based on scatter matrix is employed to measure the class separability of training samples in feature space. The KSC is a general form of KA and it can be written in the form [42]:

$$J^\varphi = \frac{\text{tr}(\mathbf{S}_B^\varphi)}{\text{tr}(\mathbf{S}_W^\varphi)} \quad (12)$$

where \mathbf{S}_B^φ and \mathbf{S}_W^φ , respectively, stand for between-class scatter matrix and within-class scatter matrix in kernel space and the traces of them are obtained as

$$\begin{aligned} \text{tr}(\mathbf{S}_B^\varphi) &= \sum_{i=1}^c n_i [(\mathbf{m}_i^\varphi - \mathbf{m}^\varphi)^\top (\mathbf{m}_i^\varphi - \mathbf{m}^\varphi)] \\ \text{tr}(\mathbf{S}_W^\varphi) &= \sum_{i=1}^c \sum_{j=1}^{n_i} [(\varphi(\mathbf{x}_{ij}) - \mathbf{m}_i^\varphi)^\top (\varphi(\mathbf{x}_{ij}) - \mathbf{m}_i^\varphi)] \end{aligned} \quad (13)$$

where n_i ($i = 1, \dots, c$) denotes the number of training samples in the i th class, $N = \sum_{i=1}^c n_i$, and \mathbf{x}_{ij} is the j th sample in the related class. \mathbf{m}_i and \mathbf{m} are the mean vector for i th class and the mean vector for all training samples, respectively. φ is the mapping function from the input space to the feature space as described in the beginning of Section 2.1. A larger value of J^φ signifies superior class separability in the training set. A maximization problem may thus be created to obtain optimal kernels and their parameters or eliminate weak kernels [43], but, in this paper, we directly exploit the value of (12) as the measure of discriminability; hence

$$\alpha_j = J^{\varphi_j}, \quad (14)$$

where $\varphi_j(\mathbf{x})^\top \varphi_j(\mathbf{y}) = K^j(\mathbf{x}, \mathbf{y})$.

After all, we extend the noniterative A-KPCA algorithm using kernel class separability measure with a semisupervised strategy. The proposed noniterative M-KPCA technique is given in Algorithm 2.

4. Experiments

In this section, we investigate the performance of the proposed M-KPCA algorithm compared with a number of conventional and state-of-the-art techniques on two Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) hyperspectral data sets. Our experiments are conducted on a machine with an Intel Core i5-2410M CPU at 2.30GHz and 8GB DDR-III RAM.

4.1. Data Sets and Experimental Setup. The first set is an airborne remote sensing data captured by the AVIRIS sensor over northwest Indiana on June 12, 1992. Indian Pines data has 16 labeled classes and 145 lines/scene and 145 pixels/line. Originally, the scene has 220 spectral bands (10 nm spectral bandwidth from 0.4 to 2.5 μm); after discarding the water absorption and noise bands, based on [44, 45], only 159 bands were used in the experiments.

Airborne hyperspectral data is acquired by AVIRIS sensor at 18 m spatial resolution over Kennedy Space Center (KSC) during March 1996 and has been employed as a second data source. Noisy bands and water absorption bands are removed. The remaining of the HSI data has 176 bands for 13 wetland and upland classes. Figure 1 shows the RGB image of Indian Pines and false color image of the KSC. Table 1 lists the summary of the data sets in our experiments. All samples in each data are adjusted in the range $[0, 1]$, as suggested in [46].

Input: Given training set \mathbf{x}_k ($k = 1, 2, \dots, N$) with labels $\mathbf{y}_k \in \{1, 2, \dots, c\}$.

(a) Obtain α_i 's for corresponding pre-selected kernels using Eqs. (12) and (13).

(b) Create the kernel matrix \mathbf{K}_k for each \mathbf{x}_k as in Eq. (9).

(c) Calculate the eigenvectors $\{v_i\}$ and eigenvalues $\{\lambda_i\}$ of $\mathbf{A}_L = \sum_{k=1}^N \mathbf{K}_k \mathbf{K}_k^T$. Sort the eigenvectors according to the decreasing order of λ_i and select first t eigenvectors (v_1, \dots, v_t) .

(d) Calculate the eigenvectors $\{r_i\}$ and eigenvalues $\{\delta_i\}$ of $\mathbf{A}_N = \sum_{k=1}^N \mathbf{K}_k^T \mathbf{K}_k$. Sort the eigenvectors according to the decreasing order of δ_i and select first p eigenvectors (r_1, \dots, r_p) .

(e) The final subspaces are $\mathbf{L} = (v_1, \dots, v_t)$ and $\mathbf{N} = (r_1, \dots, r_p)$.

Output: \mathbf{L} and \mathbf{N} .

ALGORITHM 2: M-KPCA Algorithm.

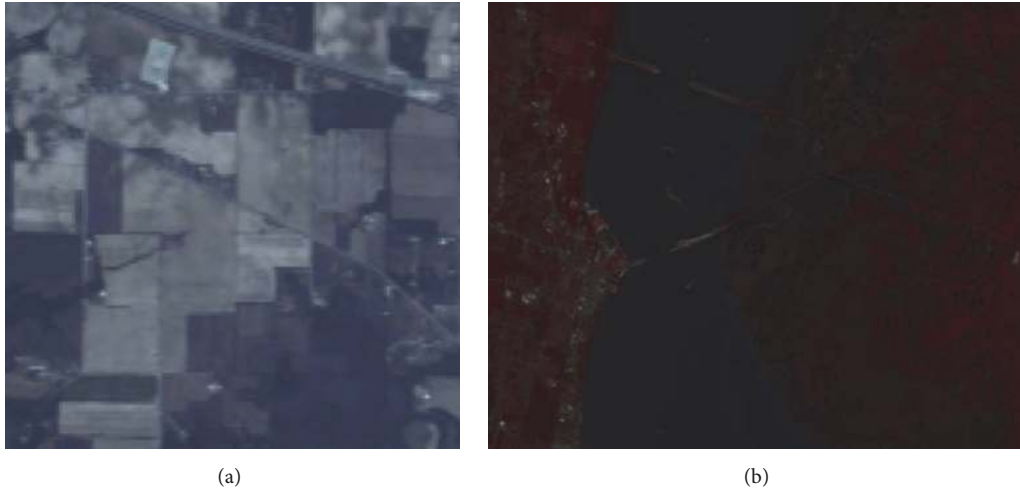


FIGURE 1: AVIRIS data sets: (a) RGB image of Indian Pines obtained from default bands {R:24, G:14, B:8} and (b) false color composite image of KSC obtained from bands {R:40, G:29, B:20}.

TABLE 1: Information for Indian Pines and KSC data sets.

No	Indian Pines	Samples		KSC	Samples	
	Class name	Train	Test	Class name	Train	Test
1	Alfalfa	14	40	Scrub	190	571
2	Corn-no till	359	1075	Willow swamp	61	182
3	Corn-min till	209	625	Cabbage hamm	64	192
4	Corn	59	175	Cabbage palm	63	189
5	Grass-pasture	124	373	Slash pine	40	121
6	Grass-trees	187	560	Oak	57	172
7	Grass-past. moved	6	20	Hardwood swamp	264	791
8	Hay-windrowed	122	367	Graminoid marsh	108	323
9	Oats	5	15	Spartina marsh	130	390
10	Soybean-no till	242	726	Cattail marsh	101	303
11	Soybean-min till	617	1851	Salt marsh	105	314
12	Soybean-clean till	154	460	Mud flats	126	377
13	Wheat	53	159	Water	232	695
14	Woods	324	970	—		
15	Bldg-Grass-Tree	95	285	—		
16	Stone-steel towers	24	71	—		
Total		2594	7772		1541	3670

TABLE 2: Employed kernel functions.

Kernels	Formula
RBF	$K_{RBF}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\ \mathbf{x} - \mathbf{y}\ ^2}{2\sigma^2}\right)$
Laplacian	$K_{Lap}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\ \mathbf{x} - \mathbf{y}\ }{\sigma}\right)$
Cauchy	$K_{Cau}(\mathbf{x}, \mathbf{y}) = \frac{1}{(1 + \ \mathbf{x} - \mathbf{y}\ ^2/\sigma^2)}$
Histogram Intersection (HIST)	$K_{HIST}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^N \min(x_i, y_i)$

The PCA and KPCAs are implemented using the SIMFEAT toolbox [47]. In each experiment, a single kernel is selected for the KPCA. In addition to Gaussian radial basis function (RBF) kernel which is formulated as $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/2\sigma^2)$, we have employed three more kernels (see Table 2). Before solving the eigenvalue problem, the parameter $\sigma \in \mathbb{R}^+$ in the RBF, Laplacian and Cauchy kernels should be selected or optimized. Unless otherwise stated the kernel parameter is set to

$$\sigma = \sqrt{\frac{\sum_N \|\mathbf{x}^k - \bar{\mathbf{x}}\|^2}{N}}, \quad (15)$$

where $\bar{\mathbf{x}}$ is the centroid of the total N training data [48]. The kernel parameter σ is nonoptimized and same for each exploited kernel given in Table 2. However, the aim of combination of nonoptimized kernels is to yield a better FE technique for classification.

The valuable parts of the obtained cumulative eigenvalues after eigen-decomposition for each method are shown in Figure 2. According to the cumulative eigenvalues of PCA, two principal components reach 99% of total variance for Indian Pines case. Nevertheless, in KSC case, three principal components are needed to reach 99% of information. According to these results, the new dimensions of Indian Pines and KSC for classification experiment are, respectively, defined as 2 and 3. However, hyperspectral information cannot be represented utilizing only the second-order statistics as it is pointed out in Section 1. From Figure 2, it can be derived that more kernel principal components (KPCs) are needed to realize the same amount of variance as for PCA. Note that the total number of components with PCA is equal to the number of bands, i.e., 159 for Indian Pines, while it is equal for KPCA to the size of the number of training samples, i.e., 2594, which is significantly higher. For the Indian Pines data set, the first 11, 51, 31, and 33 KPCs are needed to accomplish 99% of the cumulative variance with RBF, Laplacian, Cauchy, and histogram intersection (HIST) kernels, respectively. We observe that 8 KPCs are needed with the RBF, 56 with the Laplacian, 18 with the Cauchy, and 55 with the HIST kernel to achieve same amount of information considering to KSC results. In the case of A-KPCA and M-KPCA, p is set to 1 for kernel selection. For the Indian Pines data set, 35 adaptive KPCs and 20 multiple KPCs contain 99% of information and only 14 adaptive KPCs and 12 multiple KPCs for the KSC.

In order to demonstrate the first principal components (PCs) more efficiently, a subimage of size 100×100 in the KSC hyperspectral cube is selected. The first PCs for all of the methods are depicted in Figure 3.

After FE, SVM classifier has been employed for classification. For nonlinear SVMs, we have used the RBF kernel which is formulated in Table 2. The classification experiments and the optimization of parameters, C and σ , of SVMs are achieved using LIBSVM [49] with 5-fold cross validation technique. Since SVMs are designed to solve binary problems, various approaches have been proposed for multiclass situations such as remote sensing applications. The most popular approaches for multiclass classification are one-against-all (1AA) and one-against-one (1A1). In this paper, we have applied the 1AA strategy for each class. Each test sample is finally labeled as the class whose output score is maximum.

Finally, we compare the proposed M-KPCA algorithm against five state-of-the-art dimension reduction algorithms, i.e., linear discriminant analysis (LDA) [50], LPP, probabilistic PCA (pPCA) [51], RP, and t-SNE. LDA, LPP, pPCA, and t-SNE are implemented using the MATLAB toolbox [52] for dimensionality reduction, and RP algorithm is designed based on Wang's work [53].

4.2. Comparison with KPCA and A-KPCA. The original data sets, termed as raw, are also classified for comparisons. Tables 3 and 4 compare the performance of all models numerically (class accuracies and overall accuracy (OA) in percentages) and statistically (kappa test) for the Indian Pines and the KSC data sets, respectively.

Inspection of Table 3 reveals A-KPCA outperforms PCA and all the four KPCAs. Further analysis shows that the KPCA performs significantly better than the conventional PCA. Regarding the OAs, it is clear that the M-KPCA based classification produces more accurate results when compared to the A-KPCA based classification. RBF kernel gives the best results for KPCA among the other kernel functions as seen in Table 3.

The results for the KSC data set are reported in Table 4. Regarding the PCA and KPCA results, FE does not improve the accuracies significantly. The comparison between KPCA and PCA shows that KPCA performs better than the PCA in terms of classification accuracies. Moreover, classification of the A-KPCA features is more precise than the one yielded employing the all KPCs. As with the previous experiment, the best results are obtained with the M-KPCA. Figures 4 and 5 represent the available labeled scenes and classification maps of all models for the Indian Pines and KSC data sets, respectively.

In the last experiment, we increase the number of KPCAs in both A-KPCA and M-KPCA utilizing different scale parameters in the same kernel. Tables 3 and 4 show that the best single kernel for each data set is different. Therefore, we, respectively, adopt the seven RBF and Cauchy kernel functions for Indian Pines and KSC such as their scale parameters in the range $[\sigma_0/8, 8\sigma_0]$. The central parameter (σ_0) is determined by (15). The SVM is again employed for classification after FE. The selection of eigenvalues for each method

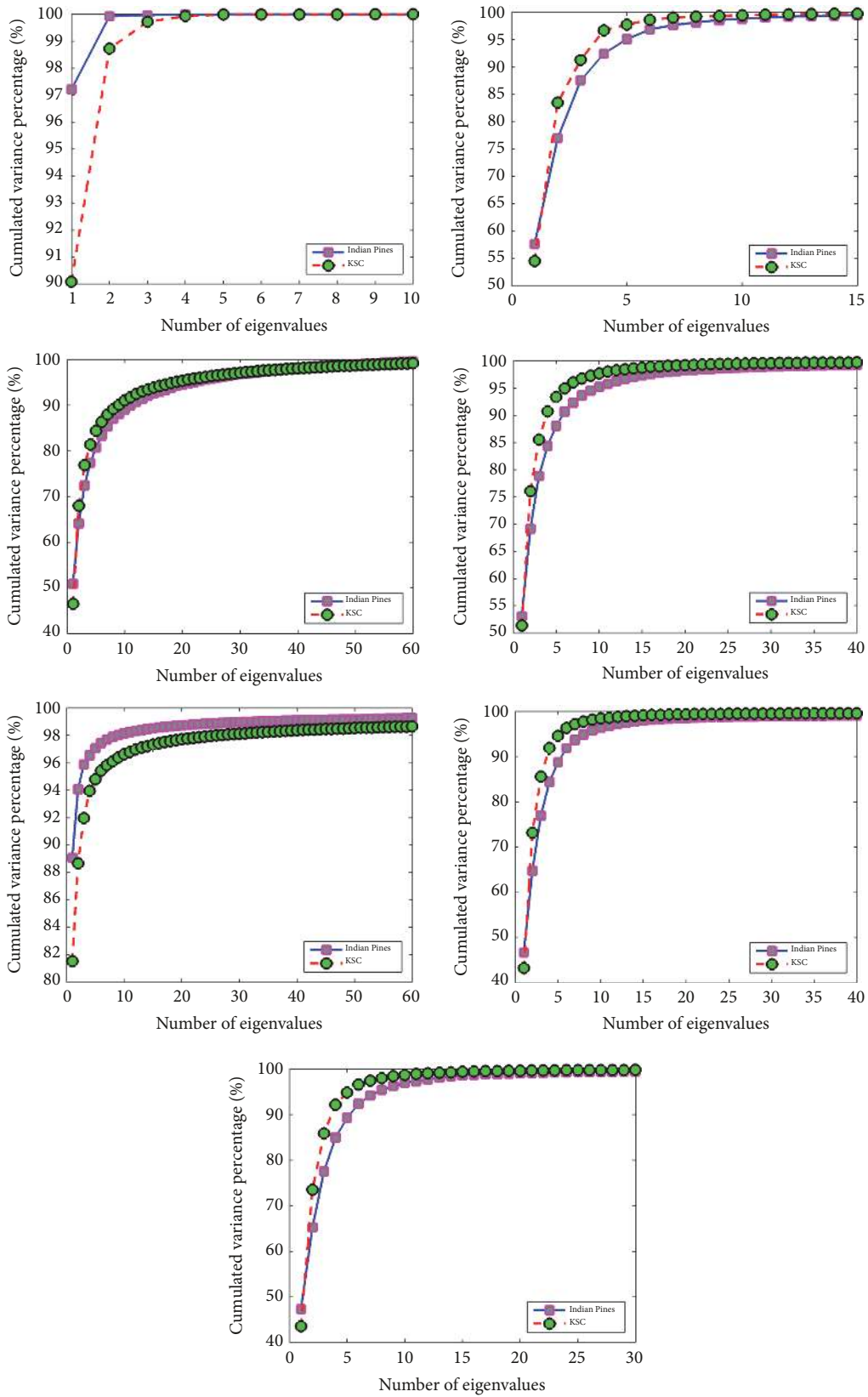


FIGURE 2: Percentages of cumulative variances of the initial data retained in the components of the new representation for the two data sets: (left to right) first row: PCA and $KPCA_{RBF}$; second row: $KPCA_{Lap}$ and $KPCA_{Cau}$; third row: $KPCA_{HIST}$ and A-KPCA; bottom: M-KPCA.

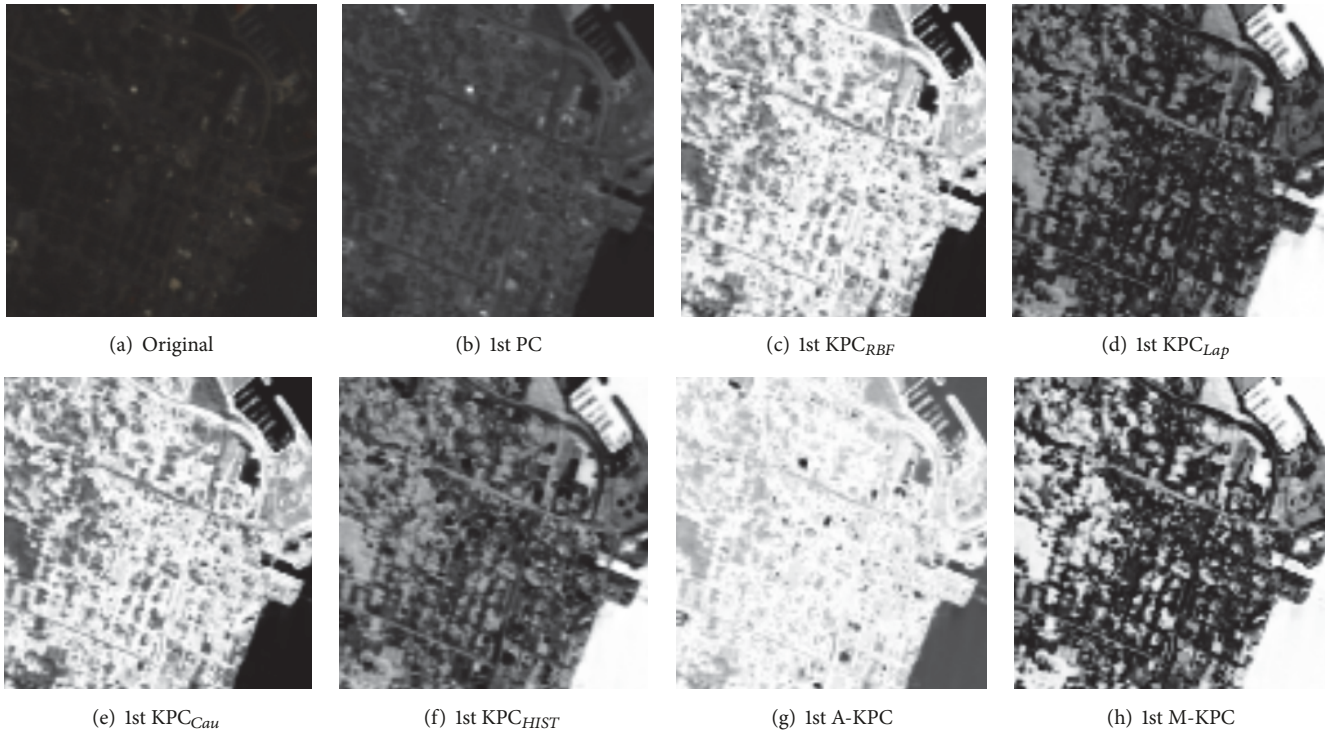


FIGURE 3: (a) False color image of KSC subscene and most informative principal components (PCs) for the subimage of KSC data: (b) 1st PC; (c) 1st KPC_{RBF} ; (d) 1st KPC_{Lap} ; (e) 1st KPC_{Cau} ; (f) 1st KPC_{HIST} ; (g) 1st A-KPC; (h) 1st M-KPC.

TABLE 3: Overall accuracy, OA (%), and Kappa statistic, κ , of the Indian Pines data set with different models. The best scores for each class are highlighted in **bold** face font, and the second best is underlined.

Feature	Raw	PCA	$KPCA_{RBF}$	$KPCA_{Lap}$	$KPCA_{Cau}$	$KPCA_{HIST}$	A-KPCA	M-KPCA
# of features	159	2	11	51	31	33	35	20
SVM C	430.5	279.17	534.67	663.98	663.98	663.98	603.46	346.67
params σ	2.18	15.22	13.37	13.28	13.27	13.28	16.59	4.74
1	87.04	33.34	88.89	74.07	77.78	<u>90.74</u>	<u>90.74</u>	94.44
2	88.77	40.59	86.82	75.80	82.08	83.19	<u>89.82</u>	89.96
3	86.09	06.95	86.81	65.47	82.61	78.30	<u>88.49</u>	88.60
4	<u>92.31</u>	35.90	91.03	68.38	81.62	82.48	91.45	96.58
5	<u>98.19</u>	58.95	97.79	88.73	95.77	90.74	97.59	98.59
6	98.93	91.43	98.39	97.59	97.99	96.52	<u>99.06</u>	99.33
7	<u>92.31</u>	30.77	<u>92.31</u>	88.46	<u>92.31</u>	84.62	<u>92.31</u>	96.15
8	<u>99.39</u>	95.50	98.98	97.75	98.98	97.34	<u>99.39</u>	99.59
9	90.00	00.00	95.00	90.00	85.00	90.00	95.00	95.00
10	89.67	66.32	91.12	80.58	90.50	88.74	<u>91.22</u>	91.84
11	91.90	73.30	<u>93.27</u>	86.95	91.13	89.55	93.11	94.25
12	94.30	21.01	95.11	80.78	90.88	88.27	<u>95.44</u>	95.60
13	<u>99.06</u>	73.58	<u>99.06</u>	98.11	98.58	97.64	<u>99.06</u>	99.53
14	97.30	86.79	98.07	96.91	<u>97.99</u>	96.52	97.84	97.60
15	80.00	15.53	77.11	53.42	70.26	58.95	<u>82.63</u>	88.95
16	92.63	90.53	94.74	94.74	91.58	90.53	90.53	94.74
OA	92.47	59.78	92.69	83.92	90.16	88.30	<u>93.44</u>	94.28
κ	0.9130	0.5399	0.9210	0.8215	0.8967	0.8749	<u>0.9317</u>	0.9348

TABLE 4: Overall accuracy, OA (%), and Kappa statistic, κ , of the KSC data set with different models. The best scores for each class are highlighted in **bold** face font, and the second best is underlined.

Feature		Raw	PCA	KPCA _{RBF}	KPCA _{Lap}	KPCA _{Cau}	KPCA _{HIST}	A-KPCA	M-KPCA
# of features		176	3	8	56	18	55	14	12
SVM	C	534.66	279.17	430.54	663.98	181.02	346.69	117.38	165.69
params	σ	3.77	9.87	34.95	5.58	10.69	5.58	5.61	5.01
1		97.50	92.64	97.50	95.66	96.58	96.32	<u>98.03</u>	98.29
2		<u>94.24</u>	84.77	93.83	90.12	91.77	91.36	<u>94.24</u>	95.47
3		<u>93.75</u>	89.84	93.36	91.80	<u>93.75</u>	<u>93.75</u>	93.36	97.26
4		78.17	49.21	80.16	71.43	78.17	75.40	<u>81.35</u>	93.25
5		77.64	56.52	60.87	55.28	61.49	57.76	<u>78.88</u>	88.82
6		75.11	50.66	58.08	55.02	75.11	58.95	<u>82.53</u>	86.90
7		93.34	73.33	89.52	91.43	90.48	91.43	<u>94.29</u>	95.24
8		<u>96.29</u>	78.65	94.20	90.26	95.13	91.18	96.75	96.28
9		99.04	97.50	98.65	96.73	98.65	97.50	<u>99.23</u>	99.61
10		100	91.58	99.75	97.77	99.26	97.77	100	99.26
11		<u>99.05</u>	97.85	<u>99.05</u>	98.09	<u>99.05</u>	98.81	<u>99.05</u>	99.28
12		<u>99.06</u>	81.11	96.82	95.23	97.42	95.63	99.01	99.20
13		100	99.89	100	99.89	99.89	99.89	99.89	100
OA		95.51	86.53	93.78	91.65	94.34	92.59	<u>96.14</u>	97.52
κ		0.9522	0.8493	0.9347	0.9106	0.9407	0.9213	<u>0.9601</u>	0.9724

TABLE 5: Overall accuracy, OA (%), and Kappa statistic, κ , of SVMs with different features. The best scores for each data set are highlighted in **bold** face font, and the second best is underlined.

Data	Kernel	Metric	KPCA							A-KPCA	M-KPCA
			$\sigma_0/8$	$\sigma_0/4$	$\sigma_0/2$	σ_0	$2\sigma_0$	$4\sigma_0$	$8\sigma_0$		
Indian Pines	RBF	OA	83.54	88.69	90.52	92.69	57.80	52.41	19.47	<u>94.09</u>	96.03
		κ	0.8006	0.8703	0.8918	0.9202	0.5036	0.4405	0.1267	<u>0.9327</u>	0.9459
KSC	Cauchy	OA	91.83	93.78	93.24	94.34	80.27	72.54	69.76	<u>95.93</u>	96.36
		κ	0.9089	0.9307	0.9247	0.9409	0.7794	0.6918	0.6605	<u>0.9547</u>	0.9574

is defined in 99% confidence interval. Table 5 summarizes the classification accuracies of this experiment. The results show that the M-KPCA-based features are better than the individual KPCAs and A-KPCA features on all data sets, no matter which kernel parameter is applied.

4.3. M-KPCA versus Other Dimension Reduction Algorithms.

In this section, we compare our method (M-KPCA) with the five FE methods, i.e., LDA, LPP, pPCA, RP, and t-SNE. M-KPCA is constructed with the subkernels indicated in Table 2, and kernel parameters are determined from (15). Different values of the dimensionality number of the new subspaces are tested for the SVM classifier across the two data sets. A set of values {1, 2, 5, 10, 20, 50} are independently generated for the subspace dimension. The classification accuracy is reported for each model, and we plot the results in Figure 6.

Inspection of Figure 6 reveals that proposed method regularly outperforms the competing FE methods for multiclass classification in higher dimensions. For instance, if the number of extracted features is set to 50, M-KPCA improves over the best competing method RP by 5.19% in terms of OA on Indian Pines and by 3.78% on KSC. It can be also found from Figure 6 that t-SNE method is highly stable against any

dimensional changes. On comparison of methods, we also observe that the performance of LDA and pPCA is limited for both data sets. Considering the lower dimensions (i.e., when the number of new dimension is assigned a value smaller than 10), the best features are produced by the t-SNE which is also the most time-consuming method. The rest of the methods are sorted as RP, LDA, LPP, pPCA, and M-KPCA in ascending order according to the average computation times.

5. Conclusion

In this paper, a novel semisupervised KPCA framework named multiple KPCA (M-KPCA) is proposed for effective feature extraction of hyperspectral images. It applies ensemble strategy to favor good candidate kernels during nonlinear projections. A noniterative algorithm is developed to simultaneously feature extraction and kernel combination based on a kernel class separability criteria. In terms of the number of kernels, KPCA uses only one base kernel with predefined parameter(s) (if existing). In terms of the kernel quality, A-KPCA has no measurement procedure to evaluate the efficiency of kernels. M-KPCA overcomes these drawbacks of both KPCA and A-KPCA.

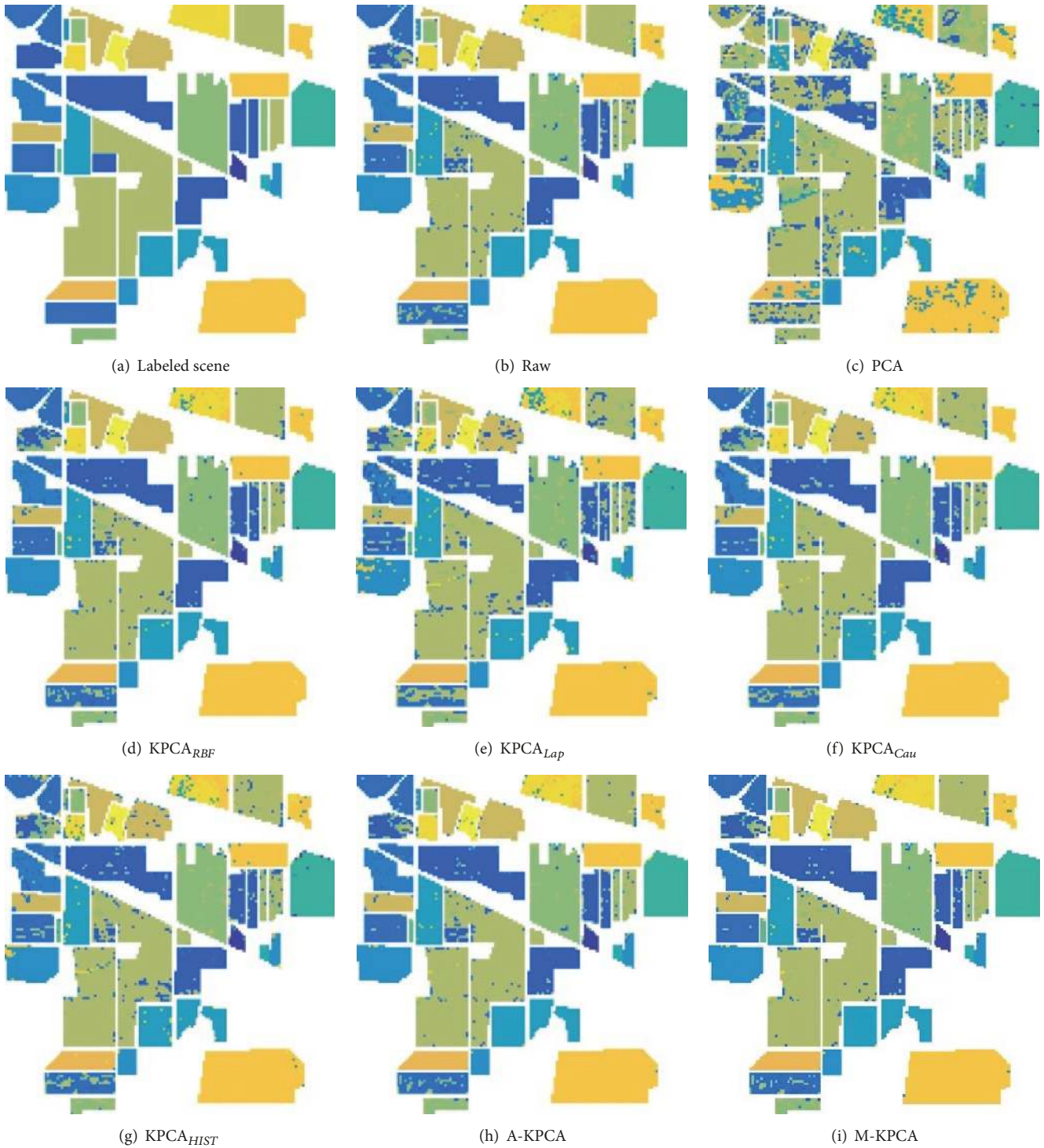


FIGURE 4: Classification results obtained with the Indian Pines: (a) labeled scene and classification maps using the (b) raw data; (c) PCA; (d) $KPCA_{RBF}$; (e) $KPCA_{Lap}$; (f) $KPCA_{Cau}$; (g) $KPCA_{HIST}$; (h) A-KPCA; (i) M-KPCA. The classification was done by SVM with an RBF kernel.

Dimension reduced HSI data is classified by nonlinear SVMs to compare classification performance for several models. Experiments on two real HSI data sets demonstrate that the best kernel type varies according to data (see

Tables 3 and 4). In the first test, KPCA presents better performance compared to the conventional PCA. Overall evaluation for dimension reduction performance of PCA, KPCAs, A-KPCA, and M-KPCA techniques shows that the

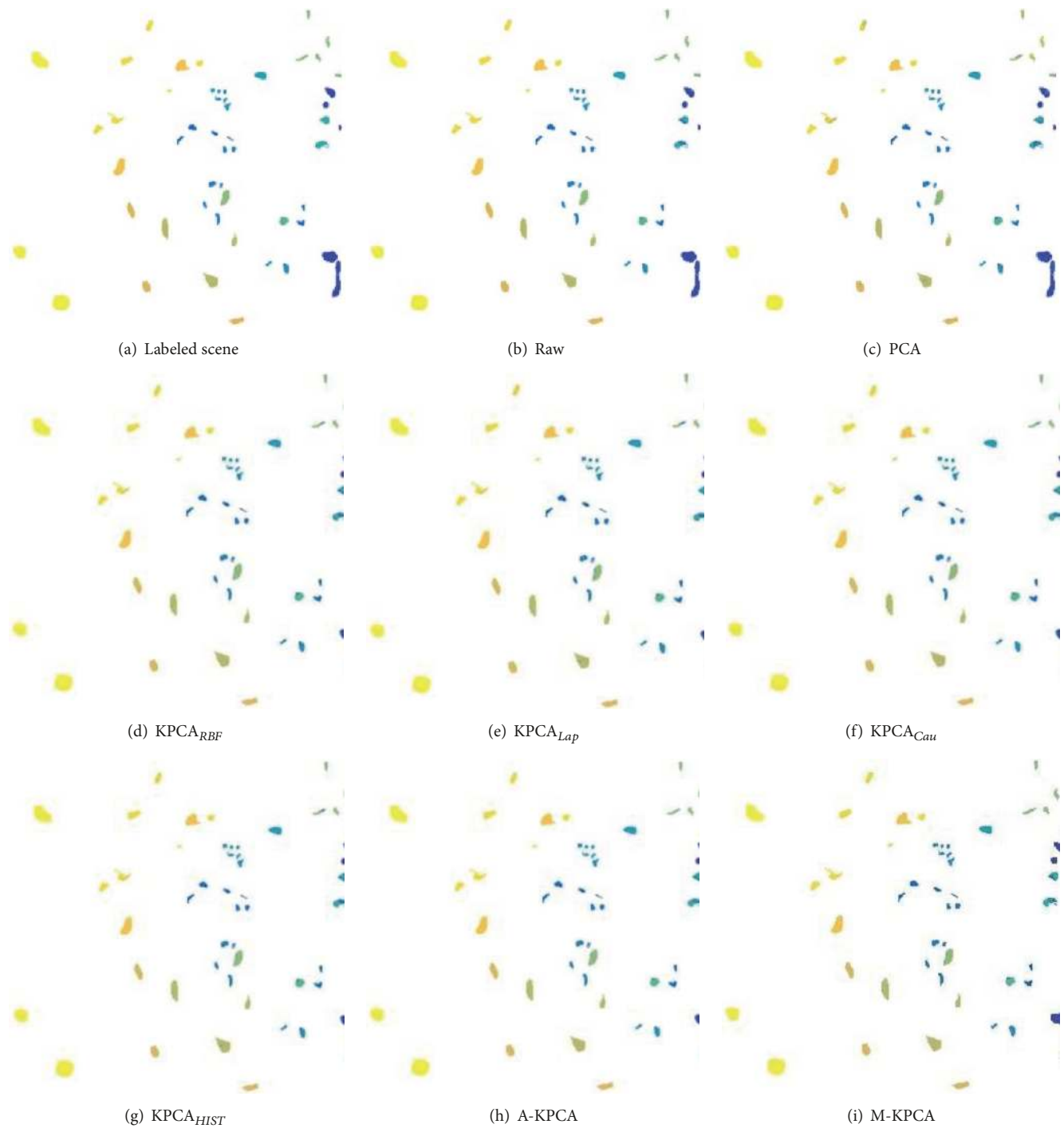


FIGURE 5: Classification results obtained with the KSC: (a) labeled scene and classification maps using the (b) raw data; (c) PCA; (d) $KPCA_{RBF}$; (e) $KPCA_{Lap}$; (f) $KPCA_{Cau}$; (g) $KPCA_{HIST}$; and (h) A-KPCA; (i) M-KPCA. The classification was done by SVM with an RBF kernel.

M-KPCA is more successful than the others. In the second experiment, we have employed seven candidate kernel functions using different kernel parameters for each data. These KPCAs are then utilized to construct the A-KPCA and M-KPCA. Experiments on the AVIRIS data sets confirm that the M-KPCA outperforms the individual KPCAs and the A-KPCA in terms of both OA and Kappa coefficient. Moreover,

the comparative results in Section 4.3 demonstrate that M-KPCA experimentally accomplished superior or competitive classification accuracy more than the other unsupervised state-of-the-art FE methods.

The results clearly validate that semisupervised learning of kernels with M-KPCA increases the robustness of nonoptimized KPCAs. One and probably most important limitation

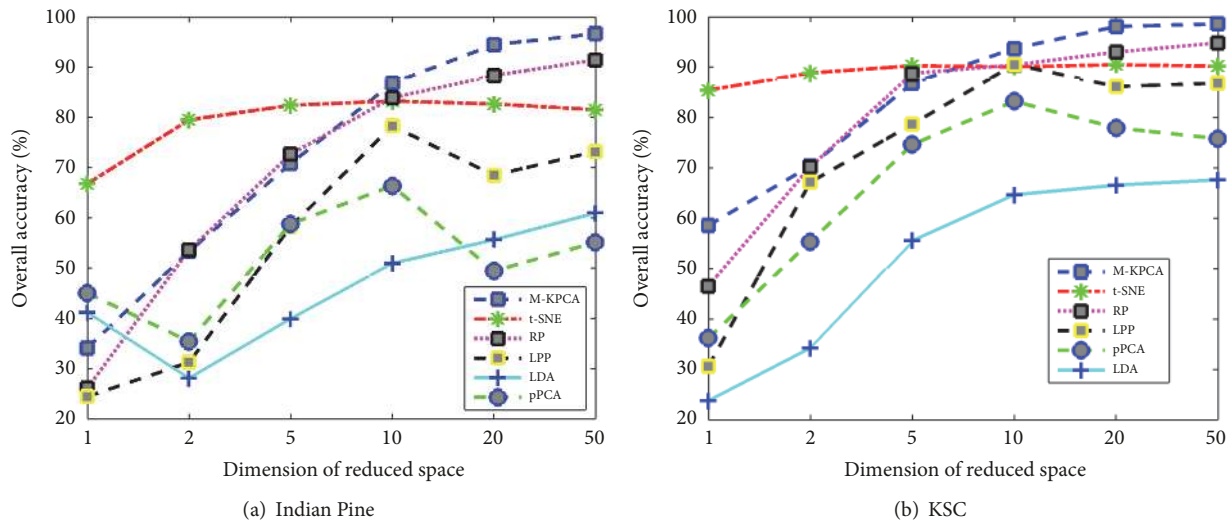


FIGURE 6: Classification results of SVM on the two data sets for varying subspace dimension produced by six FE methods.

of the M-KPCA is its computational complexity, related to the number of samples used for constructing the kernel matrix. Therefore, our future work aims to address the problem of reducing the complexity. It is also possible to extend the proposed method to a selective approach which eliminates weak kernels before feature extraction.

Data Availability

The Indian Pines and KSC data that support the findings of this study are, respectively, available in <https://engineering.purdue.edu/~biehl/> and http://www.ehu.es/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes.

Conflicts of Interest

The author declares that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Daoqiang Zhang and Zhi-Hua Zhou for providing a part of the source code for the A-KPCA.

References

- [1] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification via kernel sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 217–231, 2013.
- [2] D. W. Scott, *The Curse of Dimensionality and Dimension Reduction in Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, 1992.
- [3] H. Huang, F. Luo, J. Liu, and Y. Yang, "Dimensionality reduction of hyperspectral images based on sparse discriminant manifold embedding," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 106, pp. 42–54, 2015.
- [4] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding," *Pattern Recognition*, vol. 48, no. 10, pp. 3102–3112, 2014.
- [5] L. Zhang, L. Zhang, D. Tao, and X. Huang, "Tensor discriminative locality alignment for hyperspectral image spectral-spatial feature extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 242–256, 2013.
- [6] P. Nagabhushan, D. S. Guru, and B. H. Shekar, "Visual learning and recognition of 3D objects using two-dimensional principal component analysis: A robust and an efficient approach," *Pattern Recognition*, vol. 39, no. 4, pp. 721–725, 2006.
- [7] L. O. Jimenez, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 6, pp. 2653–2667, 1999.
- [8] S. Kaewpijit, J. L. Moigne, and T. El-Ghazawi, "Automatic reduction of hyperspectral imagery using wavelet spectral analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 4, pp. 863–871, 2003.
- [9] X. Jia, B.-C. Kuo, and M. M. Crawford, "Feature mining for hyperspectral image classification," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 676–697, 2013.
- [10] C. Lee and D. A. Landgrebe, "Analyzing high-dimensional multispectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 31, no. 4, pp. 792–800, 1993.
- [11] J. A. Richards, *Remote Sensing Digital Image Analysis*, Springer-Verlag, Berlin, Germany, 1993.
- [12] D. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*, John Wiley & Sons, Hoboken, NJ, USA, 1999.
- [13] H. Binol, S. Ochilov, M. S. Alam, and A. Bal, "Target oriented dimensionality reduction of hyperspectral data by Kernel Fukunaga–Koontz Transform," *Optics and Lasers in Engineering*, vol. 89, pp. 123–130, 2015.
- [14] M. Lennon, G. Mercier, M. Mouchot, L. Hubert-Moy, and S. B. Serpico, "Curvilinear component analysis for nonlinear dimensionality reduction of hyperspectral images," in *Proceedings of the Image and Signal Processing for Remote Sensing VII*, vol. 4541, pp. 157–168, Toulouse, France.

- [15] L. Journaux, X. Tizon, I. Foucherot, and P. Gouton, "Dimensionality reduction techniques: An operational comparison on multispectral satellite images using unsupervised clustering," in *Proceedings of the 7th Nordic Signal Processing Symposium, NORSIG 2006*, pp. 242–245, Reykjavik, Iceland, June 2006.
- [16] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
- [17] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
- [18] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Kernel principal component analysis for feature reduction in hyperspectral images analysis," in *Proceedings of the 7th Nordic Signal Processing Symposium, NORSIG 2006*, pp. 238–241, Reykjavik, Iceland, June 2006.
- [19] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 783194, 2009.
- [20] H.-Y. Huang and B.-C. Kuo, "Double nearest proportion feature extraction for hyperspectral-image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4034–4046, 2010.
- [21] B.-C. Kuo, C.-H. Li, and J.-M. Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 4, pp. 1139–1155, 2009.
- [22] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 131–159, 2002.
- [23] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, "On kernel-target alignment," in *Advances in Neural Information Processing Systems*, pp. 367–373, 2002.
- [24] C. S. Ong, A. J. Smola, and R. C. Williamson, "Hyperkernels," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds., pp. 478–485, 2003.
- [25] J. T. Kwok and I. W. Tsang, "Learning with idealized kernels," in *Proceedings of the 20th International Conference on Machine Learning*, pp. 400–407, Washington, Wash, USA, 2003.
- [26] I. W.-H. Tsang and J. T.-Y. Kwok, "Efficient Hyperkernel learning using second-order cone programming," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 17, no. 1, pp. 48–58, 2006.
- [27] H. Binol, A. Bal, and H. Cukur, "Differential evolution algorithm-based kernel parameter selection for Fukunaga-Koontz Transform subspaces construction," in *Proceedings of the High-Performance Computing in Remote Sensing V*, vol. 9646, Toulouse, France, September 2015.
- [28] N. Li and Y. Yang, "Ensemble Kernel principal component analysis for improved nonlinear process monitoring," *Industrial & Engineering Chemistry Research*, vol. 54, no. 1, pp. 318–329, 2014.
- [29] D. Zhang, Z.-H. Zhou, and S. Chen, "Adaptive kernel principal component analysis with unsupervised learning of kernels," in *Proceedings of the 6th International Conference on Data Mining, ICDM 2006*, pp. 1178–1182, Hong Kong, China, December 2006.
- [30] F. S. Uslu, H. Binol, M. Ilarslan, and A. Bal, "Improving SVDD classification performance on hyperspectral images via correlation based ensemble technique," *Optics and Lasers in Engineering*, vol. 89, pp. 169–177, 2017.
- [31] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," *Connection Science*, vol. 8, no. 3-4, pp. 385–404, 1996.
- [32] S. Mao, L. Jiao, L. Xiong, and S. Gou, "Greedy optimization classifiers ensemble based on diversity," *Pattern Recognition*, vol. 44, no. 6, pp. 1245–1261, 2011.
- [33] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, vol. 16, The MIT Press, Cambridge, MA, USA, 2004.
- [34] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*, pp. 245–250, San Francisco, Calif, USA, August 2001.
- [35] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2625, 2008.
- [36] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 62, no. 3, pp. 236–248, 2007.
- [37] H. Binol, G. Bilgin, S. Dinc, and A. Bal, "Kernel Fukunaga-Koontz Transform Subspaces for Classification of Hyperspectral Images with Small Sample Sizes," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 6, pp. 1287–1291, 2015.
- [38] J. Ye, "Generalized low rank approximations of matrices," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 887–894, Banff, Canada, July 2004.
- [39] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the empirical feature space," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 16, no. 2, pp. 460–474, 2005.
- [40] T. Sun, L. Jiao, F. Liu, S. Wang, and J. Feng, "Selective multiple kernel learning for classification with ensemble strategy," *Pattern Recognition*, vol. 46, no. 11, pp. 3081–3090, 2013.
- [41] H. Binol, H. Cukur, and A. Bal, "A supervised discriminant subspaces-based ensemble learning for binary classification," *International Journal of Advanced Computer Research*, vol. 6, no. 27, pp. 209–214, 2016.
- [42] L. Wang, "Feature selection with kernel class separability," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1534–1546, 2008.
- [43] L. Wang and K. L. Chan, "Learning kernel parameters by using class separability measure," in *Proceedings of the Advances in Neural Information Processing Systems, Sixth workshop on Kernel Machines*, Canada, 2002.
- [44] R. I. Faulconbridge and M. R. Pickering, "Unsupervised band removal leading to improved classification accuracy of hyperspectral images," in *Proceedings of the 29th Australasian Computer Science Conference*, 2006.
- [45] O. Rajadell and P. Garcia-Sevilla, "Textural features for hyperspectral pixel classification," in *IbPRIA 2009: Pattern Recognition and Image Analysis*, vol. 5524 of *Lecture Notes in Computer Science*, pp. 208–216, Springer, Berlin, Germany, 2009.
- [46] C.-H. Li, B.-C. Kuo, C.-T. Lin, and C.-S. Huang, "A spatial-contextual support vector machine for remotely sensed image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 3, pp. 784–799, 2012.

- [47] J. Arenas-Garcia, K. B. Petersen, G. Camps-Valls, and L. K. Hansen, "Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 16–29, 2013.
- [48] H. Binol, F. S. Uslu, and A. Bal, "Unsupervised nonlinear feature extraction method and its effects on target detection," *International Journal of Electrical Electronics and Data Communication*, vol. 3, no. 8, pp. 43–46, 2015.
- [49] C. Chang and C. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.
- [50] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of the Optical Society of America A: Optics, Image Science & Vision*, vol. 14, no. 8, pp. 1724–1733, 1997.
- [51] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.
- [52] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, *Matlab Toolbox for Dimensionality Reduction*, MICC, Maastricht University, 2007.
- [53] S. Wang, "A practical guide to randomized matrix computations with MATLAB implementations," 2015, <https://arxiv.org/abs/1505.07570>.

