remote sensing

MDPI

Article

Discriminative Sparse Representation for Hyperspectral Image Classification: A Semi-Supervised Perspective

Zhaohui Xue [1,*], Peijun Du [2,3,4], Hongjun Su [1] and Shaoguang Zhou [1]

1 School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China; hjsu1@163.com (H.S.); zhousg1966@126.com (S.Z.)
2 Key Laboratory for Satellite Mapping Technology and Applications of National Administration of Surveying, Mapping and Geoinformation of China, Nanjing University, Nanjing 210023, China; dupjrs@gmail.com
3 Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing University, Nanjing 210023, China
4 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing University, Nanjing 210023, China
* Correspondence: zhaohui.xue@hhu.edu.cn

Academic Editors: Gonzalo Pajares Martinsanz and Prasad S. Thenkabail
Received: 27 November 2016; Accepted: 16 April 2017; Published: 19 April 2017

Abstract: This paper presents a novel semi-supervised joint dictionary learning (S^2JDL) algorithm for hyperspectral image classification. The algorithm jointly minimizes the reconstruction and classification error by optimizing a semi-supervised dictionary learning problem with a unified objective loss function. To this end, we construct a semi-supervised objective loss function which combines the reconstruction term from unlabeled samples and the reconstruction–discrimination term from labeled samples to leverage the unsupervised and supervised information. In addition, a *soft-max loss* is used to build the reconstruction–discrimination term. In the training phase, we randomly select the unlabeled samples and loop through the labeled samples to comprise the training pairs, and the first-order stochastic gradient descents are calculated to simultaneously update the dictionary and classifier by feeding the training pairs into the objective loss function. The experimental results with three popular hyperspectral datasets indicate that the proposed algorithm outperforms the other related methods.

Keywords: hyperspectral image classification; discriminative sparse representation; semi-supervised joint dictionary learning (S^2JDL); *soft-max loss*

1. Introduction

Hyperspectral remote sensing sensors can provide plenty of useful information that increases the accurate discrimination of spectrally similar materials of interest and allow for the acquisition of hundreds of contiguous bands for the same area on the surface of the Earth [1]. The acquired hyperspectral images have been extensively exploited for classification tasks [2–5], which aim at assigning each pixel with one thematic class for an object in a scene.

Recently, sparse representation has emerged as an effective way to solve various computer vision tasks, such as face recognition, image super-resolution, motion tracking, image segmentation, image denoising and inpainting, background modeling, photogrammetric stereo, and image classification [6], where the concept of *sparsity* often leads to state-of-the-art performances. Sparse representation has also been used for hyperspectral image classification [4,7], target detection [7,8], unmixing [9], pansharpening [10], image decomposition [11,12], and dimensionality reduction [13,14], where the high-dimensional pixel vectors can be sparsely represented by a few training samples (*atoms*) from

a given dictionary and the encoded sparse vectors carry out the class-label information. In order for sparse representation to yield superior performances, the desired dictionary should have good representation power [15].

Traditional methods for learning a representative and compact dictionary have been extensively exploited [16–19]. The method of optimal directions (MOD) [16] is an improvement of matching pursuit with an iterative optimization strategy, which iteratively calculates the optimal adjustment for the *atoms*. Singular value decomposition generalized from K-means (K-SVD) [17] is an iterative method that alternates between sparse coding of instances based on the current dictionary and updating the dictionary to better represent the data. The majorization method (MM) [18] is an optimization strategy that substitutes the original objective function with a surrogate function updated in each optimization step. The recursive least squares algorithm (RLS-DLA) [19] adopts a continued update approach as each *atom* is being processed. These methods have shown good performances in various computer vision tasks. However, these dictionary learning algorithms are designed for reconstruction tasks but not for the purpose of classification. Moreover, these algorithms often result in high computational complexity, less representative power for specific class, and lower discriminative power.

Advanced dictionary learning algorithms have been recently proposed to incorporate a discriminative term into the objective function in the dictionary learning problem [20–30], which has been regarded as discriminative sparse representation (DSR) in our previous work [31]. DSR allows for jointly learning reconstructive and discriminative parts instead of only the reconstructive one. Existing DSR models include the following categories:

(i) The seminal studies that paved the way for considering discrimination in dictionary learning. Mairal [20] adopted MOD and K-SVD to update the dictionary by using a truncated *Newton* iteration method. However, this method is not strictly convex and it did not explore the discrimination capability of sparse coefficients. Later, Mairal [22] adopted a *logistic loss* function to build a binary classification problem. This work also illustrated the possibility of extending the proposed binary classification problem to multi-class classification problems using *soft-max* loss function. Pham [21] designed a constrained optimization problem by adopting a linear classifier with *quadratic loss* and ℓ_2-norm regularization to jointly minimize the reconstruction and classification errors. This approach may suffer from a local minimum issue due to the fact that it iteratively alternates between reconstruction and classification terms.

(ii) Exploiting a different loss function and discriminative criterion. Lian [23] adopted a *hinge loss* function inspired from support vector machines (SVMs) to design a unified objective loss function that links classification with dictionary learning. Such a framework is able to further increase the margins of a binary classifier, which consequently decreases the error bound of the classifier. Yang [26] presented a novel discrimination dictionary learning (FDDL) method by using a *Fisher* discrimination criterion for penalizing the sparse coefficients, where a structured dictionary was used for minimizing the reconstruction error. Henao [32] developed a new Bayesian formulation for nonlinear SVM, based on a Gaussian process and with the *hinge loss* expressed as a scaled mixture of normals.

(iii) Incorporating K-SVD with a DSR model. Following the work [21], Zhang [25] proposed a discriminative K-SVD (D-KSVD) by incorporating the classification error term into the K-SVD-based objective function. However, this approach does not guarantee the ability of discrimination when acting on a small training set. In order to overcome this issue, Jiang [29] presented a label consistent K-SVD (LC-KSVD) algorithm, where the class-label information is associated with dictionary atoms to enforce discriminative property, and the optimal solution is efficiently obtained by using the K-SVD algorithm.

(iv) Exploiting the hybrid supervised and unsupervised DSR model. Lian [24] presented a probabilistic model that combines an unsupervised model (i.e., Gaussian mixture model) and a supervised model (i.e., logistic regression) for supervised dictionary learning. Marial [33] presented an online dictionary learning (OnlineDL) algorithm. Following this work, Zhang [30] presented an online

semi-supervised dictionary learning (OnlineSSDL) algorithm by optimizing the reconstruction error from labeled and unlabeled data, and the classification error from labeled data. However, for the sake of simplicity, OnlineSSDL droops the weight decay for classifier parameters and makes the problem strictly convex, resulting in a suboptimal solution.

(v) Exploiting structured *sparsity* in the DSR model. Compared with traditional sparse representation methods, structured *sparsity*-based methods are always more robust to noise due to the stability associated with group structure [34]. In addition, the structured *sparsity*-inducing dictionary learning methods require a smaller sample size to obtain the optimal solution [35–38]. Based on graph topology, Jiang [27] proposed a submodular dictionary learning (SDL) algorithm by optimizing an objective function that accounts for the entropy rate of random walk on a graph and a discriminative term. This dictionary learning problem can be considered as a graph partitioning problem, where the dictionary is updated by finding a graph topology that maximizes the objective function.

These works also stated that a good dictionary learning method should find a proper balance between reconstruction, discrimination, and compactness. Particularly, some studies have exploited DSR for hyperspectral image processing. Charles [39] modified an existing unsupervised learning method to learn the dictionary for hyperspectral image classification. Later, Castrodad [40] exploited DSR, where block-structured dictionary learning and subpixel unsupervised abundance mapping were jointly considered. More recently, Wang [41] designed a *hinge loss* function inspired from learning vector quantization to address the discriminative dictionary learning problem. Wang [42] proposed a semi-supervised classification method by jointly learning the classifier and dictionary in a task-driven framework, where *logistic loss* function is adopted to build the discriminative term. Our previous work presented a new method for DSR by learning a reconstructive dictionary and a discriminative classifier in a sparse representation model regularized with total variation [31].

Despite the good performances of these dictionary learning methods, some shortcomings can be observed. On the one hand, most of these approaches deal with supervised dictionary learning problems, and the performances of the learnt dictionary for classification greatly depend on the number of labeled samples. Unfortunately, the collection of labeled training samples is generally difficult, expensive and time-consuming, whereas unlabeled training samples can be generated in a much easier way, which has fostered the idea of exploiting semi-supervised learning (SSL) for hyperspectral image classification [43]. On the other hand, the loss function adopted in most of these approaches is *square loss* that considers classification as a regression problem. In addition, *square loss* often suffers from one critical flaw that the data outliers are punished too heavily when squaring the errors.

In this paper, we consider the above issues by jointly learning a reconstructive and discriminative dictionary in a semi-supervised fashion. To this end, we first employ a *soft-max loss* function to build the multi-class discriminative term to address the multi-class classification problem. Different from *square loss*, *soft-max loss* is overparameterized, which means for any hypothesis that needs to fit, there are multiple parameter settings giving rise to exactly the same hypothesis function mapping from inputs to the predictions. We then calculate the first-order stochastic gradient descents (SGD) [44] to simultaneously update the dictionary and classifier. The dictionary learning phase is iteratively performed in a semi-supervised learning fashion with the obtained labeled and unlabeled training pairs. The ultimate goal of this study is classification, while dictionary is an implicit variable when applying the proposed DSR model on hyperspectral image classification. Note that the recent study [42] is related to our work. However, we adopt *soft-max loss* to build the discriminative term, whereas [42] used *logistic loss*.

Although our previous studies [11,14,45,46] have exploited sparse representation for hyperspectral image classification, the methodologies are quite different from this work. Xue [11] focused on hyperspectral image decomposition for spectral-spatial classification, Xue [14] addressed hyperspectral image dimensionality reduction using sparse graph embedding, and Xue [45,46] exploited sparse graph regularization for hyperspectral image classification with very few labeled samples.

In this context, the main contribution of our work is the proposed semi-supervised joint dictionary learning (S^2JDL) algorithm, which leverages the information from labeled and unlabeled samples, allowing more accurate classification performance. Note that the proposed algorithm is unique compared to previously proposed approaches in the hyperspectral image classification community. In addition, we adopted a *soft-max loss* function to build the DSR problem, which is beneficial to hyperspectral image classification since the multi-class classification problem is very common in this community.

2. Background

Let $\mathbf{X} = [\mathbf{X}^l, \mathbf{X}^u] = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{n \times N}$ be a hyperspectral dataset with an n-dimensional signal for each pixel $\mathbf{x}_i = [x_1, ..., x_n]^{\mathrm{T}}, i \in 1, ..., N$. Let superscripts l and u be the labeled and unlabeled sample or dataset, and let the subscripts u and s be unsupervised or supervised objective loss function (i.e., Γ or $\mathcal{L}$). Let $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{m \times N}$ represent the label matrix for the input data, where the index of the nonzero value (i.e., 1) of $\mathbf{y}_i$ represents its label. Let $\mathcal{Y}(\mathbf{x}_i) \in \{1, ..., m\}$ denote the label of $\mathbf{x}_i$. Let $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_K] \in \mathbb{R}^{n \times K}$ be the dictionary. Let $\mathbf{W} = [\mathbf{w}_1^{\mathrm{T}}, ..., \mathbf{w}_m^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{m \times K}$ be the classifier. Let $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_N] \in \mathbb{R}^{K \times N}$ be the sparse coefficients for $\mathbf{X}$.

2.1. Sparse Representation

In the context of sparse representation, the sparse coefficients of $\mathbf{X}$ with respect to dictionary $\mathbf{D}$ can be obtained by optimizing an ℓ_1-norm regularization problem [6]

$$\arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_F^2 + \lambda\|\mathbf{z}_i\|_1, \tag{1}$$

where λ is a regularization parameter controlling the tradeoff between reconstruction error and *sparsity*.

2.2. Dictionary Learning for Classification

For various dictionary learning algorithms, the construction of $\mathbf{D}$ can be achieved by minimizing the reconstruction error and satisfying the *sparsity* constraint as

$$\arg\min_{\mathbf{D},\mathbf{Z}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_F^2 + \lambda\|\mathbf{z}_i\|_1. \tag{2}$$

K-SVD [17], which iteratively alternates between sparse coding and dictionary updating to better fit the data, is an efficient algorithm generalized from the K-means clustering process to solve Equation (2). However, K-SVD is not explicitly designed for classification tasks, as it only focuses on minimizing the reconstruction error.

Separating dictionary learning from classification may result in a suboptimal $\mathbf{D}$ for classification. Therefore, it is generally preferred to jointly learn the dictionary and classifier by solving [21,22,25,28–30]

$$\arg\min_{\mathbf{D},\mathbf{W},\mathbf{Z}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_F^2 + \frac{1}{N}\sum_{i=1}^{N}\Gamma[\mathbf{y}_i, f(\mathbf{z}^*(\mathbf{x}_i, \mathbf{D}), \mathbf{W})] + \frac{\lambda_1}{2}\|\mathbf{W}\|_F^2 + \lambda\|\mathbf{z}_i\|_1, \tag{3}$$

where the classifier can be obtained by optimizing the model parameter $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_m]^{\mathrm{T}} \in \mathbb{R}^{m \times K}$ as

$$\mathbf{W} = \arg\min_{\mathbf{W}} \frac{1}{N}\sum_{i=1}^{N}\Gamma[\mathbf{y}_i, f(\mathbf{z}^*(\mathbf{x}_i, \mathbf{D}), \mathbf{W})] + \frac{\lambda_1}{2}\|\mathbf{W}\|_F^2, \tag{4}$$

where Γ denotes the objective loss function which can be of *square loss*, *logistic loss*, *soft-max loss*, and *hinge loss* forms, $\mathbf{z}^*(\mathbf{x}_i, \mathbf{D})$ denotes the sparse code $\mathbf{z}_i$ obtained by solving Equation (1), and λ_1 is another regularization parameter preventing overfitting.

2.3. Related Work

Recently proposed joint dictionary learning methods mainly focus on supervised dictionary learning, which take the form [22]

$$\arg\min_{\mathbf{D},\mathbf{W},\mathbf{Z}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{Z}\|_F^2 + \frac{1}{N}\sum_{i=1}^{N}\{\Gamma[-\mathbf{y}_i, f(\mathbf{z}^*(\mathbf{x}_i,\mathbf{D}),\mathbf{W})] - \Gamma[\mathbf{y}_i, f(\mathbf{z}^*(\mathbf{x}_i,\mathbf{D}),\mathbf{W})]\} + \frac{\lambda_1}{2}\|\mathbf{W}\|_F^2 + \lambda\|\mathbf{z}_i\|_1. \quad (5)$$

However, Equation (5) is designed for binary classification with $\mathbf{y}_i \in \{[1\ 0]^{\mathrm{T}}; [0\ 1]^{\mathrm{T}}\}$.

K-SVD can be extended to discriminative K-SVD (D-KSVD) [25] by reconstructing an augmented dictionary with augmented training data, which can be formulated as

$$\begin{aligned} &\arg\min_{\mathbf{D},\mathbf{W},\mathbf{Z}} \alpha\|\mathbf{X}-\mathbf{D}\mathbf{Z}\|_F^2 + \beta\|\mathbf{Y}-\mathbf{W}\mathbf{Z}\|_F^2 + \lambda\|\mathbf{z}_i\|_1, \\ \Rightarrow &\arg\min_{\tilde{\mathbf{D}},\mathbf{W},\mathbf{Z}} \|\tilde{\mathbf{X}}-\tilde{\mathbf{D}}\mathbf{Z}\|_F^2 + \lambda\|\mathbf{z}_i\|_1, \end{aligned} \quad (6)$$

where $\tilde{\mathbf{X}} = [\sqrt{\alpha}\mathbf{X}^{\mathrm{T}}, \sqrt{\beta}\mathbf{Y}^{\mathrm{T}}]^{\mathrm{T}}$, $\tilde{\mathbf{D}} = [\sqrt{\alpha}\mathbf{D}^{\mathrm{T}}, \sqrt{\beta}\mathbf{W}^{\mathrm{T}}]^{\mathrm{T}}$, α and β are two scalars controlling the related contributions of the corresponding terms.

Recently, label consistent K-SVD (LC-KSVD) [29] has emerged as an effective way to solve Equation (6) by jointly adding a classification term and a label consistent regularization term into the *square loss* objective function, which is of the form

$$\arg\min_{\mathbf{D},\mathbf{W},\mathbf{G},\mathbf{Z}} \|\mathbf{X}-\mathbf{D}\mathbf{Z}\|_F^2 + \alpha\|\mathbf{Q}-\mathbf{G}\mathbf{Z}\|_F^2 + \beta\|\mathbf{Y}-\mathbf{W}\mathbf{Z}\|_F^2 + \lambda\|\mathbf{z}_i\|_1, \quad (7)$$

where the term $\|\mathbf{Q}-\mathbf{G}\mathbf{Z}\|_F^2$ signifies the discriminative sparse code error, $\mathbf{Q} = [\mathbf{q}_1, ..., \mathbf{q}_N] \in \mathbb{R}^{K\times N}$ refers to the discriminative sparse code (0 or 1) corresponding to input data $\mathbf{Z}$. The nonzero values (i.e., 1) of $\mathbf{q}_i = [q_i^1, ..., q_i^K]^{\mathrm{T}} \in \mathbb{R}^K$ occur at those indices where the input signal $\mathbf{z}_i$ and the dictionary *atom* $\mathbf{d}_k$ share the same label. $\mathbf{G}\in \mathbb{R}^{K\times K}$ is a linear transformation matrix, which transforms the original sparse code to be the most discriminative one in sparse feature space $\mathbb{R}^K$. Similar to D-KSVD, LC-KSVD is solved by using K-SVD with an augmented dictionary $\tilde{\mathbf{D}} = [\sqrt{\alpha}\mathbf{D}^{\mathrm{T}}, \sqrt{\beta}\mathbf{W}^{\mathrm{T}}, \sqrt{\gamma}\mathbf{G}^{\mathrm{T}}]^{\mathrm{T}}$ [29].

More recently, based on LC-KSVD, Zhang [30] tried to solve Equation (7) in an online SSL fashion by using a block-coordinate gradient descent algorithm to update the dictionary, which is named as OnlineSSDL and formulated as

$$\arg\min_{\mathbf{D},\mathbf{W},\mathbf{G},\mathbf{Z}} \|\mathbf{X}^u-\mathbf{D}\mathbf{Z}^u\|_F^2 + \alpha\|\mathbf{X}^l-\mathbf{D}\mathbf{Z}^l\|_F^2 + \beta\|\mathbf{Y}-\mathbf{W}\mathbf{Z}^l\|_F^2 + \gamma\|\mathbf{Q}-\mathbf{G}\mathbf{Z}^l\|_F^2 + \lambda\|\mathbf{z}_i\|_1. \quad (8)$$

Mairal [28] represented semi-supervised dictionary learning as an extension in the task-driven dictionary learning problem, which takes the form

$$\arg\min_{\mathbf{D},\mathbf{W}} (1-\mu)\mathbb{E}_{\mathbf{x}}\left[\Gamma_u(f(\mathbf{z}^*(\mathbf{x}^u,\mathbf{D})))\right] + \mu\mathbb{E}_{\mathbf{y},\mathbf{x}}\left[\Gamma_s(\mathbf{y}, f(\mathbf{z}^*(\mathbf{x}^l,\mathbf{D}),\mathbf{W}))\right] + \frac{\lambda_1}{2}\|\mathbf{W}\|_F^2, \quad (9)$$

where the loss functions Γ_s and Γ_u are, respectively, for supervised and unsupervised learning fashions, and $\mu \in (0,1)$ is a new parameter controlling the tradeoff between them.

However, Equation (9) adopts the *logistic loss* with a one-versus-all strategy and addresses classification as a regression problem, resulting in the scalability issues and large memory burden.

3. Proposed Method

In the proposed method, we first define a semi-supervised joint dictionary learning problem. Then, the optimization phase includes initialization, sparse coding, dictionary updating, and classifier updating. Finally, the class labels of unknown data are predicted by using the learnt classifier and the sparse coefficients. Figure 1 graphically illustrates the main idea.

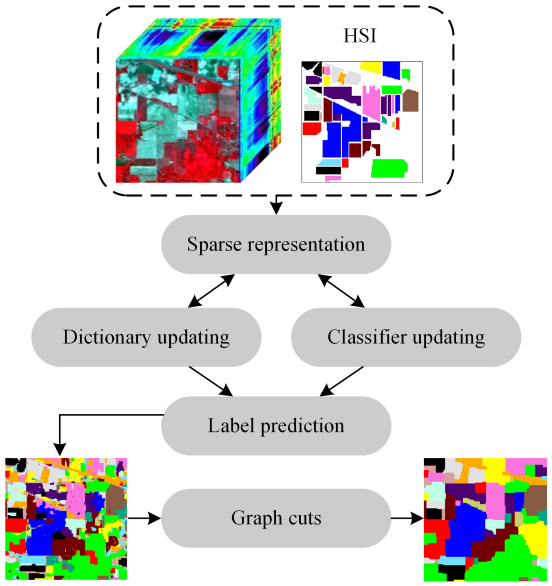


Figure 1. Graphical illustration of the proposed method.

3.1. Model Assumption

An attractive and promising research line for jointly learning the dictionary and classifier is to incorporate SSL. Inspired by the Equations (8) and (9), we now reformulate the semi-supervised joint dictionary learning problem into an improved form

$$\begin{aligned}&\arg\min_{\mathbf{D},\mathbf{W},\mathbf{Z}}(1-\mu)\Big\{\frac{1}{N}\sum_{i=1}^{N}\Gamma_u[f(\mathbf{z}^*(\mathbf{x}_i^u,\mathbf{D}))]+\lambda\psi(\mathbf{z}_i^u)\Big\}\\&+\mu\Big\{\frac{1}{2}\|\mathbf{X}^l-\mathbf{D}\mathbf{Z}^l\|_F^2+\frac{1}{N}\sum_{i=1}^{N}\Gamma_s[\mathbf{y}_i,f(\mathbf{z}^*(\mathbf{x}_i^l,\mathbf{D}),\mathbf{W})]+\frac{\lambda_1}{2}\|\mathbf{W}\|_F^2+\lambda\psi(\mathbf{z}_i^l)\Big\},\end{aligned}\tag{10}$$

where ψ is a *sparsity*-inducing function.

We adopt an ℓ_1-norm for ψ, and adopt a *soft-max loss* to design the supervised objective loss function, which takes the form

$$\Gamma_s[\mathbf{y}_i,f(\mathbf{z}^*(\mathbf{x}_i^l,\mathbf{D}),\mathbf{W})]\triangleq-\sum_{j=1}^{m}\mathbf{1}\{\mathcal{Y}(\mathbf{x}_i^l)=j\}\log\frac{\exp(\mathbf{w}_j^{\mathrm{T}}\mathbf{z}_i^l)}{\sum_{p=1}^{m}\exp(\mathbf{w}_p^{\mathrm{T}}\mathbf{z}_i^l)},\tag{11}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function, so that $\mathbf{1}\{\text{a true statement}\}=1$ and $\mathbf{1}\{\text{a false statement}\}=0$, and $j\in\{1,2,...,m\}$ refers to class.

Finally, the designed semi-supervised joint dictionary learning problem can be defined as

$$\begin{aligned}&\arg\min_{\mathbf{D},\mathbf{W},\mathbf{Z}}(1-\mu)\Big\{\frac{1}{2}\|\mathbf{X}^u-\mathbf{D}\mathbf{Z}^u\|_F^2+\lambda\|\mathbf{z}_i^u\|_1\Big\}+\mu\Big\{\frac{1}{2}\|\mathbf{X}^l-\mathbf{D}\mathbf{Z}^l\|_F^2\\&-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m}\mathbf{1}\{\mathcal{Y}(\mathbf{x}_i^l)=j\}\log\frac{\exp(\mathbf{w}_j^{\mathrm{T}}\mathbf{z}_i^l)}{\sum_{p=1}^{m}\exp(\mathbf{w}_p^{\mathrm{T}}\mathbf{z}_i^l)}+\frac{\lambda_1}{2}\|\mathbf{W}\|_F^2+\lambda\|\mathbf{z}_i^l\|_1\Big\}\end{aligned}\tag{12}$$

3.2. Optimization

3.2.1. Initialization

Let us assume that we have a small labeled dataset $\mathbf{X}^l$ spanning all classes and a large unlabeled dataset $\mathbf{X}^u$. Two variables need to be initialized since $\mathbf{Y}$ can be seen as a *prior*. For $\mathbf{D}_0$, we intend

to initialize such a dictionary in a way that its atoms are uniformly allocated to each class, with the number of atoms proportional to the dictionary size. Thus, we randomly select multiple class-specific dictionaries with equal size from the training data. The initialization process is completely supervised and the class labels attached to the dictionary remain fixed during the dictionary learning process. As for $\mathbf{W}_0$, we employ a multivariate ridge regression model [47] as

$$\arg\min_{\mathbf{W}} \tfrac{1}{2}\|\mathbf{Y}-\mathbf{W}\mathbf{Z}\|_F^2 + \tfrac{\lambda_1}{2}\|\mathbf{W}\|_F^2, \tag{13}$$

which is equipped with *square loss* and ℓ_2-norm regularization, and yields the following solution

$$\mathbf{W} = \mathbf{Y}\mathbf{Z}^{\mathrm{T}}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \tfrac{\lambda_1}{2}\mathbf{I}_K)^{-1}, \tag{14}$$

where $\mathbf{I}_K$ denotes the identity diagonal matrix with degree K.

We employ a spectral unmixing by variable splitting and augmented Lagrangian (*SUnSAL*) algorithm [48] to obtain the sparse code $\mathbf{Z}$ for the input data $\mathbf{X}$ with respect to the initialized dictionary $\mathbf{D}_0$. Then, the initial $\mathbf{W}_0$ can be computed by using Equation (14). Our previous studies have validated the good performance of *SUnSAL* for hyperspectral image processing [11,14,31,45,46].

3.2.2. Variables Updating

We can resort to the SGD algorithm to optimize Equation (12) since the objective loss function in our problem is highly nonlinear. In order to achieve semi-supervised optimization, we first regard the optimization process as two independent ingredients (i.e, unsupervised learning and supervised learning) and then calculate their gradients respectively. Then, we can combine the gradients with weighted summation strategy to obtain the final update.

At iteration t, we first select the t-th labeled sample $\mathbf{x}_t^l$ from $\mathbf{X}^l$. Assume currently, the dictionary $\mathbf{D}_t$ and the label matrix $\mathbf{y}_t$ are all given. We then randomly select an unlabeled sample $\mathbf{x}_t^u$ from $\mathbf{X}^u$. Next, we calculate the sparse codes $(\mathbf{z}_t^u, \mathbf{z}_t^l)$ for the training pair $(\mathbf{x}_t^u, \mathbf{x}_t^l)$ with respect to the current dictionary by adopting the *SUnSAL* algorithm. At last, $\mathbf{D}_t$ and $\mathbf{W}_t$ can be updated by feeding the training pair into the semi-supervised objective loss function. To this end, the gradients should be firstly formulated, which poses a critical challenge in optimizing the proposed dictionary learning problem.

For an unlabeled sample $\mathbf{x}_t^u$, assume $\mathcal{L}_u[f(\mathbf{z}^*(\mathbf{x}_t^u, \mathbf{D}_t))] \triangleq \frac{1}{2}\|\mathbf{x}_t^u - \mathbf{D}_t\mathbf{z}_t^u\|_2^2 + \lambda\|\mathbf{z}_t^u\|_1$, then we compute the gradients of $\mathcal{L}_u$ for $\mathbf{x}_t^u$ with respect to $\mathbf{D}_t$ as

$$\nabla_{\mathbf{D}_t}\mathcal{L}_u[f(\mathbf{z}^*(\mathbf{x}_t^u, \mathbf{D}_t))] = (1-\mu)(\mathbf{D}_t\mathbf{z}_t^u - \mathbf{x}_t^u)\mathbf{z}_t^{u\mathrm{T}}. \tag{15}$$

For a labeled sample $\mathbf{x}_t^l$, we solve the following problem

$$\arg\min_{\mathbf{D}_t, \mathbf{W}_t, \mathbf{z}_t^l} \frac{1}{2}\|\mathbf{x}_t^l - \mathbf{D}_t\mathbf{z}_t^l\|_2^2 + \lambda\|\mathbf{z}_t^l\|_1 - \sum_{j=1}^{m} \mathbf{1}\{\mathcal{Y}(\mathbf{x}_t^l) = j\} \log \frac{\exp(\mathbf{w}_j^{\mathrm{T}}\mathbf{z}_t^l)}{\sum_{p=1}^{m}\exp(\mathbf{w}_p^{\mathrm{T}}\mathbf{z}_t^l)} + \frac{\lambda_1}{2}\|\mathbf{W}_t\|_F^2. \tag{16}$$

The skeleton optimization process of the presented semi-supervised joint dictionary learning (S^2JDL) method is summarized in Algorithm 1. More details for the dictionary and classifier updating can be found in the Appendix A.

Algorithm 1 Semi-Supervised Joint Dictionary Learning (S^{2}JDL).

Input: $\mathbf{X}^l$, $\mathbf{X}^u$, K, T, λ, λ_1, λ_2, μ, ρ.
Output: $\mathbf{D}$ and $\mathbf{W}$.
Initialization: Initialize $\mathbf{D}_0$ with K atoms from $\mathbf{X}^l$ and obtain $\mathbf{W}_0$ by Equation (14).
for each $\mathbf{x}_t^l$ in $\mathbf{X}^l$ **do**
  Randomly select an unlabeled sample $\mathbf{x}_t^u$ from $\mathbf{X}^u$.
  Obtain $(\mathbf{z}_t^l, \mathbf{z}_t^u)$ for $(\mathbf{x}_t^l, \mathbf{x}_t^u)$ with the current dictionary $\mathbf{D}_t$ using *SUnSAL*.
  Find the support set Λ_t for $\mathbf{z}_t^l$.
  Calculate the learning rate: $\rho_t \leftarrow \min(\rho, \rho t_0/t)$.
  Update $\mathbf{D}$ and $\mathbf{W}$ by Equations (A4) and (A5), respectively.
  Remove the selected unlabeled sample: $\mathbf{X}^u \leftarrow \mathbf{X}^u \setminus \mathbf{x}_t^u$.
end for
return $\mathbf{D}_{t+1}$ and $\mathbf{W}_{t+1}$.

3.3. Classification

Once we have obtained the learnt dictionary $\widehat{\mathbf{D}}$ and the classifier parameter $\widehat{\mathbf{W}}$ from Algorithm 1, we can predict a new incoming test sample $\mathbf{x}_{test}$. To this end, we first compute its sparse code $\mathbf{z}_{test}$ using *SUnSAL* and then assign its label by the position corresponding to the largest value (also the most possible value) in the label vector by

$$\mathcal{Y}(\mathbf{x}_{test}) = \mathcal{H}(\mathbf{x}_{test}, \widehat{\mathbf{W}}) \triangleq \arg\min_j \widehat{\mathbf{W}}\mathbf{z}_{test}. \tag{17}$$

Since the proposed method can produce probability as Equation (17), we adopt graph cuts [49] to obtain smoother classification maps. Graph cuts are as an energy minimization algorithm, which can tackle the combinatorial optimization problem involving unary and pairwise interaction terms, i.e., the maximum a posteriori (MAP) segmentation problem using multinomial logistic regression with multilevel logistic prior. Graph cuts can yield very good approximations to the MAP segmentation and are quite efficient from a computational point of view [50]. Overall accuracy (OA) and the Kappa value (κ) generated from the confusion matrix are used to evaluate the classification performance based on the ground-truth [51]. In addition, the Kappa [52] and McNemar z-score statistical tests [53] are also adopted for accuracy assessment.

4. Experimental Results and Discussion

4.1. Experimental Settings

For performance comparison, some strongly related dictionary learning algorithms are considered. The unsupervised methods are MOD, K-SVD, D-KSVD, and OnlineDL. The supervised category includes LC-KSVD and SDL. We also implemented a semi-supervised method, semi-supervised joint dictionary learning with logistic loss function (S^{2}JDL-Log, "Log" for Logistic loss function), which is a variant of the proposed method. We then denote by S^{2}JDL-Sof ("Sof" for soft-max loss function) the proposed method to differentiate them. It is worth noting that all the methods employ a multivariate ridge regression model for classifier learning (see Equation (14)), and adopt orthogonal matching pursuit (OMP) [54] for sparse coding.

It is worth noting that MOD, K-SVD, D-KSVD, LC-KSVD, OnlineDL, and SDL are implemented based on their original source codes released to the public by their owners, and they share the common parameters of sparsity level ($T = 5$), reconstruction error ($E = 1 \times 10^{-4}$), and ℓ_1-norm penalty ($\lambda = 1 \times 10^{-5}$). In addition, three other parameters in SDL have been set as the recommended values. We also note that these parameters have been selected by careful cross-validation. In this context, the parameter settings always ensure a fair comparison even if they are suboptimal. For dictionary and classifier update, the initial learning rate is set to $\rho = 1 \times 10^{-3}$, and the tradeoff between the unsupervised and supervised learning ingredients is set to 0.5.

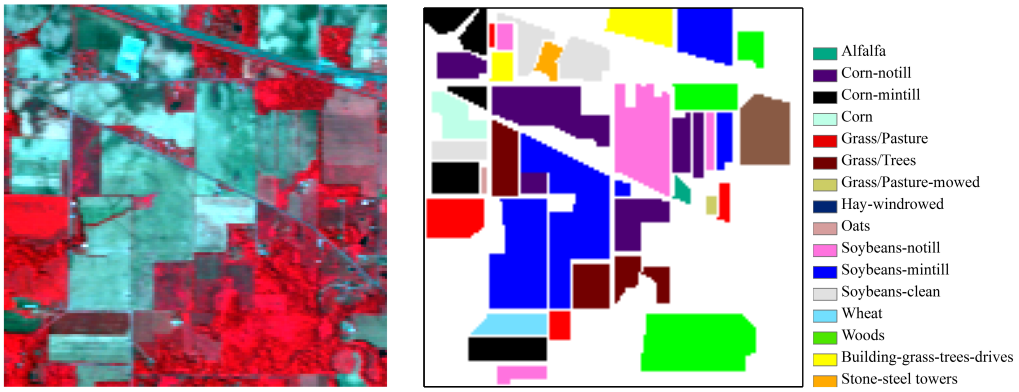
We randomly select labeled training samples from the ground-truth to initialize the dictionary and classifier. The ℓ_2-norm penalty parameter is set to $\lambda_1 = 1 \times 10^{-5}$ when initializing the classifier using Equation (14).

For classification, we set the maximum number of iterations to $k_{max} = 10$, which means that the reported overall accuracies (OAs), average accuracies (AAs), kappa statistic (κ), and class individual accuracies are derived by averaging the results after conducting ten independent Monte Carlo runs with respect to the initialized labeled training set. In addition, the smoothness parameter (τ) in graph cuts is set to 3, and we adopt graph cuts for all the methods.

Finally, it is worth noting that all the implementations were carried out using Matlab R2015b on a desktop PC equipped with an Intel Xeon E3 CPU (at 3.4 GHz) and 32 GB of RAM. It should be emphasized that all the results are derived from our own implementations and the involved parameters for these methods have been carefully optimized in order to ensure a fair comparison.

4.2. Hyperspectral Datasets

Three real hyperspectral datasets [55] collected by Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and the Reflective Optics Spectrographic Imaging System (ROSIS) are used in our experiments. The first hyperspectral image was acquired by the AVIRIS sensor over the Indian Pines region in northwestern Indiana in 1992. The image size in pixels is 145 × 145, with moderate spatial resolution of 20 m. The number of data channels in the acquired image is 220 (with spectral range from 0.4 to 2.5 μm). A total of 200 radiance channels are used in the experiments by removing several noisy and water absorbed bands. A three-band false color composite image and the ground-truth map are shown in Figure 2. A total of 10,366 samples containing 16 classes are available.



(**a**) (**b**)

Figure 2. (**a**) False color composition of the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) Indian Pines scene (R: 57, G: 27, B: 17); (**b**) Ground truth-map containing 16 mutually exclusive land-cover classes.

The second hyperspectral image was acquired by the ROSIS sensor over the urban area of the University of Pavia, Italy. The image size in pixels is 610 × 340, with very high spatial resolution of 1.3 m. The number of data channels in the acquired image is 103 (with spectral range from 0.43 to 0.86 μm). A three-band false color composite image and the ground-truth map are shown in Figure 3. A total of 42,776 samples containing nine classes are available.

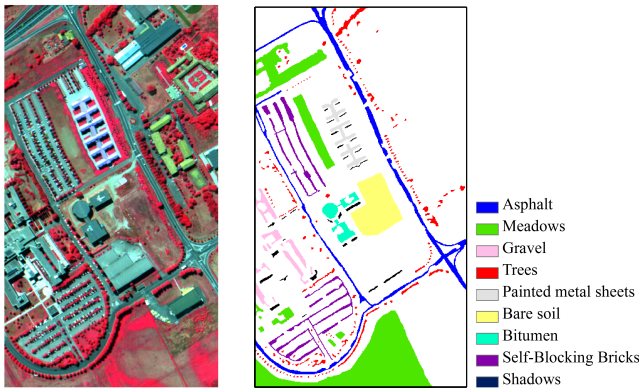


Figure 3. (**a**) False color composition of the Reflective Optics Spectrographic Imaging System (ROSIS) University of Pavia scene (R: 102, G: 56, B: 31); (**b**) Ground truth map containing nine mutually exclusive land-cover classes.

The third hyperspectral image was acquired by the AVIRIS sensor over Salinas Valley in southern California, USA. The image size in pixels is 512 × 217, with a spatial resolution of 3.7 m. The number of data channels in the acquired image is 224 (with spectral range from 0.4 to 2.5 µm). A total of 204 radiance channels are used in the experiments by removing the noisy and water absorbed bands. A three-band false color composite image and the ground-truth map are shown in Figure 4. A total of 54,129 samples containing 16 classes are available.

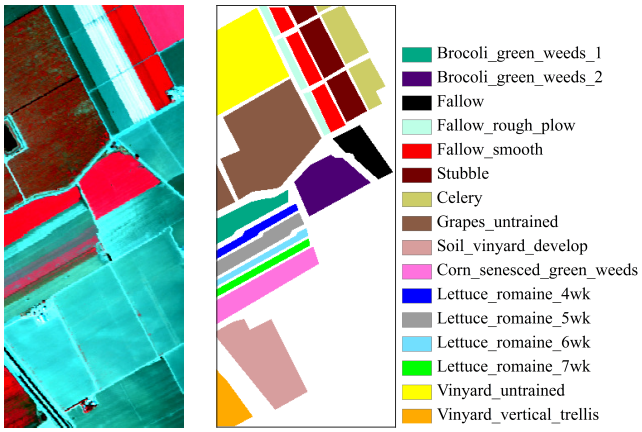


Figure 4. (**a**) False color composition of the AVIRIS Salinas scene (R: 57, G: 27, B: 17); (**b**) Ground truth map containing 16 mutually exclusive land-cover classes.

4.3. Experiments with AVIRIS Indian Pines Dataset

Experiment 1: We first analyze the sensitivity of parameters for the proposed method. Figure 5 illustrates the sensitivity of the proposed method to parameters λ, λ_1, and μ under the condition that 10% of labeled samples per class are used for training and 15 labeled samples per class are used to build the dictionary. As we can see from Figure 5a, the OA is insensitive to λ and λ_1. Therefore, we roughly set $\lambda = 1 \times 10^{-5}$ and $\lambda_1 = 1 \times 10^{-5}$. This observation is reasonable since λ controls the uniqueness of sparse coding and λ_1 determines the initial performance of the classifier parameter. The impacts of the two parameters are reduced in a iterative learning scheme since sparse coding and classifier update are alternatively conducted. According to Figure 5b, we found that OA is sensitive to μ, and the optimal value is set to $\mu = 3$. This observation is in accordance with that made in [50].

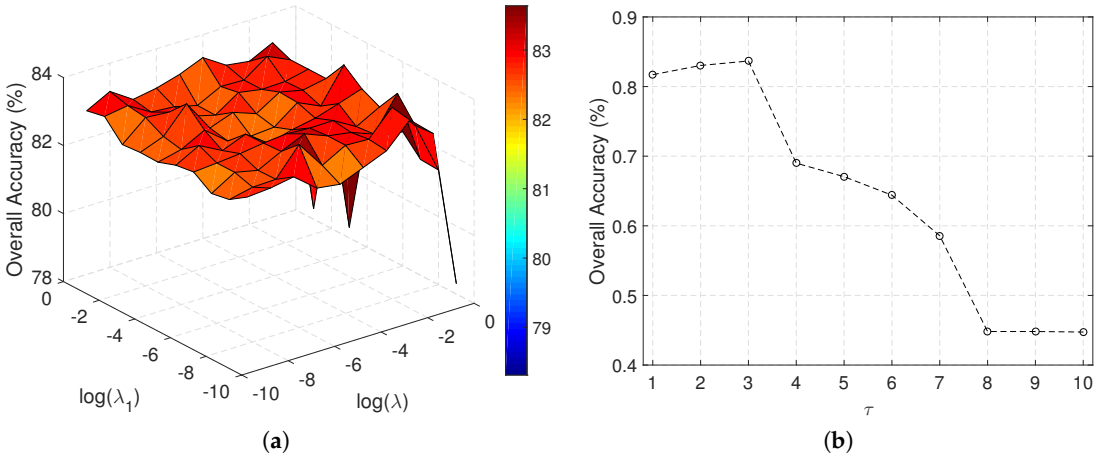


Figure 5. Parameter sensitivity analysis of the proposed method for the AVIRIS Indian Pines dataset (10% of labeled samples per class are used for training and 15 labeled samples per class are used to build the dictionary). (**a**) Overall accuracy (OA) as a function of λ and λ_1; (**b**) Overall accuracy (OA) as a function of τ. (**a**) OA vs. λ and λ_1; (**b**) OA vs. τ.

Experiment 2: In this test, Gaussian random noise with a pre-defined signal-to-noise ratio (SNR) (SNR $\triangleq 10\log_{10}\left(\mathbb{E}[\|\mathbf{X}\|_2^2]/\mathbb{E}[\|N\|_2^2]\right)$) is generated and added to the original imagery. Figure 6 illustrates the evolution of OA with SNR for different classifiers. As shown in the figure, the proposed method outperforms others in most cases with SNR = 5, 10, etc. The interval between the curve of the proposed method and others visually indicates the significances, which confirms the robustness of the proposed method to data noise.

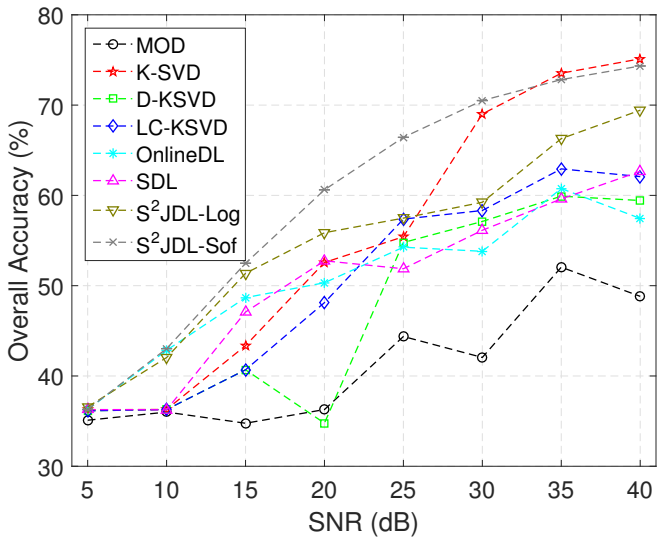


Figure 6. The evolution of overall accuracy with SNR for different classifiers (10% of labeled samples per class are used for training and 15 labeled samples per class are used to build the dictionary).

Experiment 3: We then analyze the impact of training data size on classification accuracy. To this end, we randomly choose 10% of labeled samples per class (a total of 1027 samples) to initialize the training data and evaluate the impact of the number of atoms on the classification performance. Figure 7 shows the OAs as a function of the number of atoms per class obtained by different methods. As we can see from the figure, the proposed method obtains the highest accuracies in all cases. Another observation is that, as the number of atoms increases, the OAs also increase. When the number of

atoms per class is equal to 15, the proposed method reaches a stable level, with an OA higher than 80%. It is interesting to note that D-KSVD and LC-KSVD obtain very similar results.

Figure 8 shows the OAs as a function of the ratio of labeled samples per class for different methods. As we can see from the figure, the proposed method obtains higher accuracies when the ratio is larger than 5%. It is worth noting that the proposed method can stably obtain improved classification performance with additional labeled samples. However, the other methods cannot produce higher OAs as the ratio ranges from 3% to 10%. This observation can be explained by the increase of the number of labeled samples; the proposed method can exploit more information from both the labeled and unlabeled training samples, whereas the supervised dictionary learning methods can only rely on the labeled information, and the performance cannot be improved significantly. Another observation is that S^2JDL-Log yields a very competitive classification performance. This is due to the fact that S^2JDL-Log is based on semi-supervised learning fashion.

Experiment 4: Table 1 reports the OA, AA, individual classification accuracies, and κ statistic. The results of the proposed algorithm are listed in the last column of the table and we have marked in bold typeface the best result for each class and the best results of OA, AA, and κ statistic. Our method achieves the best results compared to the other supervised dictionary learning methods. The improvements of classification accuracy are around 10–40% when compared to other methods. Especially, when classifying the class *Wheat*, our method performs very well. Although our method may not always obtain the best accuracy in some specific classes, AA, OA, and kappa are more convincing metrics measuring the classification performance. In addition, the time costs of different methods are listed in the table, where we can see that the proposed method is more efficient than K-SVD, D-KSVD, and LC-KSVD. However, MOD, OnlineDL, and SDL take less time. The time cost of the proposed method mainly comes from the optimization step, which can be reduced by exploiting more efficient optimization strategy.

Table 2 reports the statistical tests between-classifier in terms of Kappa z-score and McNemar z-score. The critical value of z-score is 1.96 at a confidence level of 0.95, and all the tests are significant with 95% confidence, which indicates that the proposed method significantly outperforms the other methods. Another observation is that the lower value of z-score demonstrates the closer classification results, e.g., the Kappa z-score value for S^2JDL-Log/S^2JDL-Sof is 4.4. Similar observation can be made for the tests using McNemar z-score.

For illustrative purposes, Figure 9 shows the obtained classification maps (corresponding to the best one after ten Monte Carlo runs). The advantages obtained by adopting the semi-supervised dictionary learning approach with regard to the corresponding supervised and unsupervised cases can be visually appreciated in the classification maps displayed in Figure 9, where the classification OAs obtained for each method are reported in the parentheses. Compared to the ground-truth map, the proposed method obtains a more accurate and smoother classification map. Significant differences can be observed when classifying the class *Wheat* in this scene, which is in accordance with the former results. However, the classification maps obtained by D-KSVD and LC-KSVD are much less homogeneous than the other methods. This observation can be explained by the fact that Graph cuts are adopted as a post processing strategy, which largely relies on the initial classification probabilities obtained by the classifiers. If the initial classification results are poor, then the classification improvements may not be satisfied. That is the case for D-KSVD and LC-KSVD with an initial OA = 60.07% for the former and an initial OA = 60.52% for the latter.

Table 1. Overall (OA), average (AA), and individual class accuracies (%), kappa statistics (κ), and the standard deviation of ten conducted Monte Carlo runs obtained for different classification methods for the AVIRIS Indian Pines dataset with a balanced training set (10% of labeled samples per class are used for training and 15 labeled samples per class are used to build the dictionary).

Class	#Samples		MOD	K-SVD	D-KSVD	LC-KSVD	OnlineDL	SDL	S^{2}JDL-Log	S^{2}JDL-Sof
	Train	Test								
Alfalfa	5	41	3.66 ± 11.57	0.00 ± 0.00	17.80 ± 17.36	**33.90 ± 30.60**	10.24 ± 31.55	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Corn-notill	143	1285	14.44 ± 8.94	58.44 ± 11.20	40.60 ± 6.87	33.70 ± 6.59	36.90 ± 15.71	59.28 ± 7.61	**83.03 ± 5.67**	82.40 ± 3.89
Corn-mintill	83	747	0.00 ± 0.00	48.63 ± 9.39	26.85 ± 11.35	29.84 ± 11.72	19.38 ± 26.51	53.13 ± 5.31	50.96 ± 12.25	**55.85 ± 5.85**
Corn	24	213	0.05 ± 0.15	2.35 ± 5.52	13.85 ± 8.73	15.02 ± 15.53	0.00 ± 0.00	**15.35 ± 29.16**	1.08 ± 3.41	5.77 ± 5.11
Grass-pasture	48	435	0.02 ± 0.07	71.91 ± 19.54	70.46 ± 12.43	64.85 ± 17.43	11.03 ± 10.64	24.34 ± 1.78	85.03 ± 3.84	**88.41 ± 4.44**
Grass-trees	73	657	83.84 ± 20.12	99.83 ± 0.24	88.57 ± 10.82	87.15 ± 12.77	89.82 ± 31.56	**99.73 ± 0.51**	98.83 ± 0.95	99.44 ± 0.46
Grass-pasture-mowed	3	25	0.80 ± 1.69	26.80 ± 43.17	**38.00 ± 44.51**	24.80 ± 39.05	2.40 ± 7.59	26.00 ± 42.09	0.00 ± 0.00	0.00 ± 0.00
Hay-windrowed	48	430	75.19 ± 32.08	**100.00 ± 0.00**	97.88 ± 2.22	96.95 ± 3.59	89.81 ± 31.56	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**
Oats	2	18	0.00 ± 0.00	0.00 ± 0.00	**8.89 ± 18.56**	7.78 ± 10.54	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Soybeans-notill	97	875	0.17 ± 0.54	73.65 ± 1.82	49.38 ± 7.25	49.20 ± 14.91	20.11 ± 27.13	**74.26 ± 1.51**	55.66 ± 8.59	53.69 ± 5.87
Soybeans-mintill	246	2209	**99.84 ± 0.36**	99.49 ± 0.67	79.00 ± 10.66	81.21 ± 8.01	88.47 ± 31.25	98.25 ± 2.34	98.71 ± 2.11	97.15 ± 0.76
Soybean-clean	59	534	0.00 ± 0.00	20.64 ± 13.06	26.70 ± 12.92	36.25 ± 13.04	2.68 ± 8.47	14.38 ± 12.53	55.21 ± 26.95	**75.67 ± 8.67**
Wheat	21	184	9.84 ± 31.11	99.13 ± 0.46	75.49 ± 22.95	93.15 ± 5.13	87.88 ± 30.91	87.12 ± 30.79	**99.29 ± 0.52**	99.18 ± 0.46
Woods	127	1138	99.85 ± 0.18	**99.88 ± 0.22**	92.50 ± 4.30	92.30 ± 4.69	89.53 ± 31.46	**99.88 ± 0.11**	99.85 ± 0.25	99.67 ± 0.32
Bldg-grass-tree-drives	39	347	0.06 ± 0.12	31.73 ± 16.71	15.48 ± 9.09	23.95 ± 12.60	9.86 ± 17.51	2.05 ± 6.47	59.91 ± 15.95	**60.35 ± 5.73**
Stone-steel-towers	9	84	56.79 ± 36.55	**99.64 ± 0.80**	99.05 ± 0.75	87.38 ± 30.90	78.69 ± 29.03	34.29 ± 44.78	37.98 ± 49.15	98.57 ± 2.68
Average accuracy	-	-	27.78 ± 4.50	58.26 ± 4.07	52.53 ± 4.10	53.59 ± 4.82	39.80 ± 12.01	49.25 ± 3.78	57.85 ± 4.35	**63.51 ± 0.66**
Overall accuracy	-	-	48.48 ± 3.35	75.79 ± 2.18	62.11 ± 2.83	62.65 ± 1.97	55.02 ± 19.54	71.77 ± 1.54	80.46 ± 2.52	**82.25 ± 1.08**
κ statistic	-	-	0.365 ± 0.04	0.715 ± 0.03	0.561 ± 0.03	0.567 ± 0.02	0.478 ± 0.17	0.667 ± 0.02	0.771 ± 0.03	**0.793 ± 0.01**
Time (Seconds)	-	-	2.64 ± 0.41	138.76 ± 8.02	148.11 ± 6.20	143.17 ± 9.03	1.96 ± 0.50	**0.83 ± 0.13**	53.97 ± 4.66	59.64 ± 4.67

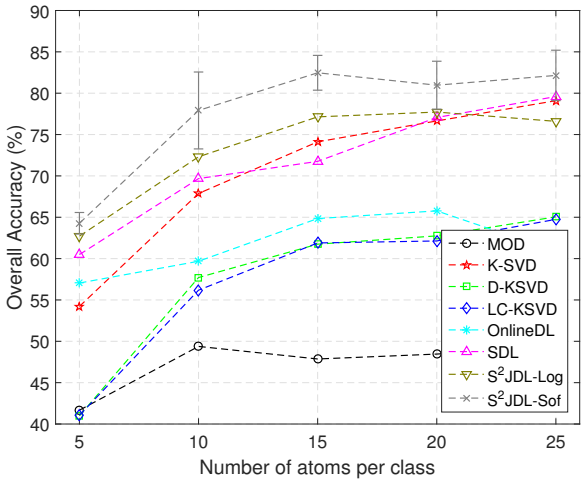


Figure 7. Overall accuracy (OA) as a function of the number of atoms per class for the AVIRIS dataset (10% of labeled samples per class are used for training). Error bars indicate the standard deviations obtained by the proposed method.

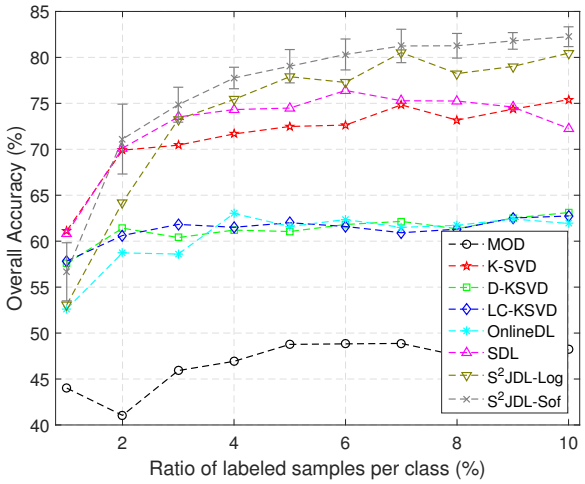


Figure 8. Overall accuracy (OA) as a function of the ratio of labeled samples per class for the AVIRIS Indian Pines dataset (15 labeled samples per class are used to build the dictionary). Error bars indicate the standard deviations obtained by the proposed method.

Table 2. Pairwise statistical test in terms of Kappa z-score and McNemar z-score for the AVIRIS Indian Pines dataset with a balanced training set (10% of labeled samples per class are used for training and 15 labeled samples per class are used to build the dictionary).

Between-Classifier	κ (z-score)	McNemar (z-score)
MOD/S^{2}JDL-Sof	57.4	3.3×10^3
K-SVD/S^{2}JDL-Sof	12.2	2.7×10^2
D-KSVD/S^{2}JDL-Sof	31.7	1.3×10^3
LC-KSVD/S^{2}JDL-Sof	32.3	1.3×10^3
OnlineDL/S^{2}JDL-Sof	45.2	2.5×10^3
SDL/S^{2}JDL-Sof	18.4	5.8×10^2
S^{2}JDL-Log/S^{2}JDL-Sof	4.4	5.5×10^1

The critical value of z-score is 1.96 at a confidence level of 0.95, and all the tests are significant with 95% confidence.

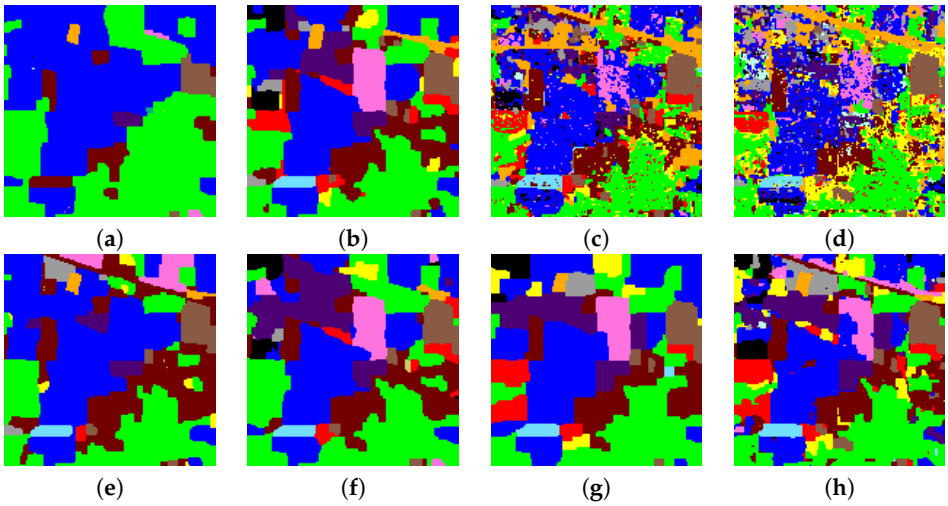


Figure 9. Classification maps obtained by different methods for the AVIRIS Indian Pines dataset. The OA in each case is reported in the parentheses. (**a**) MOD (48.48%); (**b**) K-SVD (75.79%); (**c**) D-KSVD (62.11%); (**d**) LC-KSVD (62.65%); (**e**) OnlineDL (55.02%); (**f**) SDL (71.77%); (**g**) S^2JDL-Log (80.46%); (**h**) S^2JDL-Sof (82.25%).

Experiment 5: In the last experiment, we analyze the mechanism of the proposed method. Firstly, we plot the stem distributions of sparse coefficients obtained by different methods. As we can see from Figure 10, the distributions between-classifier are significantly different. Precisely, the corresponding atoms belonging to the same land cover type contribute more than the others, thus making the associated coefficients sparsely locate at those atoms. For example, the atoms indexed by 146–160 in the dictionary belong to the class *Wheat*, and the sparse coefficients will mainly locate at these atoms if this class is well represented. Obviously, the proposed method produces more accurate sparse coefficients since the stem distributions mainly locate at the corresponding atoms (see Figure 10g). As for the other methods, the associated sparse atoms cannot be accurately found, i.e., the stem distributions obtained by OnlineDL partially locate at the class *Woods*. Figure 11 spatially exhibits the sparse coefficients relative to the class *Wheat* for different methods. As shown in Figure 11, the proposed method yields more accurate sparse coefficients relative to the class *Wheat*. Therefore, the aggregation characteristics of sparse coefficients naturally enlarge the discrimination between different land cover types, and the spatial variations of sparse coefficients explain the accuracy of the prosed method for sparse representation. The above observations demonstrate the good performance of the proposed method in dictionary learning, and the discrimination performance of our method has been validated in the former experiments.

Secondly, we analyze the denoising power of the proposed method by plotting the original spectrum, the reconstructed spectrum, and the noise for the class *Wheat*. From the results shown in Figure 12, we can see the original spectrum can be accurately reconstructed with a very small Root-Mean-Square Error (RMSE) (The RMSE for two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ can be defined as: $\text{RMSE} = \text{sqrt}\left(\sum_{b=1}^{B}(x_{i,b} - x_{j,b})^2 / B\right)$), which is the difference between the original spectrum ($\mathbf{x}$) and the reconstructed spectrum ($\mathbf{Dz}$). It is worth noting that the proposed method obtains a very small RMSE value. In this context, the proposed method can accurately reconstruct the original spectrum with high fidelity by removing noise, which explains the robustness to noise of the proposed method in the former experiment. We then evaluate the global reconstruction performance of the proposed method by considering all classes. As reported in Table 3, the proposed method obtains the smallest RMSE value. This experiment also hints at the good performance of the proposed method in dictionary learning.

Table 3. Global reconstruction performance for different methods by considering all classes.

	MOD	K-SVD	D-KSVD	LC-KSVD	OnlineDL	SDL	S^{2}JDL-Log	S^{2}JDL-Sof
RMSE	2.60×10^{-3}	1.72×10^{-4}	1.53×10^{-4}	1.25×10^{-4}	3.69×10^{-4}	3.05×10^{-4}	3.45×10^{-5}	$\mathbf{3.42} \times 10^{-5}$

Finally, we attempt to analyze the dictionary structure by visually illustrating the matrix **D** learnt by different methods. As shown in Figure 13, S^{2}JDL-Sof and S^{2}JDL-Log yield similar data structure, and the atoms belonging to the same class are more similar to each other, while the atoms belonging to different classes are more distinctive between each other. Similar observations can be made for D-KSVD, LC-KSVD, and SDL. However, the dictionaries learnt by OnlineDL model very little data structure, as shown in the figure. Note that we cannot currently explain the factors inducing the differences of dictionary structure.

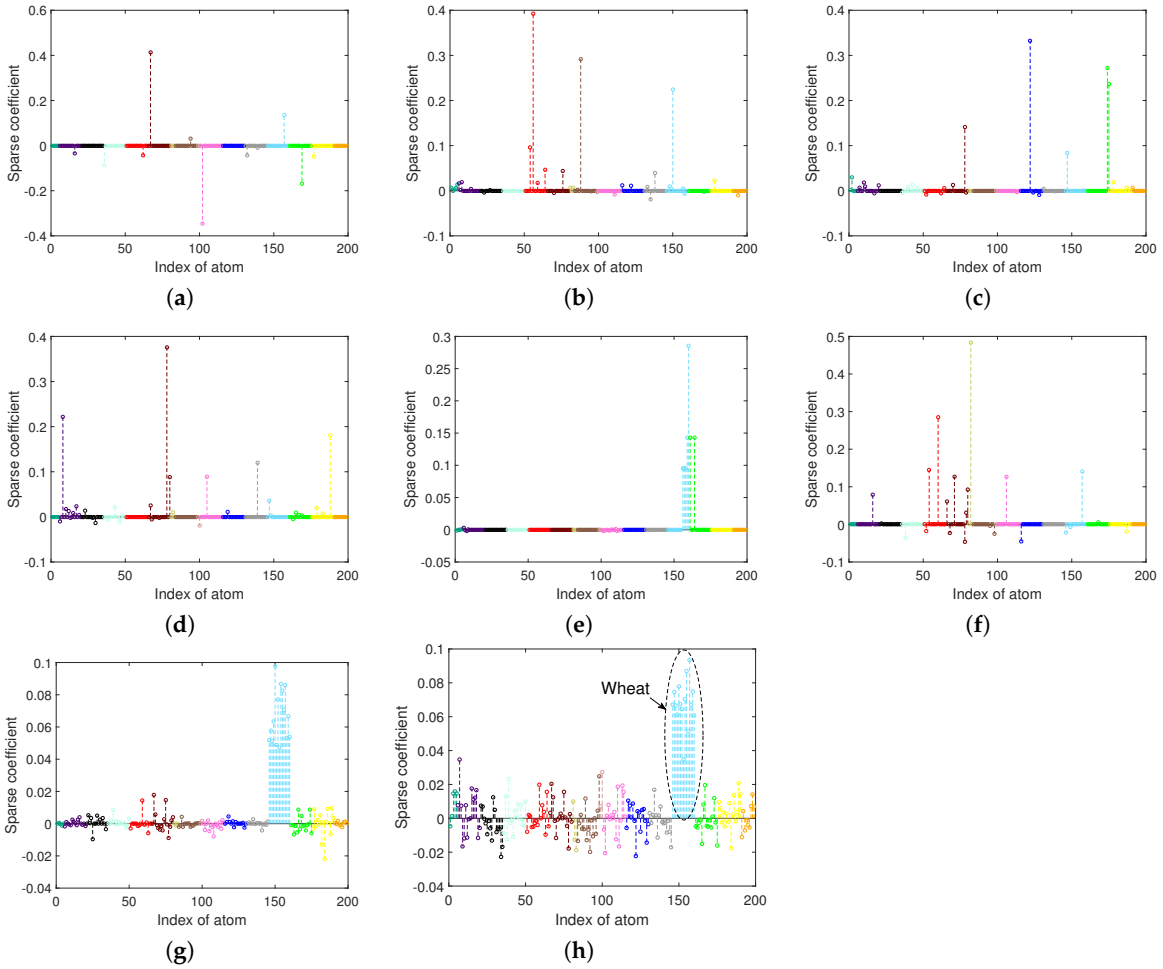


Figure 10. Stem distributions of sparse coefficients relative to the class *Wheat* obtained by different methods for the AVIRIS Indian Pines dataset. The circles terminating different stems represent the sparse coefficients relative to the associated atoms which are marked with different colors representing different classes. (**a**) MOD; (**b**) K-SVD; (**c**) D-KSVD; (**d**) LC-KSVD; (**e**) OnlineDL; (**f**) SDL; (**g**) S^{2}JDL-Log; (**h**) S^{2}JDL-Sof.

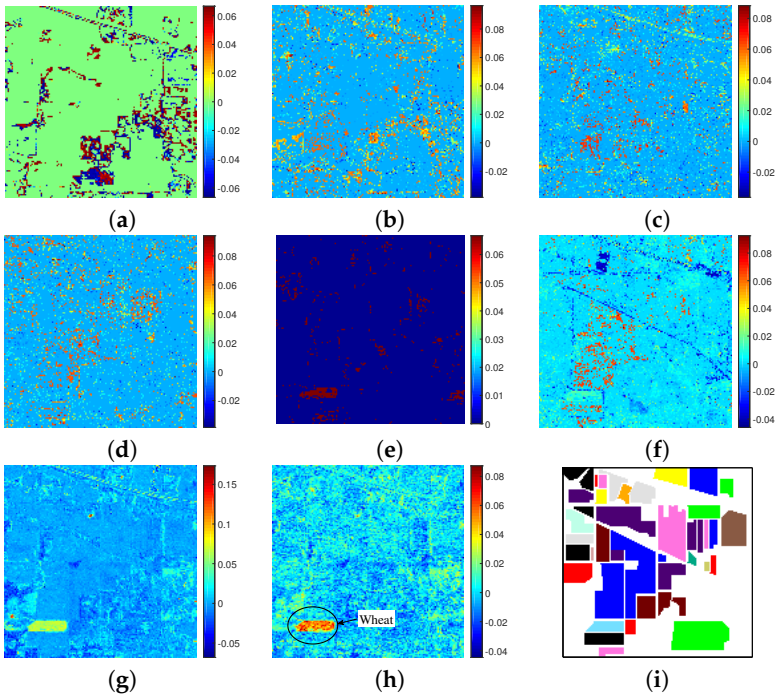


Figure 11. Graphical illustration of sparse coefficients relative to the class *Wheat* obtained by different methods for the AVIRIS Indian Pines dataset. (**a**) MOD; (**b**) K-SVD; (**c**) D-KSVD; (**d**) LC-KSVD; (**e**) OnlineDL; (**f**) SDL; (**g**) S^{2}JDL-Log; (**h**) S^{2}JDL-Sof; (**i**) Ground-truth.

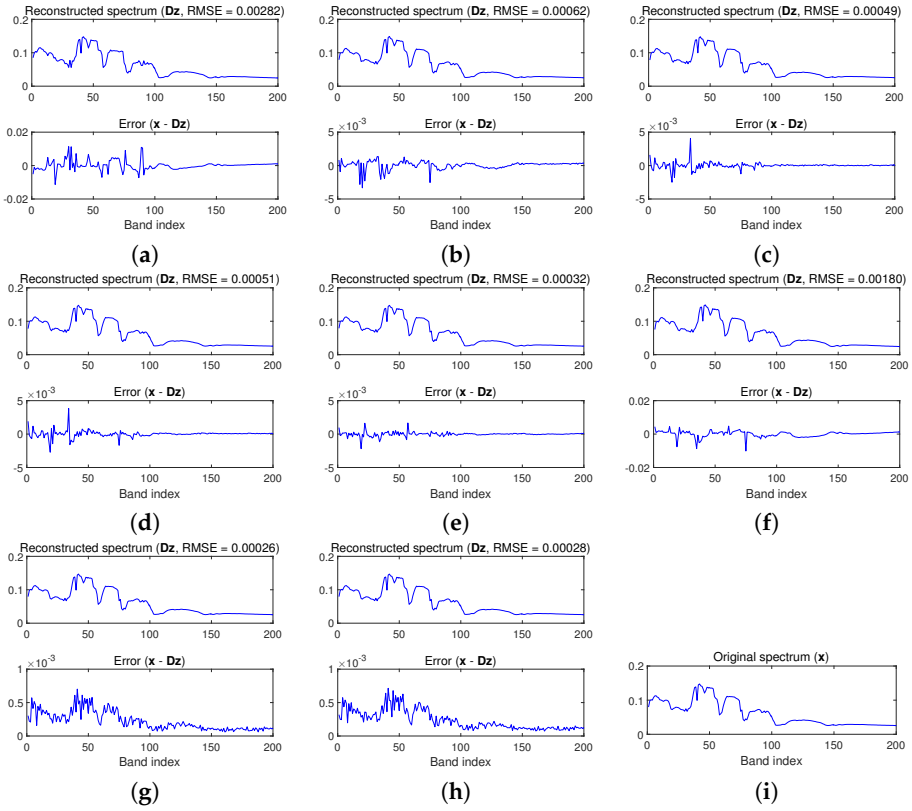


Figure 12. Reconstruction and denoising power of sparse representation for different methods by taking the class *Wheat* as an example. The original spectrum (top), reconstructed spectrum with RMSE value (middle), and noise (bottom) are given for each case. (**a**) MOD; (**b**) K-SVD; (**c**) D-KSVD; (**d**) LC-KSVD; (**e**) OnlineDL; (**f**) SDL; (**g**) S^{2}JDL-Log; (**h**) S^{2}JDL-Sof; (**i**) Original.

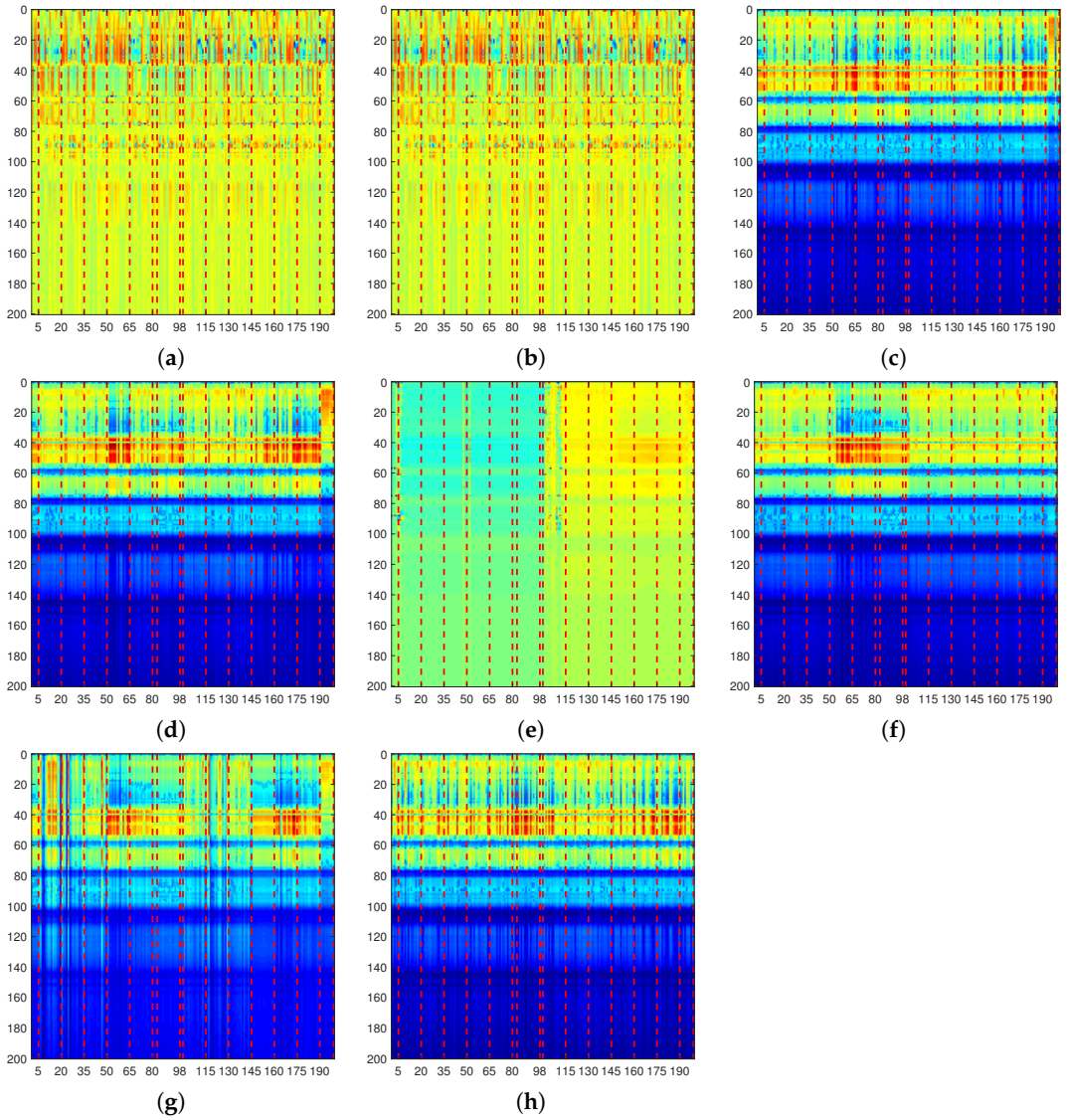


Figure 13. Graphical illustration of the dictionary structure learnt by different methods. The vertical dashed lines in each figure separate different atoms belonging to different classes. (**a**) MOD; (**b**) K-SVD; (**c**) D-KSVD; (**d**) LC-KSVD; (**e**) OnlineDL; (**f**) SDL; (**g**) S^2JDL-Log; (**h**) S^2JDL-Sof.

4.4. Experiments with ROSIS University of Pavia Dataset

Experiment 1: We first analyze the impact of training data size on classification accuracy. We randomly choose 5% of labeled samples per class (a total of 2138 samples) to initialize the training data and evaluate the impact of the number of atoms on classification performance achieved by the proposed method for the ROSIS University of Pavia dataset. Figure 14 shows the OAs as a function of the number of atoms per class obtained by different methods. Again, the proposed method obtains the highest accuracies in all cases. Another observation is that, for most of the methods, the OAs increase as the number of atoms also increase. Different from the former experiments, when the number of atoms per class is larger than 10, the OAs obtained by D-KSVD and LC-KSVD become lower. In this scene, MOD does not perform very well in all cases.

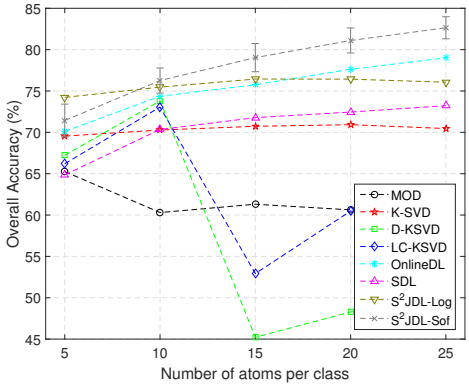


Figure 14. Overall accuracy (OA) as a function of the number of atoms per class for the ROSIS University of Pavia dataset (5% of labeled samples per class are used for training). Error bars indicate the standard deviations obtained by the proposed method.

Figure 15 depicts the OAs as a function of the ratio of labeled samples per class for different methods. As we can see from the figure, the proposed method obtains the highest accuracies in all cases. It is interesting to note that the proposed method cannot stably obtain improved classification performances with additional labeled samples. This may be due to the fact that the homogeneity in this scene is not so significant, and graph cuts reduce the effects of the learning phase since the classification accuracy cannot be significantly improved by using additional labeled samples. In addition, similar observations can be made for other methods. In this scene, D-KSVD does not perform very well in different cases. Again, S^2JDL-Log provides competitive performance.

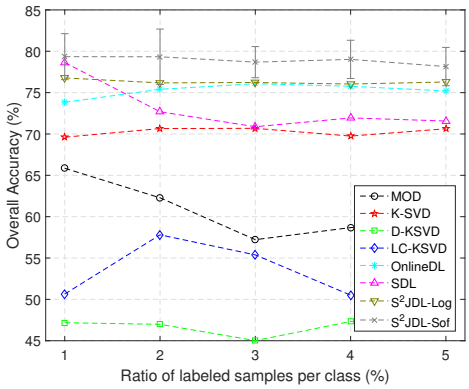


Figure 15. Overall accuracy (OA) as a function of the ratio of labeled samples per class for the ROSIS University of Pavia dataset (15 labeled samples per class are used to build the dictionary). Error bars indicate the standard deviations obtained by the proposed method.

Experiment 2: Table 4 lists the OA, AA, individual classification accuracies, and κ statistic. As reported in the table, the proposed method achieves the best results compared to the other supervised dictionary learning methods. The improvements of classification accuracy are around 3–30% when compared to other methods. When classifying the class *Bare soil*, our method obtains the highest accuracy, with an OA of 23.10%. Although this accuracy is not very high, it demonstrates the merit of the proposed method since *Bare soil* is very difficult to accurately classify. In addition, the time costs of different methods are listed in the table, where we can see that the proposed method is more efficient than K-SVD, D-KSVD, and LC-KSVD. Again, MOD, OnlineDL, and SDL take less time.

Table 4. Overall (OA), average (AA), and individual class accuracies (%), kappa statistics (κ), and the standard deviation of ten conducted Monte Carlo runs obtained for different classification methods for the ROSIS University of Pavia dataset with a balanced training set (5% of labeled samples per class are used for training and 15 labeled samples per class are used to build the dictionary).

Class	#Samples		MOD	K-SVD	D-KSVD	LC-KSVD	OnlineDL	SDL	S^2JDL-Log	S^2JDL-Sof
	Train	Test								
Asphalt	332	6299	78.61 ± 21.49	93.67 ± 1.51	11.92 ± 4.72	11.50 ± 14.42	98.58 ± 0.95	**99.32 ± 0.38**	98.22 ± 0.52	98.38 ± 0.98
Meadows	932	177,17	**100.00 ± 0.00**	**100.00 ± 0.00**	87.58 ± 8.56	81.89 ± 15.62	**100.00 ± 0.00**	**100.00 ± 0.00**	99.80 ± 0.38	100.00 ± 0.01
Gravel	105	1994	0.00 ± 0.00	15.31 ± 21.13	9.35 ± 10.97	13.23 ± 17.28	14.06 ± 14.40	8.57 ± 14.87	23.61 ± 19.49	**42.54 ± 9.39**
Trees	153	2911	2.70 ± 4.50	50.84 ± 4.37	16.54 ± 11.44	27.02 ± 19.63	36.81 ± 8.29	55.36 ± 3.77	**73.48 ± 2.40**	67.28 ± 6.93
Painted metal sheets	67	1278	99.66 ± 0.40	97.68 ± 0.90	27.07 ± 32.75	93.32 ± 10.51	99.49 ± 0.34	99.05 ± 0.72	**99.87 ± 0.13**	99.86 ± 0.20
Bare soil	251	4778	0.70 ± 1.56	22.12 ± 2.65	6.64 ± 7.37	15.23 ± 10.49	21.52 ± 4.09	18.28 ± 5.98	**23.73 ± 2.57**	23.10 ± 4.96
Bitumen	67	1263	0.00 ± 0.00	0.00 ± 0.00	17.03 ± 21.09	**32.71 ± 37.04**	5.22 ± 16.50	6.67 ± 14.36	0.00 ± 0.00	13.93 ± 18.64
Self-Blocking Bricks	184	3498	1.50 ± 2.96	24.08 ± 6.00	10.61 ± 12.66	3.69 ± 3.70	**61.70 ± 14.99**	40.54 ± 12.37	60.43 ± 6.44	59.83 ± 14.27
Shadows	47	900	22.44 ± 20.08	0.00 ± 0.00	**100.00 ± 0.00**	99.89 ± 0.05	84.66 ± 7.02	1.63 ± 3.89	0.00 ± 0.00	72.27 ± 11.78
Average accuracy	-	-	33.96 ± 3.77	44.85 ± 2.18	31.86 ± 6.67	42.05 ± 7.46	58.00 ± 2.69	47.71 ± 2.87	53.24 ± 2.53	**64.13 ± 2.70**
Overall accuracy	-	-	59.82 ± 3.64	70.25 ± 1.09	46.96 ± 4.07	48.34 ± 9.47	75.21 ± 1.45	72.37 ± 1.50	76.29 ± 1.24	**78.79 ± 1.10**
κ statistic	-	-	0.382 ± 0.08	0.572 ± 0.02	0.337 ± 0.04	0.365 ± 0.09	0.645 ± 0.02	0.603 ± 0.02	0.665 ± 0.02	**0.701 ± 0.02**
Time (Seconds)	-	-	8.30 ± 2.68	341.32 ± 42.90	93.09 ± 8.54	100.12 ± 9.40	6.17 ± 1.20	**6.02 ± 0.90**	521.07 ± 27.39	79.58 ± 8.17

Table 5 also reports the statistical tests between-classifier in terms of Kappa z-score and McNemar z-score. Again, the results indicate that the proposed method significantly outperforms the other methods. In this scene, we observe that OnlineDL is closer to our method, i.e., the Kappa z-score value for OnlineDL/S^{2}JDL is 14.2, which is in accordance with the results reported in Table 4.

Table 5. Pairwise statistical test in terms of Kappa z-score and McNemar z-score for the ROSIS University of Pavia dataset with a balanced training set (5% of labeled samples per class are used for training and 15 labeled samples per class are used to build the dictionary).

Between-Classifier	κ (z-score)	McNemar (z-score)
MOD/S^{2}JDL-Sof	74.5	7.6×10^3
K-SVD/S^{2}JDL-Sof	33.9	2.8×10^3
D-KSVD/S^{2}JDL-Sof	98.8	1.2×10^4
LC-KSVD/S^{2}JDL-Sof	93.4	1.2×10^4
OnlineDL/S^{2}JDL-Sof	14.2	5.6×10^2
SDL/S^{2}JDL-Sof	24.5	2.5×10^3
S^{2}JDL-Log/S^{2}JDL-Sof	8.9	2.0×10^2

The critical value of z-score is 1.96 at a confidence level of 0.95, and all the tests are significant with 95% confidence.

Figure 16 visually depicts the obtained classification maps. The advantages obtained by adopting the semi-supervised dictionary learning approach with regard to the corresponding supervised case can be visually appreciated in the classification maps displayed in Figure 16, which also reports the classification OAs obtained for each method in the parentheses. As shown in the figure, the homogeneity is very clear for this scene, and the proposed method depicts a more accurate and smoother classification map. As expected, D-KSVD and LC-KSVD obtain poor classification maps for this scene.

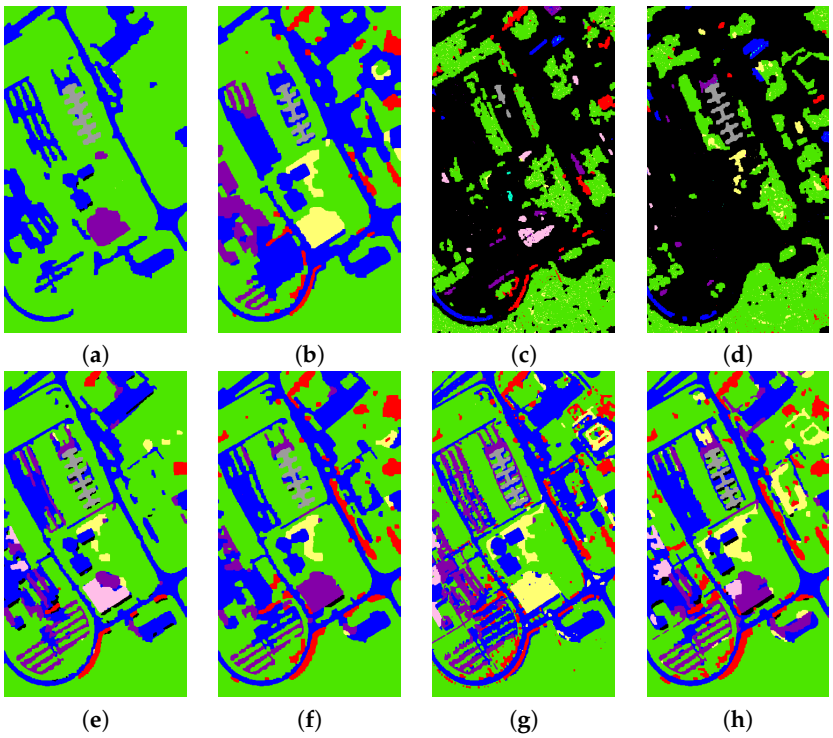


Figure 16. Classification maps obtained by different methods for the ROSIS University of Pavia dataset. The OA in each case is reported in the parentheses. (**a**) MOD (59.82%); (**b**) K-SVD (70.25%); (**c**) D-KSVD (46.96%); (**d**) LC-KSVD (48.34%); (**e**) OnlineDL (75.21%); (**f**) SDL (72.37%); (**g**) S^{2}JDL-Log (76.29%); (**h**) S^{2}JDL-Sof (78.79%).

4.5. Experiments with AVIRIS Salinas Dataset

Experiment 1: Similarly, we first analyze the impact of training data size on classification accuracy. A total of 5% of labeled samples per class (a total of 2706 samples) are randomly selected to initialize the training data. We evaluate the impact of the number of atoms on the classification performance achieved by the proposed method for the AVIRIS Salinas dataset. Figure 17 shows the OAs as a function of the number of atoms per class obtained by different methods. Similar to the experiments implemented for the AVIRIS Indian Pines dataset, the OAs become stable when 15 atoms per class are used to build the dictionary. Another observation is that when the number of atoms per class is larger than 10, our method stably outperforms the other methods. It is interesting to note that D-KSVD and LC-KSVD obtain similar results even when the latter is incorporated with the class-label information. For most of the cases, the OAs increase as the number of atoms also increases.

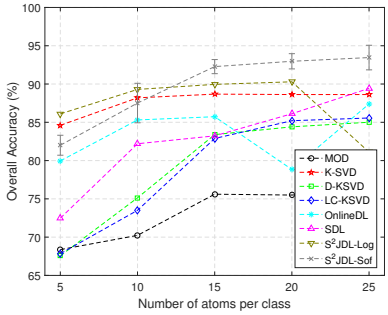


Figure 17. Overall accuracy (OA) as a function of the number of atoms per class for the AVIRIS Salinas dataset (5% of labeled samples per class are used for training). Error bars indicate the standard deviations obtained by the proposed method.

Figure 18 plots the OAs as a function of the ratio of labeled samples per class for different methods. As we can see from the figure, the proposed method obtains the highest accuracies in all cases. Different from the former experiments, the proposed method can stably obtain improved classification performance with the additional labeled samples in this scene. However, the additional labeled samples deteriorate the classification performance for SDL.

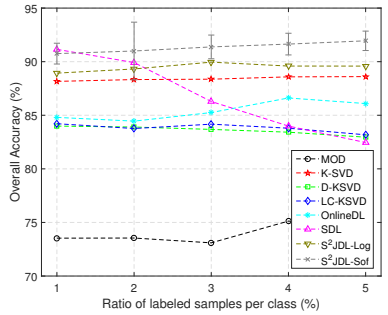


Figure 18. Overall accuracy (OA) as a function of the ratio of labeled samples per class for the AVIRIS Salinas dataset (15 labeled samples per class are used to build the dictionary). Error bars indicate the standard deviations obtained by the proposed method.

Experiment 2: Table 6 gives the OA, AA, individual classification accuracies, and κ statistic. As reported in the table, the proposed method achieves the best results compared to the other supervised dictionary learning methods. The improvements of classification accuracy are around 10–20% when compared to the other methods. As for the specific classification accuracy, the proposed method obtains higher accuracy when classifying the class *Lettuce_romaine_6wk*. In addition, the time costs of different methods are listed in the table, where we can see that the proposed method is more efficient than D-KSVD and LC-KSVD. Again, MOD, OnlineDL, and SDL take less time.

Table 6. Overall (OA), average (AA), and individual class accuracies (%), kappa statistics (κ), and the standard deviation of ten conducted Monte Carlo runs obtained for different classification methods for the AVIRIS Salinas dataset with a balanced training set (5% of labeled samples per class are used for training and 15 labeled samples per class are used to build the dictionary).

Class	#Samples		MOD	K-SVD	D-KSVD	LC-KSVD	OnlineDL	SDL	S^2JDL-Log	S^2JDL-Sof
	Train	Test								
Brocoli_green_weeds_1	100	1909	99.25 ± 1.02	99.13 ± 0.91	97.52 ± 1.33	97.77 ± 1.15	99.66 ± 0.75	**99.98 ± 0.07**	99.18 ± 0.22	99.70 ± 0.38
Brocoli_green_weeds_2	186	3540	99.41 ± 0.87	90.19 ± 31.02	96.98 ± 4.93	98.16 ± 0.48	99.90 ± 0.16	98.63 ± 0.84	**100.00 ± 0.00**	99.25 ± 0.27
Fallow	99	1877	3.31 ± 7.36	98.00 ± 5.98	84.61 ± 14.29	92.10 ± 6.24	31.13 ± 10.68	80.61 ± 6.51	**100.00 ± 0.00**	90.51 ± 5.46
Fallow_rough_plow	70	1324	52.39 ± 39.19	4.46 ± 5.79	99.27 ± 0.55	99.19 ± 1.09	86.62 ± 23.04	99.32 ± 0.28	83.05 ± 3.93	**99.49 ± 0.33**
Fallow_smooth	134	2544	74.28 ± 41.09	**99.94 ± 0.07**	93.06 ± 6.20	90.39 ± 5.85	99.49 ± 0.37	99.24 ± 0.44	99.33 ± 0.11	98.79 ± 0.53
Stubble	198	3761	99.60 ± 0.24	99.90 ± 0.03	99.84 ± 0.13	99.84 ± 0.11	99.37 ± 0.23	**99.99 ± 0.01**	99.91 ± 0.02	99.97 ± 0.04
Celery	179	3400	93.70 ± 15.59	99.94 ± 0.03	97.24 ± 2.62	98.96 ± 0.68	**99.92 ± 0.08**	99.31 ± 0.21	99.90 ± 0.02	99.48 ± 0.26
Grapes_untrained	564	10,707	99.35 ± 0.34	95.35 ± 0.80	66.13 ± 7.08	67.38 ± 4.39	99.17 ± 1.09	**99.46 ± 0.27**	94.01 ± 0.85	98.21 ± 2.19
Soil_vinyard_develop	310	5893	99.92 ± 0.11	**100.00 ± 0.00**	97.39 ± 0.74	97.36 ± 0.86	99.89 ± 0.14	99.06 ± 0.90	**100.00 ± 0.00**	99.57 ± 0.39
Corn_senesced_green_weeds	164	3114	73.17 ± 26.15	**95.28 ± 0.85**	76.27 ± 7.63	76.61 ± 9.85	83.24 ± 3.73	83.16 ± 7.18	95.17 ± 0.49	94.24 ± 3.21
Lettuce_romaine_4wk	53	1015	42.24 ± 47.68	94.56 ± 0.89	91.28 ± 5.15	92.71 ± 3.77	94.09 ± 3.76	30.38 ± 41.43	94.47 ± 0.41	**95.79 ± 3.66**
Lettuce_romaine_5wk	96	1831	95.09 ± 6.40	**99.95 ± 0.02**	97.13 ± 3.94	96.35 ± 10.33	99.49 ± 0.72	98.96 ± 1.23	93.01 ± 8.53	99.89 ± 0.28
Lettuce_romaine_6wk	46	870	78.51 ± 31.42	0.00 ± 0.00	76.45 ± 40.62	90.97 ± 20.62	95.38 ± 2.64	96.93 ± 3.85	4.71 ± 7.70	**97.83 ± 0.37**
Lettuce_romaine_7wk	54	1016	68.47 ± 34.98	**97.92 ± 0.44**	82.44 ± 29.82	83.97 ± 10.14	94.29 ± 3.34	85.03 ± 22.89	97.08 ± 0.31	96.41 ± 1.82
Vinyard_untrained	363	6905	0.01 ± 0.02	57.27 ± 2.35	**58.95 ± 4.15**	56.54 ± 6.23	38.54 ± 7.07	3.77 ± 11.28	53.14 ± 2.29	55.73 ± 10.69
Vinyard_vertical_trellis	90	1717	81.68 ± 10.05	**99.46 ± 0.10**	89.17 ± 12.00	91.77 ± 8.42	88.53 ± 9.35	94.27 ± 4.22	98.78 ± 0.25	97.92 ± 0.64
Average accuracy	-	-	72.52 ± 6.55	83.21 ± 2.38	87.73 ± 2.64	89.38 ± 2.17	88.04 ± 1.81	85.51 ± 2.80	88.24 ± 0.93	**95.17 ± 0.82**
Overall accuracy	-	-	75.35 ± 2.96	87.89 ± 2.37	82.90 ± 1.47	83.56 ± 1.41	86.88 ± 1.48	82.98 ± 1.81	89.59 ± 0.53	**92.50 ± 1.63**
κ statistic	-	-	0.720 ± 0.03	0.865 ± 0.03	0.810 ± 0.02	0.817 ± 0.02	0.853 ± 0.02	0.808 ± 0.02	0.884 ± 0.01	**0.916 ± 0.02**
Time (Seconds)	-	-	11.93 ± 3.42	448.47 ± 25.98	617.16 ± 23.52	630.04 ± 31.32	5.72 ± 0.64	**5.38 ± 0.92**	593.68 ± 38.96	500.24 ± 16.38

Similarly, we conduct the statistical tests between-classifier in terms of Kappa z-score and McNemar z-score in this scene. According to the results reported in Table 7, we observe that the proposed method significantly outperforms the other methods since all the tests are significant with 95% confidence. Another observation is that K-SVD and the proposed method produce similar results, with the Kappa z-score value of 22.9.

Table 7. Pairwise statistical test in terms of Kappa z-score and McNemar z-score for the AVIRIS Salinas dataset with a balanced training set (5% of labeled samples per class are used for training and 15 labeled samples per class are used to build the dictionary).

Between-Classifier	κ (z-score)	McNemar (z-score)
MOD/S^{2}JDL-Sof	77.8	8.7×10^3
K-SVD/S^{2}JDL-Sof	22.9	9.3×10^2
D-KSVD/S^{2}JDL-Sof	46.7	3.0×10^3
LC-KSVD/S^{2}JDL-Sof	42.3	2.3×10^3
OnlineDL/S^{2}JDL-Sof	29.5	2.3×10^3
SDL/S^{2}JDL-Sof	44.7	4.1×10^3
S^{2}JDL-Log/S^{2}JDL-Sof	14.6	4.6×10^2

The critical value of z-score is 1.96 at a confidence level of 0.95, and all the tests are significant with 95% confidence.

The classification maps are given in Figure 19, where the OAs obtained for each method are reported in the parentheses. As shown in the figure, we can see clear differences between different methods. For example, when classifying the class *Lettuce_romaine_6wk*, the proposed method is more accurate compared to the other methods. In addition, the homogeneity is very clear for this scene, which is similar to the AVIRIS Indian Pines dataset. Also, our method produces a more accurate and smoother classification map compared to the other methods.

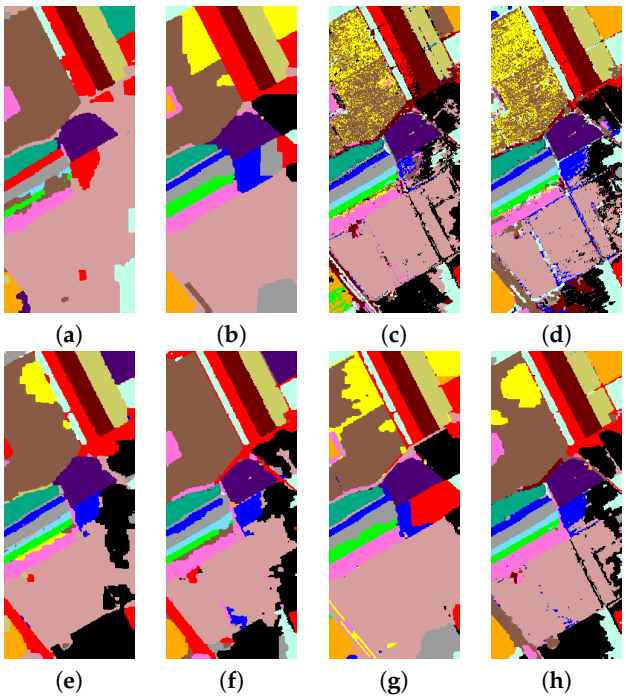


Figure 19. Classification maps obtained by different methods for the AVIRIS Salinas dataset. The OA in each case is reported in the parentheses. (**a**) MOD (75.35%); (**b**) K-SVD (87.89%); (**c**) D-KSVD (82.90%); (**d**) LC-KSVD (83.56%); (**e**) OnlineDL (86.88%); (**f**) SDL (82.98%); (**g**) S^{2}JDL-Log (89.59%); (**h**) S^{2}JDL-Sof (92.50%).

5. Conclusions

In this paper, we have developed a novel semi-supervised algorithm for jointly learning a reconstructive and discriminative dictionary for hyperspectral image classification. Precisely, w design a unified semi-supervised objective loss function which integrates a reconstruction term with a reconstruction–discrimination term built by *soft-max loss* to leverage the unsupervised and supervised information from the training samples. In the iteratively semi-supervised learning phase, we simultaneously update the dictionary and classifier by feeding the obtained training pairs into the unified objective function via a SGD algorithm. The experimental results obtained by using three real hyperspectral images indicate that the proposed algorithm leads to better classification performance compared to the other related methods. We should note that although dictionary learning serves as an important part in DSR, previous studies involving the DSR problem mainly focused on the classification performance. Our experiments also mainly focus on classification since it is the ultimate goal of this work. On the basis of the comprehensive experiments, we draw the following conclusions:

(i) The proposed method is insensitive to λ and λ_1, but it is sensitive to μ.
(ii) The proposed method outperforms other related algorithms in terms of classification accuracy, which demonstrates the superiority of *soft-max loss*.
(iii) Although the proposed method exhibits slightly higher computational complexity compared with MOD, OnlineDL, and SDL, its computational time is bearable.

Further experiments with additional scenes and comparison methods should be conducted in the future. Furthermore, we also envisage two future perspectives for the development of the presented work:

(i) Given the fact that the dictionary and classifier are updated by randomly selected unlabeled samples, our future work will consider exploiting active learning [43] to select the most informative samples during the learning phase in the DSR model.
(ii) Since the computational complexity of our algorithm is a bit high, in our future work we will exploit the objective function to speed up the optimization process by incorporating incremental learning [56].
(iii) Inspired by the experimental results, our future work will exploit the theoretical reason for when and why more labeled samples help in the semi-supervised joint dictionary learning tasks, which is a crucial and interesting problem. The possible reason may be training data distribution, variance across the training and test data domain, feature dimensions, sample size, number of labels samples per class, [57].
(iv) Since the spatial property is important in the proposed method, our future work will also focus on exploiting the dictionary structure in the DSR problem [37], and we will try to reveal the relation between dictionary structure and classification accuracy in DSR.

Acknowledgments: This work was mainly supported by the National Natural Science Foundation of China (NSFC) (41601347), the Natural Science Foundation of Jiangsu Province (BK20160860), the Fundamental Research Funds for the Central Universities (2015B29214), and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), which also was partially supported by the NSFC founds (41571325, 41271420). The authors would like to thank D. Landgrebe for making the Airborne Visible/Infrared Imaging Spectrometer Indian Pines hyperspectral data set available to the community and P. Gamba for providing the Reflective Optics Spectrographic Imaging System data over Pavia, Italy, along with the training and test data sets. The authors would also like to thank the Associate Editor who handled this paper and the anonymous reviewers for providing truly outstanding comments and suggestions that significantly helped improve the technical quality and presentation of this paper.

Author Contributions: Zhaohui Xue conceived and designed the methodology, Peijun Du performed the experiments, Hongjun Su analyzed the results, and Shaoguang Zhou made the conclusion, and all authors jointly wrote the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Let $\mathcal{L}_s[\mathbf{y}_t, f(\mathbf{z}^*(\mathbf{x}_t^l, \mathbf{D}_t), \mathbf{W}_t)]$ denote the unified objective loss function in Equation (16). At this point, we have to compute the gradient descents of $\mathcal{L}_s$ for $\mathbf{x}_t^l$ with respect to $\mathbf{D}_t$ and $\mathbf{W}_t$. Due to the fact that $\mathbf{D}_t$ is implicit in $\mathbf{z}^*(\mathbf{x}_t^l, \mathbf{D}_t)$ and there is no explicit analytical link between $\mathbf{z}_t^l$ and $\mathbf{D}_t$ [28], we have to compute the gradients of $\nabla_{\mathbf{D}_t}\mathcal{L}_s$ using a chain rule $\nabla_{\mathbf{x}_t^l}\mathcal{L}_s\nabla_{\mathbf{D}_t}\mathbf{x}_t^l$ (Hereinafter, we use $\mathcal{L}_s$ instead of $\mathcal{L}_s[\mathbf{y}_t, f(\mathbf{z}^*(\mathbf{x}_t^l, \mathbf{D}_t), \mathbf{W}_t)]$ for simplicity in the differential formulations), which leads to a main difficulty in solving the objective loss function in Equation (12). Following *Proposition* 1 by the work [28], we can obtain

$$\begin{aligned}
&\nabla_{\mathbf{D}_t}\mathcal{L}_s = \mu(-\mathbf{D}_t\mathbf{fl}\mathbf{z}_t^{l\mathrm{T}} + (\mathbf{x}_t^l - \mathbf{D}_t\mathbf{z}_t^l)\mathbf{fl}^{\mathrm{T}}),\\
&\mathbf{fl}_\Lambda = (\mathbf{D}_{t,\Lambda}^{\mathrm{T}}\mathbf{D}_{t,\Lambda} + \lambda_2\mathbf{I}_{K,\Lambda})^{-1}\nabla_{z_{t,\Lambda}^l}\mathcal{L}_s^{\mathbf{W}},\\
&\mathbf{fl}_{\bar{\Lambda}} = \mathbf{0},\\
&\nabla_{\mathbf{z}_{t,\Lambda}^l}\mathcal{L}_s^{\mathbf{W}} = \sum_{j=1}^{m}\left[\frac{\sum_{p=1}^{m}\mathbf{w}_{p,\Lambda}\exp(\mathbf{w}_{p,\Lambda}^{\mathrm{T}}\mathbf{z}_{t,\Lambda}^l)}{\sum_{p=1}^{m}\exp(\mathbf{w}_{p,\Lambda}^{\mathrm{T}}\mathbf{z}_{t,\Lambda}^l)} - \mathbf{1}\{\mathcal{Y}(\mathbf{x}_t^l) = j\}\mathbf{w}_{j,\Lambda}\right] + \lambda\mathbf{I}_{K,\Lambda},
\end{aligned} \tag{A1}$$

where γ is an auxiliary vector, Λ denotes the support set in sparse code $\mathbf{z}_t$ (The support set contains indices of nonzero coefficients in a sparse vector), $\bar{\Lambda}$ refers to the zero indices, and $\mathcal{L}_s^{\mathbf{W}}$ denotes $\Gamma_s[\mathbf{y}_t, f(\mathbf{z}^*(\mathbf{x}_t, \mathbf{D}_t), \mathbf{W}_t)] + \frac{\lambda_1}{2}\|\mathbf{W}_t\|_F^2$. Note that we adopt the similar optimization scheme as stated by the work [28] with the designed objective loss function to solve our problem since both of them adopt the SGD algorithm for optimization, but we have derived $\nabla_{\mathbf{z}_{t,\Lambda}^l}\mathcal{L}_s^{\mathbf{W}}$ by ourselves due to the fact that we adopt a different loss function.

On the other hand, the gradients of $\mathcal{L}_s^{\mathbf{W}}$ with respect to $\mathbf{w}_j$ can be easily obtained by

$$\nabla_{\mathbf{w}_j}\mathcal{L}_s^{\mathbf{W}} = \mathbf{z}_t^{l\mathrm{T}}\left(\frac{e^{\mathbf{w}_j^{\mathrm{T}}\mathbf{z}_t^l}}{\sum_{p=1}^{m}e^{\mathbf{w}_p^{\mathrm{T}}\mathbf{z}_t^l}} - \mathbf{1}\{\mathcal{Y}(\mathbf{x}_t^l) = j\}\right) + \lambda_1\mathbf{w}_j. \tag{A2}$$

Then, the gradient descents of $\mathcal{L}_s$ with respect to $\mathbf{W}_t$ can be written as

$$\nabla_{\mathbf{W}_t}\mathcal{L}_s = \begin{pmatrix}\nabla_{\mathbf{w}_1}\mathcal{L}_s^{\mathbf{W}}\\ \nabla_{\mathbf{w}_2}\mathcal{L}_s^{\mathbf{W}}\\ \cdots\\ \nabla_{\mathbf{w}_m}\mathcal{L}_s^{\mathbf{W}}\end{pmatrix}. \tag{A3}$$

Since both the gradients from unsupervised and supervised dictionary learning phases are obtained, we now can obtain the final update of the dictionary by the following expression

$$\mathbf{D}_{t+1} \leftarrow \mathbf{D}_t - \rho_t[(1-\mu)(\mathbf{D}_t\mathbf{z}_t^u - \mathbf{x}_t^u)\mathbf{z}_t^{u\mathrm{T}} + \mu(-\mathbf{D}_t\gamma\mathbf{z}_t^{l\mathrm{T}} + (\mathbf{x}_t^l - \mathbf{D}_t\mathbf{z}_t^l)\gamma^{\mathrm{T}})], \tag{A4}$$

where ρ_t refers to the learning rate.

Similar to the procedure adopted for updating the dictionary $\mathbf{D}$, the classification parameter $\mathbf{W}_t$ can be updated by

$$\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \rho_t\nabla_{\mathbf{W}_t}\mathcal{L}_s. \tag{A5}$$

So far, we have given the details in optimizing Equation (12). In addition, the learning rate for updating the dictionary is usually chosen according to a heuristic rule. Here, we follow the studies [28,29] by setting the learning rate to $\min(\rho, \rho t_0/t)$, where ρ is a constant that ensures the

convergence of SGD with $t_0 = T/10$ and T is the total number of iterations. Before applying another iteration, we remove $\mathbf{x}^u$ from the candidate pool $\mathbf{X}^u$. The process of sampling, sparse coding, and updating is repeated as we loop through all the labeled samples $\mathbf{x}^l$ in $\mathbf{X}^l$.

References

1. Toth, C.; Jozkow, G. Remote sensing platforms and sensors: A survey. *ISPRS J. Photogramm. Remote Sens.* **2016**, *115*, 22–36.
2. Plaza, A.; Benediktsson, J.A.; Boardman, J.W.; Brazile, J.; Bruzzone, L.; Camps-Valls, G.; Chanussot, J.; Fauvel, M.; Gamba, P.; Gualtieri, A.; et al. Recent advances in techniques for hyperspectral image processing. *Remote Sens. Environ.* **2009**, *113*, S110–S122.
3. Bioucas-Dias, J.; Plaza, A.; Camps-Valls, G.; Paul, S.; Nasrabadi, N.M.; Chanussot, J. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–36.
4. Camps-Valls, G.; Tuia, D.; Bruzzone, L.; Benediktsson, J.A. Advances in Hyperspectral Image Classification. *IEEE Signal Process. Mag.* **2014**, *31*, 45–54.
5. Ghamisi, P.; Plaza, J.; Chen, Y.; Li, J.; Plaza, A.J. Advanced Spectral Classifiers for Hyperspectral Images: A review. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–32.
6. Wright, J.; Ma, Y.; Mairal, J.; Sapiro, G.; Huang, T.S.; Yan, S.C. Sparse Representation for Computer Vision and Pattern Recognition. *Proc. IEEE* **2010**, *98*, 1031–1044.
7. Li, W.; Du, Q. A survey on representation-based classification and detection in hyperspectral remote sensing imagery. *Pattern Recognit. Lett.* **2016**, *83*, 115–123.
8. Nasrabadi, N.M. Hyperspectral Target Detection. *IEEE Signal Process. Mag.* **2014**, *31*, 34–44.
9. Ma, W.K.; Bioucas-Dias, J.M.; Chan, T.H.; Gillis, N.; Gader, P.; Plaza, A.J.; Ambikapathi, A.; Chi, C.Y. A Signal Processing Perspective on Hyperspectral Unmixing. *IEEE Signal Process. Mag.* **2014**, *31*, 67–81.
10. Loncan, L.; Almeida, L.B.D.; Bioucas-Dias, J.M.; Briottet, X.; Chanussot, J.; Dobigeon, N.; Fabre, S.; Liao, W.; Licciardi, G.A.; Simoes, M.; et al. Hyperspectral Pansharpening: A Review. *IEEE Geosci. Remote Sens. Mag.* **2015**, *3*, 27–46.
11. Xue, Z.; Li, J.; Cheng, L.; Du, P. Spectral-Spatial Classification of Hyperspectral Data via Morphological Component Analysis-Based Image Separation. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 70–84.
12. Xu, X.; Li, J.; Huang, X.; Mura, M.D.; Plaza, A. Multiple Morphological Component Analysis Based Decomposition for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3083–3102.
13. Ly, N.H.; Du, Q.; Fowler, J.E. Sparse Graph-Based Discriminant Analysis for Hyperspectral Imagery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 3872–3884.
14. Xue, Z.; Du, P.; Li, J.; Su, H. Simultaneous Sparse Graph Embedding for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6114–6133.
15. Rubinstein, R.; Bruckstein, A.M.; Elad, M. Dictionaries for Sparse Representation Modeling. *Proc. IEEE* **2010**, *98*, 1045–1057.
16. Engan, K.; Aase, S.O.; Husoy, J.H. Frame based signal compression using method of optimal directions (MOD). In Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, Orlando, FL, USA, 30 May–2 June 1999; Volume 4, pp. 1–4.
17. Aharon, M.; Elad, M.; Bruckstein, A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **2006**, *54*, 4311–4322.
18. Yaghoobi, M.; Blumensath, T.; Davies, M. Dictionary learning for sparse approximation with majorization method. *IEEE Trans. Signal Process.* **2009**, *57*, 2178–2191.
19. Skretting, K.; Engan, K. Recursive Least Squares Dictionary Learning Algorithm. *IEEE Trans. Signal Process.* **2010**, *58*, 2121–2130.
20. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. Discriminative learned dictionaries for local image analysis. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; Volume 1-12, pp. 2415–2422.
21. Pham, D.S.; Venkatesh, S. Joint learning and dictionary construction for pattern recognition. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; Volume 1-12, pp. 517–524.

22. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. Supervised dictionary learning. *arXiv* **2009**, arXiv:0809.3083.
23. Lian, X.C.; Li, Z.W.; Lu, B.L.; Zhang, L. Max-Margin Dictionary Learning for Multiclass Image Categorization. In Proceedings of the 2010 European Conference on Computer Vision ECCV, Pt IV, Hersonissos, Greece, 5–11 September 2010; Volume 6314, pp. 157–170.
24. Lian, X.C.; Li, Z.W.; Wang, C.H.; Lu, B.L.; Zhan, L. Probabilistic Models for Supervised Dictionary Learning. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 2305–2312.
25. Zhang, Q.A.; Li, B.X. Discriminative K-SVD for Dictionary Learning in Face Recognition. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 2691–2698.
26. Yang, M.; Zhang, L.; Feng, X.C.; Zhang, D. Fisher Discrimination Dictionary Learning for Sparse Representation. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 543–550.
27. Jiang, Z.L.; Zhang, G.X.; Davis, L.S. Submodular Dictionary Learning for Sparse Coding. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012, pp. 3418–3425.
28. Mairal, J.; Bach, F.; Ponce, J. Task-Driven Dictionary Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 791–804.
29. Jiang, Z.L.; Lin, Z.; Davis, L.S. Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2651–2664.
30. Zhang, G.X.; Jiang, Z.L.; Davis, L.S. Online Semi-Supervised Discriminative Dictionary Learning for Sparse Representation. In Proceedings of the 11th Asian Conference on Computer Vision (ACCV), Daejeon, Korea, 5–9 November 2012; Volume 7724, pp. 259–273.
31. Du, P.; Xue, Z.; Li, J.; Plaza, A. Learning Discriminative Sparse Representations for Hyperspectral Image Classification. *IEEE J. Sel. Top. Signal Process.* **2015**, *9*, 1089–1104.
32. Henao, R.; Yuan, X.; Carin, L. Bayesian nonlinear support vector machines and discriminative factor modeling. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 1754–1762.
33. Marial, J.; Bach, F.; Ponce, J.; Sapiro, G. Online dictionary learning for sparse coding. In Proceeding of the International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009.
34. Jenatton, R.; Audibert, J.Y.; Bach, F. Structured Variable Selection with Sparsity-Inducing Norms. *J. Mach. Learn. Res.* **2011**, *12*, 2777–2824.
35. Jenatton, R.; Mairal, J.; ObozinskiF, G.; Bach, F. Proximal methods for sparse hi erarchical dictionary learning. In Proceedings of the International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 487–494.
36. Huang, J.; Zhang, T.; Metaxas, D. Learning with structured sparsity. In Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, QC, Canada, 14–18 June 2009.
37. Mairal, J.; Jenatton, R. Obozinski, G. Learning Hierarchical and Topographic Dictionaries with Structured Sparsity. In Proceedings of the SPIE Wavelets and Sparsity XIV 81381P, San Diego, CA, USA, 21 August 2011.
38. Lian, W.; Rai, P.; Salazar, E.; Carin, L. Integrating Features and Similarities: Flexible Models for Heterogeneous Multiview Data. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 2757–2763.
39. Charles, A.S.; Olshausen, B.A.; Rozell, C.J. Learning Sparse Codes for Hyperspectral Imagery. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 963–978.
40. Castrodad, A.; Xing, Z.M.; Greer, J.B.; Bosch, E.; Carin, L.; Sapiro, G. Learning Discriminative Sparse Representations for Modeling, Source Separation, and Mapping of Hyperspectral Imagery. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4263–4281.
41. Wang, Z.W.; Nasrabadi, N.M.; Huang, T.S. Spatial-Spectral Classification of Hyperspectral Images Using Discriminative Dictionary Designed by Learning Vector Quantization. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4808–4822.
42. Wang, Z.Y.; Nasrabadi, N.M.; Huang, T.S. Semisupervised Hyperspectral Classification Using Task-Driven Dictionary Learning With Laplacian Regularization. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1161–1173.

43. Persello, C.; Bruzzone, L. Active and Semisupervised Learning for the Classification of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6937–6956.
44. Kushner, H.J.; Yin, G. *Stochastic Approximation and Recursive Algorithms and Applications*; Springer: New York, NY, USA, 2003.
45. Xue, Z.; Du, P.; Li, J.; Su, H. Sparse graph regularization for robust crop mapping using hyperspectral remotely sensed imagery with very few in situ data. *ISPRS J. Photogramm. Remote Sens.* **2017**, *124*, 1–15.
46. Xue, Z.; Du, P.; Li, J.; Su, H. Sparse Graph Regularization for Hyperspectral Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2351–2366.
47. Golub, G.H.; Hansen, P.C.; O'Leary, D.P. Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.* **1999**, *21*, 185–194.
48. Bioucas-Dias, J.M.; Figueiredo, M.A.T. Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing. In Proceedings of the 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Reykjavik, Iceland, 14–16 June 2010; pp. 1–4.
49. Boykov, Y.; Veksler, O.; Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 1222–1239.
50. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Hyperspectral Image Segmentation Using a New Bayesian Approach With Active Learning. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3947–3960.
51. Jensen, J.R. *Introductory Digital Image Processing: A Remote Sensing Perspective*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2005.
52. Congalton, R.G.; Green, K. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*; CRC Press: Boca Raton, FL, USA, 2008.
53. Foody, G.M. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 627–633.
54. Pati, Y.C.; Rezaiifar, R.; Krishnaprasad, P.S. Orthogonal Matching Pursuit-Recursive Function Approximation with Applications to Wavelet Decomposition. In Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1–3 November 1993; pp. 40–44.
55. Hyperspectral Remote Sensing Scenes. Available online: http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 25 November 2016)
56. Roux, N.L.; Schmidt, M.; Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. *arXiv* **2012**, arXiv:1202.6258.
57. Chapelle, O.; Schölkopf, B.; Zien, A. *Semi-Supervised Learning*; MIT Press: Cambridge, MA, USA, 2006.