Multiple Feature Learning for Hyperspectral Image Classification

Jun Li, *Member, IEEE*, Xin Huang, *Member, IEEE*, Paolo Gamba, *Fellow, IEEE*, José M. Bioucas Dias, *Senior Member, IEEE*, Liangpei Zhang, *Senior Member, IEEE*, Jon Atli Benediktsson, *Fellow, IEEE*, and Antonio Plaza, *IEEE Senior Member*

Abstract

Hyperspectral image classification has been an active topic of research in recent years. In the past, many different types of features have been extracted (using both linear and nonlinear strategies) for classification problems. On the one hand, some approaches have exploited the original spectral information or other features linearly derived from such information in order to have classes which are linearly separable. On the other hand, other techniques have exploited features obtained through nonlinear transformations intended to reduce data dimensionality, to better model the inherent nonlinearity of the original data (e.g., kernels), or to adequately exploit the spatial information contained in the scene (e.g., using morphological analysis). Special attention has been given to techniques able to exploit a single kind of features, such composite kernel learning or multiple kernel learning, developed in order to deal with multiple kernels. However, few approaches have been designed to integrate multiple types of features extracted from both linear and nonlinear transformations. In this paper, we develop a new framework for classification of hyperspectral scenes that pursues the combination of multiple features. The ultimate goal of the proposed framework is to be able to cope with linear and nonlinear class boundaries present in the data, thus following the two main mixing models considered for hyperspectral data interpretation. An important characteristic of the presented approach is that it does not require any regularization parameters to control the weights of considered features, so that different types of features can be efficiently exploited and integrated in a collaborative and flexible way. Our experimental results, conducted using a variety of input features and hyperspectral scenes, indicate that the proposed framework for multiple feature learning provides state-of-the-art classification results without significantly increasing computational complexity.

Index Terms

Hyperspectral imaging, multiple feature learning, linear and nonlinear features.

J. Li is with the School of Geography and Planning, Sun Yat-Sen University, Guangzhou, P. R. China. X. Huang and L. Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, P.R. China. P. Gamba is with the Telecommunications and Remote Sensing Laboratory, University of Pavia, Italy. J. M. Bioucas-Dias is with Instituto Superior Técnico, Portugal. J. A. Benediktsson is with the Faculty of Electrical and Computer Engineering, University of Iceland, Reykjavik, Iceland. A. Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, Cáceres, Spain.

I. Introduction

The recent availability of remotely sensed hyperspectral images has fostered the development of techniques able to interpret such high-dimensional data in many different application contexts [1]. It is now commonly accepted that using the spatial and the spectral information simultaneously provides significant advantages in terms of improving the performance of classification techniques. A detailed overview of recent advances in spatial-spectral classification of hyperspectral data is available in [2]. Resulting from the need to model both the spectral and the spatial information contained in the original data, different types of features have been exploited for spectral-spatial classification. These features can mainly classified into two categories:

- On the one hand, several methods exploit the original spectral information or other features linearly derived from such information. These kind of features have been widely used to exploit the linear separability of certain classes [3]. Techniques commonly used for this purpose include the maximum noise fraction (MNF) [4], independent component analysis (ICA) [5], linear spectral unmixing [6], or projection pursuit (PP) [7], among many others [8].
- On the other hand, in real analysis scenarios it is likely to find cases in which nonlinear features are more effective for class discrimination due to the existence of nonlinear class boundaries. As a result, several techniques have focused on on exploiting features obtained through nonlinear transformations to better model the inherent nonlinearity of the original data. Examples include kernel methods [9], [10] and manifold regularization [11], [12]. Other nonlinear approaches are focused on adequately exploiting the spatial information contained in the scene, e.g., using morphological analysis [13], [14].

Once relevant features have been extracted from the original data, the classification process itself can also be either linear or nonlinear. For instance, in linear discriminant analysis (LDA) [15] a linear function is used in order to maximize the discriminatory power and separate the available classes effectively. However, such a linear function may not be the best choice and nonlinear strategies such as quadratic discriminant analysis (QDA) or logarithmic discriminant analysis (LogDA) have also been used. The main problem of these supervised classifiers, however, is their sensitivity to the Hughes effect [16].

In turn, kernel methods [9] have been widely used in order to deal effectively with the Hughes phenomenon [17], [18]. The idea is to use a kernel trick that allows separation of the classes in a higher dimensional space by means of a nonlinear transformation, particularly in those cases in which the problem is not linearly separable in the original feature space. The combination of kernel methods and nonlinearly derived features (such as morphological features) has also been widely explored in the context of hyperspectral image classification [19].

Recently, a new trend has been oriented towards the composition of different kernels for improved learning, inspired by multiple kernel learning (MKL) approaches [20]–[23]. Some of these aspects were particularly discussed in [24], in which a detailed overview of machine learning in remote sensing data processing is given. For instance, a simple strategy to incorporate the spatial context into kernel-based classifiers is to define a pixel entity both in the spectral domain (using its spectral content) and also in the spatial domain, *e.g.* by applying some feature extraction

to its surrounding area which yields spatial (contextual) features, such as those derived using morphological analysis. These separated entities lead to two different kernel matrices, which can be easily computed. At this point, one can sum spectral and textural dedicated kernel matrices and introduce the cross-information between textural and spectral features in the formulation. This methodology yields a full family of composite kernel-based methods for hyperspectral data classification [25].

More recently, composite kernels have been generalized in [26] using the multinomial logistic regression (MLR) classifier [27] and extended multi-attribute profiles (EMAPs) [28]. The MLR has been recently explored in hyperspectral imaging as a technique able to model the posterior class distributions in a Bayesian framework, thus supplying (in addition to the boundaries between the classes) a degree of plausibility for such classes [29]. The resulting generalized composite kernel-based MLR can combine multiple kernels without any restriction of convexity. This introduces a different approach with regards to traditional composite kernel and MKL methods, in which composite kernels need to be convex combinations of kernels.

At this point, it is important to emphasize that both composite kernel learning and MKL focus on kernels, which are obtained either from the original (linear) spectral features or from (nonlinear) features such as MPs. These approaches exploit the information contained in the kernels using linear combinations, due to the fact that the optimization problem is much easier to solve under a linear framework. With these assumptions in mind, very good performance has been reported for MKL or other composite kernel learning approaches in different remote sensing problems [21], [22], [26]. However, these approaches focus on kernels, while kernel transformations of nonlinear features might bring redundance or lose the physical meaning of the features themselves. Instead, in certain situations it may be desirable to exploit the information carried out by each feature under its specifical physical or acquisition conditions. Inspired by these ideas, and based on the fact that it is common to have both linear and nonlinear class boundaries in the same scene, this paper develops a new framework for classification of hyperspectral images which integrates multiple features extracted from linear and nonlinear transformations.

A main characteristic of the presented approach is that it can adaptively exploit information from both linear and nonlinearly derived features, thus being able to address practical scenarios in which different classes may need different (linear or nonlinear) strategies. It should be noted that, as it is the case of MKL, the proposed approach also follows a linear optimization framework due to model complexity. However, the proposed approach has been designed in a way that it exhibits great flexibility to combine different types of features without any regularization parameters to control the weight of each feature, thus taking advantage of the complementarity that the features can provide without any *a priori* restrictions. In turn, MKL (which can be seen as a special instance of our proposed framework) generally needs to learn the weight parameters which is difficult from the viewpoint of both optimization and computational cost. Our presented approach is thus aimed at exploiting the different properties that both linear and nonlinear features can provide, with the ultimate goal of being able to characterize both linear and nonlinear boundaries independently of which type of features dominate the scene. In order to achieve the desired spectral-spatial integration that is normally expected in advanced classification problems, we consider morphological features as an important part of our framework, which also exploits kernel-based features and the original spectral

information contained in the hyperspectral scene.

The remainder of the paper is organized as follows. Section II presents the proposed classification framework, which uses the sparse MLR (SMLR) [30] as the baseline classifier. It will be shown that this classifier provides a natural framework to achieve the desired integration of multiple features. Section III reports the classification results obtained by the proposed multiple feature learning approach using different real hyperspectral data sets, which comprise a scene collected by the airborne visible infra-red imaging spectrometer (AVIRIS) over the Indian Pines region in Indiana, two scenes collected by the reflective optics spectrographic imaging system (ROSIS) over the city of Pavia, Italy, and a scene collected by the hyperspectral digital imagery collection experiment (HYDICE) over the city of Washington DC. These data sets have been widely used for evaluating the performance of hyperspectral image classification algorithms, and the results reported in this work rank among the most accurate ones ever reported for these scenes. Section IV concludes our study with some remarks and hints at plausible future research lines.

II. Proposed Framework for Multiple Feature Learning

First of all, we define the notations that will be adopted throughout the paper. Let $\mathcal{K}\equiv\{1,\ldots,K\}$ denote a set of K class labels; let $\mathcal{S}\equiv\{1,\ldots,n\}$ denote a set of integers indexing the n pixels of a hyperspectral image; let $\mathbf{x}\equiv(\mathbf{x}_1,\ldots,\mathbf{x}_n)\in\mathbb{R}^d$ denote such hyperspectral image, which is made up of d-dimensional feature vectors; let $\mathbf{y}\equiv(y_1,\ldots,y_n)$ denote an image of labels; and let $\mathcal{D}_L\equiv\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_L,y_L)\}$ be the labeled training set with L being the number of samples in $\mathcal{D}_L$. In this work, we model the posterior class probabilities using the MLR [27] as follows:

$$p(y_i=k|\mathbf{x}_i,\boldsymbol{\omega})\equiv\frac{\exp(\boldsymbol{\omega}^{(k)^T}\mathbf{h}(\mathbf{x}_i))}{\sum_{k=1}^{K}\exp(\boldsymbol{\omega}^{(k)^T}\mathbf{h}(\mathbf{x}_i))},\tag{1}$$

where $\mathbf{h}(\mathbf{x}_i)$ is the input feature, $\boldsymbol{\omega}$ denotes the regressors, and $\boldsymbol{\omega}\equiv[\boldsymbol{\omega}^{(1)^T},\cdots,\boldsymbol{\omega}^{(K-1)^T}]^T$. Since the density (1) does not depend on translations on the regressors $\boldsymbol{\omega}^{(k)}$, in this work we take $\boldsymbol{\omega}^{(K)}=\mathbf{0}$. It should be noted that the input feature $\mathbf{h}$ can be linear or nonlinear. In the former case, we have:

$$\mathbf{h}(\mathbf{x}_i)=[1,x_{i,1},\cdots,x_{i,d}]^T,\tag{2}$$

where $x_{i,j}$ denotes the j-th component of $\mathbf{x}_i$. On the other hand, the input feature $\mathbf{h}$ can also be nonlinear, in which case we have:

$$\mathbf{h}(\mathbf{x}_i)=[1,\psi_1(\mathbf{x}_i),\cdots,\psi_{l_1}(\mathbf{x}_i)]^T,\tag{3}$$

which is a feature vector with l_1 elements and which is built based on part of or the complete observation $\mathbf{x}$, with $\psi(\cdot)$ being a nonlinear function. Depending on the nonlinear function used, there are many possible ways to build nonlinear features. For instance, a kernel is some symmetric function with the form:

$$\mathbf{h}(\mathbf{x}_i)=[1,K(\mathbf{x}_i,\mathbf{x}_1),\cdots,K(\mathbf{x}_i,\mathbf{x}_l)]^T,\tag{4}$$

where:

$$K(\mathbf{x}_i,\mathbf{x}_j)=\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle,$$

and $\phi(\cdot)$ is a nonlinear mapping function. Kernels have been largely used in this context since they tend to improve data separability in the transformed space. However, other types of nonlinear functions for feature extraction may also be considered:

$$\mathbf{h}(\mathbf{x}_i) = [1, f_1(\mathbf{x}_i), \cdots, f_{l_2}(\mathbf{x}_i)]^T, \tag{5}$$

where $f(\cdot)$ is a nonlinear feature extraction transformation on the original data (for instance, the EMAP in [28]), and l_2 is the number of elements in $\mathbf{h}(\mathbf{x}_i)$. It should be noted that both the linear function $\mathbf{h}(\mathbf{x}_i) = \mathbf{x}_i$, and the kernel function $\mathbf{h}(\mathbf{x}_i) = K(\mathbf{x}_i, \mathbf{x})$ can be simply regarded as two instances of the nonlinear case.

As mentioned before, there have been some efforts in the literature to combine different types of features, such as MKL. Linear features have been generally less effective for hyperspectral image classification than nonlinear features. In turn, kernel-based features (obtained from linear or nonlinear transformations) have been more widely used. This trend has been exploited by MKL by focusing on kernel features, which are extracted from the original spectral data or the nonlinear transformed data. However, few efforts have attempted to exploit both linear and nonlinear features in simultaneous fashion, despite they can exhibit some complementary properties (e.g., some classes may be properly separated using linear boundaries, while other classes may require nonlinear boundaries for separability). In real analysis scenarios, it is likely to have both linear and nonlinear class boundaries in the same hyperspectral image. At the same time, kernel transformations of nonlinear features may lead to data redundance and loss of physical meaning for the features. It is therefore important for a methodology to be able to cope with such linear and nonlinear boundaries simultaneously and adaptively. In this regard, the proposed framework provides the possibility to interpret multiple boundaries together. Again, different features have different characteristics, and the joint exploitation of different kind of features could lead to improved data separability. Inspired by this idea, we develop a framework for the integration of multiple features, with the ultimate goal of exploiting the characteristics of each type of feature in the classification process. For this purpose, we first define:

$$\mathbf{h}(\mathbf{x}_i) = [1, \mathbf{h}_1(\mathbf{x}_i)^T, \mathbf{h}_2(\mathbf{x}_i)^T, \ldots, \mathbf{h}_l(\mathbf{x}_i)^T]^T, \tag{6}$$

a vector of l fixed functions of the input data $\mathbf{x}_i$, where $\mathbf{h}_j(\mathbf{x}_i) \equiv [h_{j,1}(\mathbf{x}_i), \cdots, h_{j,l_j}(\mathbf{x}_i)] \in \mathbb{R}^{l_j}$ (for $j = 1, \ldots, l$) is a feature obtained by a linear/nonlinear transformation, and l_j is the number of elements in $\mathbf{h}_j(\mathbf{x}_i)$. Notice that, if $\mathbf{h}_j(\mathbf{x}_i)$ is a kernel function, then (6) is a combination of multiple kernels (this is the particular case addressed by MKL). Instead, our proposed framework opens the structure to the exploitation of multiple features, not necessarily kernels. In this scenario, learning the class densities amounts to estimating the logistic regressors $\boldsymbol{\omega}$ given by the input features $\mathbf{h}(\mathbf{x})$. Following previous work [27], [29]–[31], we compute $\boldsymbol{\omega}$ by calculating the *maximum a posteriori* estimate:

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \quad \ell(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}), \tag{7}$$

where $\ell(\boldsymbol{\omega})$ is the log-likelihood function given by:

$$\begin{aligned} \ell(\boldsymbol{\omega}) &\equiv \log \prod_{i=1}^{L} p(y_i|\mathbf{x}_i, \boldsymbol{\omega}) \\ &\equiv \sum_{i=1}^{L} \left(\mathbf{h}^T(\mathbf{x}_i)\boldsymbol{\omega}^{(y_i)} - \log \sum_{k=1}^{K} \exp(\mathbf{h}^T(\mathbf{x}_i)\boldsymbol{\omega}^{(k)}) \right), \end{aligned} \tag{8}$$

and $\log p(\boldsymbol{\omega})$ is a prior over $\boldsymbol{\omega}$ which is independent from the observation $\mathbf{x}$. In order to control the machine complexity and, thus, its generalization capacity, we model $\boldsymbol{\omega}$ as a random vector with Laplacian density $p(\boldsymbol{\omega}) \propto \exp(-\lambda\|\boldsymbol{\omega}\|_1)$, where λ is the regularization parameter controlling the degree of sparsity [30], [31].

Let $\boldsymbol{\nu}_j = [\omega_{j,1}, \cdots, \omega_{j,l_j}]^T$ denote the regressors associated with feature $\mathbf{h}_j(\cdot)$. By introducing the input features in (6), problem (7) can be solved as follows:

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \sum_{i=1}^{L} \left(\mathbf{h}^T(\mathbf{x}_i)\boldsymbol{\omega}^{(y_i)} - \log \sum_{k=1}^{K} \exp(\mathbf{h}^T(\mathbf{x}_i)\boldsymbol{\omega}^{(k)}) \right) + \log p(\boldsymbol{\omega}) \tag{9}$$

$$= \arg\max_{\boldsymbol{\omega}} \sum_{i=1}^{L} \omega_1^{(y_i)} + \log p(\boldsymbol{\omega}) \tag{10}$$

$$+ \sum_{i=1}^{L} \sum_{j=1}^{l} \sum_{t=1}^{l_j} \left(h_{j,p}(\mathbf{x}_i)\omega_{j,p}^{(y_i)} - \log \sum_{k=1}^{K} \exp\left(\omega_1^{(k)} + h_{j,p}(\mathbf{x}_i)\omega_{j,p}^{(k)}\right) \right), \tag{11}$$

where the term in (10) is independent from the observation data it is also independent from the nonlinear functions used. At this point, several important observations can be made:

- First and foremost, if $\mathbf{h}(\mathbf{x}_i)$ is a combination of multiple kernels, then (9) stands for a typical MKL problem. However, as compared with the simple MKL [20] implemented on the SVM model, problem (9) require no convexity constraint for the combination of multiple kernels. From this observation, we can also see MKL as a specific instance of the proposed multiple learning framework.
- As shown in (11) we have a linear combination of multiple nonlinear features which is not restricted to kernels, and the logistic weights $\boldsymbol{\nu}_j$ are specific for each associated nonlinear feature $\mathbf{h}_j(\cdot)$ and independent from any other $\boldsymbol{\nu}_p$, for $p = 1, \cdots, l$ and $p \neq j$. This is quite important as, on the one hand, the linear combination provides great flexibility for the classifier to search for the most representative features, which could be linear or nonlinear, thus balancing the information provided by different features while reducing the computational complexity due to the possibility to use a conventional optimization approach.
- Furthermore, the linear combination in (11) provides sufficient flexibility to find the most representative feature $\mathbf{h}_j$, and also provides the potential to find the most representative elements in each feature. As a result, the final logistic weights could be derived from a combination of different features, which is a collaborative solution involving multiple (linear or nonlinear) features.
- It is finally important to point out that, by introducing the Laplacian prior $p(\boldsymbol{\omega})$ which can lead to sparse solutions, the proposed approach can deal with high-dimensional input features using limited training samples, thus addressing ill-posed problems.

To conclude this section, we emphasize that the optimization problem (9) can be solved by the SMLR in [30] and by the fast SMLR (FSMLR) in [32]. However, most hyperspectral data sets are beyond the reach of these algorithms, as their processing becomes unbearable when the dimensionality of the input features increases. This is even more critical in our framework, in which we use multiple features. In order to address this issue, we take advantage of the logistic regression via variable splitting and augmented Lagrangian (LORSAL) algorithm in [31], [33], with overall complexity $O(L \times (l_1 + \cdots + l_l) \times K)$. At this point, we recall that L is the number of training samples, K is the number of classes, and l_j is the number of elements in the j-th linear/nonlinear feature. LORSAL is able to deal with high-dimensional features and plays a central role in this work, as in previous contributions [29], [31]. A full demo with our algorithm implementation is given.

III. Experimental Results

In this section, we provide an experimental evaluation for the presented framework using four real hyperspectral datasets. In our experiments, we consider four different linear/nonlinear features as reported in Table I. Specifically, we use a linear feature $\mathbf{h}_{\text{linear}}$ (the original spectral information), a nonlinear feature $\mathbf{h}_{\text{EMAP}}$ (which uses the concept of EMAP in [28], [34]), and two kernel features constructed over the two previously mentioned sources of information (spectral and spatial, respectively) using the Gaussian radial basis function (RBF) kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$ which is widely used in hyperspectral image classification [18]. In this work, the spectral kernel $\mathbf{K}_{\text{linear}}$ is built by using the original spectral data and the spatial kernel $\mathbf{K}_{\text{EMAP}}$ is built by using the EMAP. At this point, we emphasize that the linear and nonlinear features that have been selected for experiments in this work can be considered highly representative of the spectral and spatial information contained in the scene. While $\mathbf{h}_{\text{linear}}$ is a linear feature related to the original spectral information, $\mathbf{h}_{\text{EMAP}}$ exploits the interpretation of the data in spatial terms, and $\mathbf{K}_{\text{linear}}$ and $\mathbf{K}_{\text{EMAP}}$ are nonlinear representations of the original data and EMAPs, respectively. For the considered problems, we only use four different features as these features are able to provide very good performance. However, we would like to emphasize again that any other kind of features can be included in our framework, according to the considered application. At this point, we reiterate that the proposed framework has been designed to cope with both linear and nonlinear boundaries in a general way, so that other additional features (linear and nonlinear) could be included in accordance with the specific application domain. We believe, however, that the selected features are sufficiently representative in order to demonstrate the advantages of our proposed framework.

We emphasize that, in all our experiments, the parameter values involved have been carefully optimized so that the best performance is reported for each considered method. For the EMAP-based feature extraction we have used a grid search approach to optimize parameter values, and for the LORSAL classification we have also carefully optimized the parameter λ. Nevertheless, as shown in [31], we may have a large amount of suboptimal options and the solution is insensitive to different suboptimal values. The reported figures of overall accuracy [%], average accuracy (AA) [%], κ statistic [%], and individual classification accuracies [%] are obtained by averaging the results obtained after conducting ten independent Monte Carlo runs with respect to the training set $\mathcal{D}_L$. At the same time, we include the standard deviation in order to assess the statistical significance of the results. Finally, in order to

TABLE I

TYPES OF FEATURES CONSIDERED IN THIS WORK

Feature	Description
$\mathbf{h}_{\text{linear}}$	Original spectral information: $\mathbf{h}_{\text{linear}}(\mathbf{x}_i) = \mathbf{x}_i$
$\mathbf{h}_{\text{EMAP}}$	Extended multi-attribute profiles (EMAPs) in [35]
$\mathbf{K}_{\text{linear}}$	Gaussian RBF kernel applied to the original spectral information
$\mathbf{K}_{\text{EMAP}}$	Gaussian RBF kernel applied to the EMAPs
$\mathbf{h}_{\text{all}}$	All nonlinear features considered: [$\mathbf{h}_{\text{linear}}$, $\mathbf{h}_{\text{EMAP}}$, $\mathbf{K}_{\text{linear}}$, $\mathbf{K}_{\text{EMAP}}$]

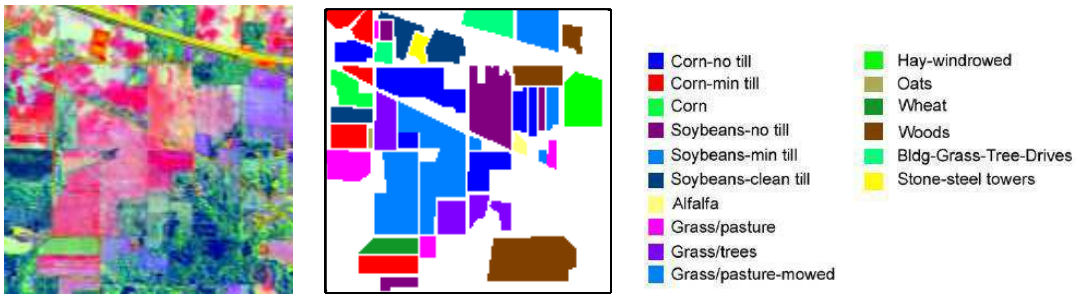
show the efficiency of the proposed framework, the computational time in seconds for learning the features is also reported in all cases (the time for deriving the features is not included for simplicity).

The remainder of the section is organized as follows. In subsection III-A we introduce the datasets used for evaluation. Subsection III-B describes the experiments with the AVIRIS Indian Pines data set. Subsection III-C conducts experiments using the ROSIS Pavia University dataset. Finally, subsection III-D presents the results obtained by the two remaining hyperspectral data sets.

A. Hyperspectral Data Sets

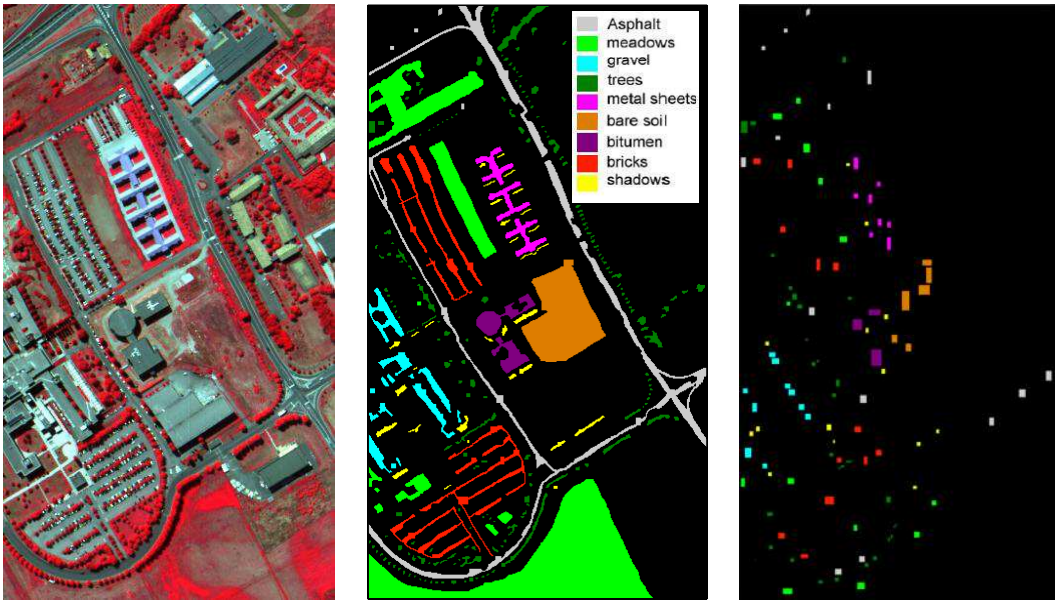
Four hyperspectral data sets collected by two different instruments are used in our experiments:

- The first hyperspectral image used in experiments was collected by the AVIRIS sensor over the Indian Pines region in Northwestern Indiana in 1992. This scene, with a size of 145 lines by 145 samples, was acquired over a mixed agricultural/forest area, early in the growing season. The scene comprises 202 spectral channels in the wavelength range from 0.4 to 2.5 μm, nominal spectral resolution of 10 nm, moderate spatial resolution of 20 meters by pixel, and 16-bit radiometric resolution. After an initial screening, several spectral bands were removed from the data set due to noise and water absorption phenomena, leaving a total of 164 radiance channels to be used in the experiments. For illustrative purposes, Fig. 1(a) shows a false color composition of the AVIRIS Indian Pines scene, while Fig. 1(b) shows the reference map available for the scene, displayed in the form of a class assignment for each labeled pixel, with 16 mutually exclusive reference classes, in total, 10366 samples. These data, including reference information, are available online from ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C.tif.zip, a fact which has made this scene a widely used benchmark for testing the accuracy of hyperspectral data classification algorithms. This scene constitutes a challenging classification problem due to the presence of mixed pixels in all available classes, and because of the unbalanced number of available labeled pixels per class.
- The second hyperspectral data set was collected by the ROSIS optical sensor over the urban area of the University of Pavia, Italy. The flight was operated by the Deutschen Zentrum for Luftund Raumfahrt (DLR, the German Aerospace Agency) in the framework of the HySens project, managed and sponsored by the



(a) (b)

Fig. 1. (a) False color composition of the AVIRIS Indian Pines scene. (b) Reference map containing 16 mutually exclusive land-cover classes (right).



(a) (b) (c)

Fig. 2. (a) False color composition of the ROSIS University of Pavia scene. (b) Reference map containing 9 mutually exclusive land-cover classes. (c) Training set used in experiments.

European Union. The image size in pixels is 610×340, with very high spatial resolution of 1.3 meters per pixel. The number of data channels in the acquired image is 103 (with spectral range from 0.43 to 0.86 μm). Fig. 2(a) shows a false color composite of the image, while Fig. 2(b) shows nine reference classes of interest, which comprise urban features, as well as soil and vegetation features. Out of the available reference pixels, 3921 were used for training [see Fig. 2(c)] and 42776 samples were used for testing.

- The third data set was also collected by the ROSIS optical sensor over a different location in the city centre of Pavia, Italy. The flight was also operated by DLR in the HySens framework. The number of data channels in the acquired image is 102 (with spectral range from 0.43 to 0.86 μm) and the spatial resolution is again 1.3 meters per pixel. These data were used in the 2008 IEEE Geoscience and Remote Sensing (GRSS) Data Fusion

Technical Committee contest. Additional details about the data and the training/test samples are available in [36].

- The fourth data set was also collected by HYDICE over the Mall area in Washington DC. The data set comprises 210 spectral bands from 0.4 to 2.4 μm. Bands in the 0.9 and 1.4 μm region where the atmosphere is opaque have been omitted from the data set, leaving 191 bands. The data set contains 1208$\times$307 pixels, with a spatial resolution of about 2.8 meters. Seven thematic land cover classes are present in the scene: roofs, street, path (graveled paths down the mall center), grass, trees, water, and shadow, with 19629 labeled samples in the ground truth image. The scene is available online from: http://cobweb.ecn.purdue.edu/$\sim$biehl/Hyperspectral_Project.zip.

B. Experiments with the AVIRIS Indian Pines Data Set

For this data set, the EMAPs were built using threshold values in the range 2.5% to 10% with respect to the mean of the individual features, with a step of 2.5% for the standard deviation attribute and thresholds of 200, 500 and 1000 for the area attribute.

1) Experiment 1: In our first set of experiments, we evaluated the classification accuracy of the proposed approach using a balanced training set per class in which around 5% of the labeled samples per class were used for training (a total of 515 samples) and the remaining labeled samples were used for testing. For very small classes we took a minimum of 3 training samples per class. Table II shows the overall, average, and individual classification accuracies (in percentage) and the κ statistic, along with the standard deviations, obtained after using the proposed framework with different types of features when applied to the AVIRIS Indian Pines scene.

From Table II, we can conclude that the proposed framework achieved the best results in terms of classification accuracies when all the considered features were used. This is expected, since in this case the proposed scheme seeks for the best solution among all the available (linear and nonlinear) features. On the other hand, the results obtained using the nonlinear feature $\mathbf{h}_{\text{EMAP}}$ are better than those obtained using the original spectral information. This is consistent with previous studies indicating that the EMAP provides a powerful tool for feature extraction, where the features extracted in the spatial domain can improve class separability [28], [35]. Another interesting observation is that the results obtained using only the nonlinear feature $\mathbf{h}_{\text{EMAP}}$ are better than those obtained from its kernel transformation $\mathbf{K}_{\text{EMAP}}$. This suggests that the kernel transformation of this particular nonlinear feature may not be able to improve the class separability.

2) Experiment 2: In our second experiment, we compare the proposed framework with composite kernel (CK) learning [25] and generalized composite kernel (GCK) learning [26]. Notice that all the experiments share exactly the same training and test sets. Table III shows that the proposed framework with $\mathbf{h}_{\text{all}}$ (i.e., using all the considered features) leads to the best classification results. However, the proposed framework exhibits the highest computational cost. Another important observation is that the results obtained by the EMAPs $\mathbf{h}_{\text{EMAP}}$ were better than those obtained by the kernel transformation $\mathbf{K}_{\text{EMAP}}$. As discussed, this is an indication that a kernel transformation of nonlinear features may not be able to improve the class separability.

TABLE II

OVERALL, AVERAGE AND INDIVIDUAL CLASSIFICATION ACCURACIES [%] OBTAINED BY THE PROPOSED FRAMEWORK (WITH DIFFERENT TYPES OF FEATURES) WHEN APPLIED TO THE AVIRIS INDIAN PINES HYPERSPECTRAL DATA SET WITH A BALANCED TRAINING SET IN WHICH 5% OF THE LABELED SAMPLES PER CLASS ARE USED FOR TRAINING (A TOTAL OF 515 SAMPLES) AND THE REMAINING LABELED SAMPLES ARE USED FOR TESTING.

Class	# Samples Training/Testing	Features				
		$\mathbf{h}_{\text{linear}}$	$\mathbf{h}_{\text{EMAP}}$	$\mathbf{K}_{\text{linear}}$	$\mathbf{K}_{\text{EMAP}}$	$\mathbf{h}_{\text{all}}$
Alfalfa	3/51	2.75±3.83	87.06±2.95	57.06±15.48	74.51±9.06	87.45±1.89
Corn-no till	71/1363	64.50±2.94	90.26±2.26	79.57±2.99	86.78±2.26	91.56±1.35
Corn-min till	41/793	35.17±6.32	91.44±2.85	62.81±2.53	88.81±2.27	92.35±2.07
Corn	11/223	13.14±3.98	90.18±3.39	48.43±11.08	67.00±11.79	90.67±2.83
Grass/pasture	24/473	75.12±3.66	93.15±3.95	89.37±2.08	91.16±2.92	94.61±2.39
Grass/tree	37/710	88.82±2.09	96.86±2.86	95.51±1.15	97.56±0.95	98.68±1.13
Grass/pasture-mowed	3/23	6.96±5.10	93.91±4.67	64.35±12.43	83.91±8.21	94.78±1.83
Hay-windrowed	24/465	95.94±1.55	98.92±1.42	98.69±0.51	98.92±0.29	99.66±0.11
Oats	3/17	7.06±7.23	99.41±1.86	78.82±16.68	83.53±17.71	97.06±5.00
Soybeans-no till	48/920	42.59±3.91	89.14±4.14	69.08±4.43	85.79±4.60	89.71±4.46
Soybeans-min till	123/2245	63.84±2.69	94.14±0.86	82.29±1.30	93.84±1.02	97.21±1.21
Soybeans-clean till	30/584	48.61±5.92	85.60±5.48	73.73±4.15	81.92±5.40	90.79±4.89
Wheat	10/202	89.51±4.13	99.01±0.57	99.31±0.26	99.51±0.40	99.60±0.31
Woods	64/1230	93.66±2.03	96.46±1.45	96.01±1.36	96.23±2.65	98.14±1.59
Bldg-grass-tree-drives	19/361	50.64±5.99	80.19±5.66	57.04±4.61	80.58±5.51	91.39±1.47
Stone-steel towers	4/91	56.37±9.19	61.76±7.36	53.52±11.26	78.46±6.86	74.29±7.29
Overall accuracy		64.60±1.01	92.25±0.33	80.59±0.60	90.42±0.63	94.59±0.58
Average accuracy		52.17±1.42	90.47±0.74	75.35±2.06	86.78±1.33	93.00±0.85
κ statistic		59.27±1.19	91.15±0.38	77.75±0.69	89.09±0.72	93.82±0.67
Time (seconds)		1.17	1.83	3.37	3.61	23.64

TABLE III

COMPARISON BETWEEN THE PROPOSED FRAMEWORK AND COMPOSITE KERNEL (CK) [25] AND GENERALIZED COMPOSITE KERNEL (GCK) [26] USING THE AVIRIS INDIAN PINES SCENE. THE PROCESSING TIME (IN SECONDS) IS ALSO REPORTED IN EACH CASE.

Accuracies	Proposed framework				GCK	SVM			
	$\mathbf{h}_{\text{EMAP}}$	$\mathbf{K}_{\text{linear}}$	$\mathbf{K}_{\text{EMAP}}$	$\mathbf{h}_{\text{all}}$	GCK$[\mathbf{K}_{\text{linear}}, \mathbf{K}_{\text{EMAP}}]$	$\mathbf{h}_{\text{EMAP}}$	$\mathbf{K}_{\text{linear}}$	$\mathbf{K}_{\text{EMAP}}$	CK$[\mathbf{K}_{\text{linear}}, \mathbf{K}_{\text{EMAP}}]$
Overall accuracy	92.25	80.59	90.42	94.59	93.87	91.46	76.95	90.52	90.85
Average accuracy	90.47	75.35	86.78	93.00	91.09	85.79	73.18	86.44	87.37
κ statistic	91.15	77.75	89.09	93.82	93.01	90.25	73.65	89.18	89.56
Time (seconds)	1.83	3.37	3.61	23.64	9.19	0.69	9.12	8.66	13.18

TABLE IV

NUMBER OF PIXELS DOMINATED BY EACH CONSIDERED TYPE OF FEATURE AND CLASSIFICATION ACCURACIES OBTAINED WHEN APPLYING THE PROPOSED FRAMEWORK TO THE AVIRIS INDIAN PINES DATA SET. IN THIS EXPERIMENT, WE USED APPROXIMATELY 30 TRAINING SAMPLES PER CLASS.

Class	# Samples	Number of pixels dominated by each feature				Total	Class Accuracy	Overall Accuracy
	Train/Test	$(\boldsymbol{\nu}^T\mathbf{h})_{\text{linear}}$	$(\boldsymbol{\nu}^T\mathbf{h})_{\text{EMAP}}$	$(\boldsymbol{\nu}^T\mathbf{K})_{\text{linear}}$	$(\boldsymbol{\nu}^T\mathbf{K})_{\text{EMAP}}$			
Alfalfa	15/39	0	38	0	0	38	96.15±3.25	92.28±1.11
Corn-no till	30/1404	205	582	629	0	1416	87.75±2.47	
Corn-min till	30/804	95	726	0	0	821	88.54±4.40	
Corn	30/204	142	119	0	0	261	97.06±1.63	
Grass/pasture	30/467	77	387	0	0	464	94.90±2.30	Average Accuracy
Grass/tree	30/717	21	686	1	0	708	97.88±0.72	
Grass/pasture-mowed	15/11	0	17	0	0	17	97.27±4.39	95.03±0.59
Hay-windrowed	30/459	0	467	0	0	467	99.83±0.17	
Oats	15/20	0	10	1	0	12	100	
Soybeans-no till	30/938	111	742	0	96	949	87.95±4.29	
Soybeans-min till	30/2438	1145	966	160	0	2271	89.78±3.66	κ statistic
Soybeans-clean till	30/584	42	520	29	58	649	93.61±2.62	
Wheat	30/182	2	182	0	0	184	99.62±0.27	91.19±1.25
Woods	30/1264	0	1227	0	0	1227	96.98±1.95	
Bldg-grass-tree-drives	30/350	36	247	42	41	365	95.31±1.46	
Stone-steel towers	30/65	-	-	-	-	91	97.69±1.95	

3) Experiment 3: In our third experiment, we analyze the relevance of linear and nonlinear features in the final classification results, with the ultimate goal of analyzing their capacity to characterize different complex classes in the scene. That is, in the set of all nonlinear features $\mathbf{h}_{\text{all}}$, we would like to analyze which feature has the most significant contribution. Here, we will use approximately 30 training samples per class, which is an unbalanced scenario in comparison with the one considered in the former experiment. Let $(\boldsymbol{\nu}_j)^T\mathbf{h}_j$ be the numerator of the MLR in (1). For $p = 1, \cdots, K$ and $p \neq j$, if $(\boldsymbol{\nu}_j)^T\mathbf{h}_j \geq (\boldsymbol{\nu}_p)^T\mathbf{h}_p$, then we conclude that the classification is dominated by $\mathbf{h}_j$. Table IV reports the total number of pixels in the scene which are dominated by each kind feature.

Several conclusions can be observed from Table IV. First and foremost, it is remarkable that for most classes the dominating feature according to Table IV is $\mathbf{h}_{\text{EMAP}}$. This is consistent with previous works revealing the power of EMAP for separating most classes which are nonlinearly separable in the spatial domain [28], [35]. Furthermore, it is remarkable that the original spectral information is highly relevant. This is due to the fact that some of the classes, *e.g.*, *Soybeans-min till*, are likely to be linearly separable. It is also observable that the kernel version of the spectral information provides important contributions, especially for the *Corn-no till*. This is because no-till is an

agricultural technique which increases the amount of water that infiltrates into the soil and increases organic matter retention and cycling of nutrients in the soil. Along with the complexity of corn itself, this may lead to nonlinearities that appear to be better explained by kernel-based features, as indicated in Table IV. However, the kernel version of EMAP rarely dominates the classification. This confirms our introspection that a kernel transformation of the nonlinear EMAP feature may not significantly improve the class separability, which is already fully exploited by the original EMAP itself.

In order to further illustrate the relative weights of the logistic regressors in the MLR classification, Fig. 3 shows the specific regressors calculated for classes *Corn-no till*, *Soybeans-min till*, and *Woods*, which are respectively dominated by $\mathbf{K}_{\text{linear}}$, $\mathbf{h}_{\text{linear}}$ and $\mathbf{h}_{\text{EMAP}}$. Fig. 3(d) shows the regressors calculated for class *Soybeans-clean till*, which has combined contributions from all features. From Fig. 3, it is clear that the original spectral information and the EMAP features are more relevant than the other tested features. A final observation resulting from this experiment is that, given the high computational complexity associated to using all the features $\mathbf{h}_{\text{all}}$, we can obtain a suitable subset of features including using only $\mathbf{h}_{\text{linear}}$ and $\mathbf{h}_{\text{EMAP}}$, *i.e.*, $\mathbf{h}_{\text{subset}} = [\mathbf{h}_{\text{linear}}, \mathbf{h}_{\text{EMAP}}]$, which leads to a comparable solution with very competitive computational cost. In this case, the kernel transformations are not relevant for improving classification accuracies and the combination of the original (spectral and EMAP-based) features can lead to very similar performance.

For illustrative purposes, Fig. 4 shows some of the obtained classification maps after applying the proposed framework to the AVIRIS Indian Pines scene using approximately 30 training samples per class. These maps correspond to one of the 10 Monte Carlo runs conducted for each considered type of feature. As we can observe in Fig. 4, the best classification accuracies are obtained using $\mathbf{h}_{\text{all}}$, but the accuracies obtained using $\mathbf{h}_{\text{EMAP}}$ and $\mathbf{K}_{\text{EMAP}}$ are also significant. Finally, the accuracies obtained using the original spectral information only ($\mathbf{h}_{\text{linear}}$) are low in comparison with the other approaches, while the introduction of the kernel version $\mathbf{K}_{\text{linear}}$ improves the obtained results but not to the levels achieved when EMAP features are also used for the proposed framework. In turn, EMAP-based features alone can lead to significant accuracies without the need for a kernel-based transformation.

C. Experiments with ROSIS University of Pavia Data Set

1) Experiment 1: In our first experiment with the ROSIS Pavia University scene, we evaluate the classification accuracies achieved by the proposed framework. In this experiment we consider, in addition to the features used in the previous experiment, a subset given by $\mathbf{h}_{\text{subset}} = [\mathbf{h}_{\text{linear}}, \mathbf{h}_{\text{EMAP}}]$ for comparison. Table V shows the overall, average, individual classification accuracies (in percentage) and the κ statistic obtained by the proposed framework using different types of input features. In all cases, we used the fixed training set in Fig. 2(c) to train the classifier. The EMAPs in this particular experiment were built using threshold values in the range 2.5% to 10% with respect to the mean of the individual features, and with a step of 2.5% for the definition of the criteria based on the standard deviation attribute. Values of 100, 200, 500 and 1000 were selected as references for the area attribute. The threshold values considered for the area attribute were chosen according to the resolution of the data and, thus, the size of the objects present in the scene.

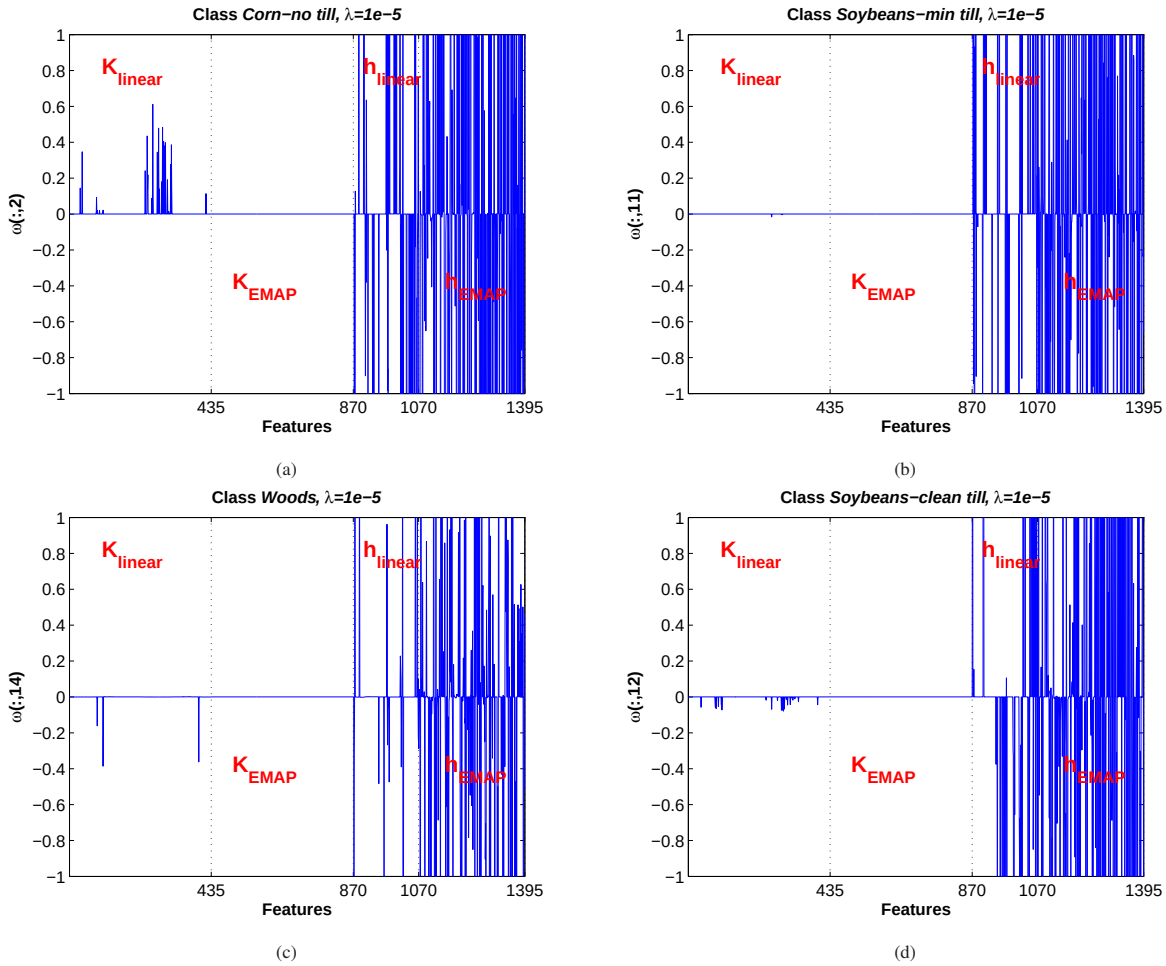


Fig. 3. Logistic regressors of the MLR classifier obtained from the AVIRIS Indian Pines dataset corresponding to the experiment reported in Table IV. (a) Class *Corn-no till* is dominated by the spectral kernel $\mathbf{K}_{\text{linear}}$. (b) Class *Soybeans-min till* is dominated by the original spectral information $\mathbf{h}_{\text{linear}}$. (c) Class *Woods* is dominated by the EMAP features $\mathbf{h}_{\text{EMAP}}$. (d) Class *Soybean-clean till* has contributions from all the considered features.

As shown by Table V, the classification accuracies obtained by the proposed framework are very high. Furthermore, as it was already the case in the previous experiment, the results using $\mathbf{h}_{\text{subset}}$ are comparable to those obtained using the full set of features, $\mathbf{h}_{\text{all}}$. However, in the case of $\mathbf{h}_{\text{subset}}$ the results can be obtained with much less computational complexity when compared to $\mathbf{h}_{\text{all}}$. This confirms our introspection that, even though our multiple learning framework can adequately exploit all available features, a selection of the most relevant features for classification (in this case, the original spectral information and the spatial characterization provided by EMAPs) can lead to similar results but with less computational complexity. In this experiment, as it was already the case in our experiment with the AVIRIS Indian Pines scene, kernel transformations cannot bring relevant additional information for classification.

2) Experiment 2: In our second experiment, we provide a comparison between the proposed framework with CK [25] and GCK [26]. Table VI shows the obtained results, in which all the experiments share exactly the same

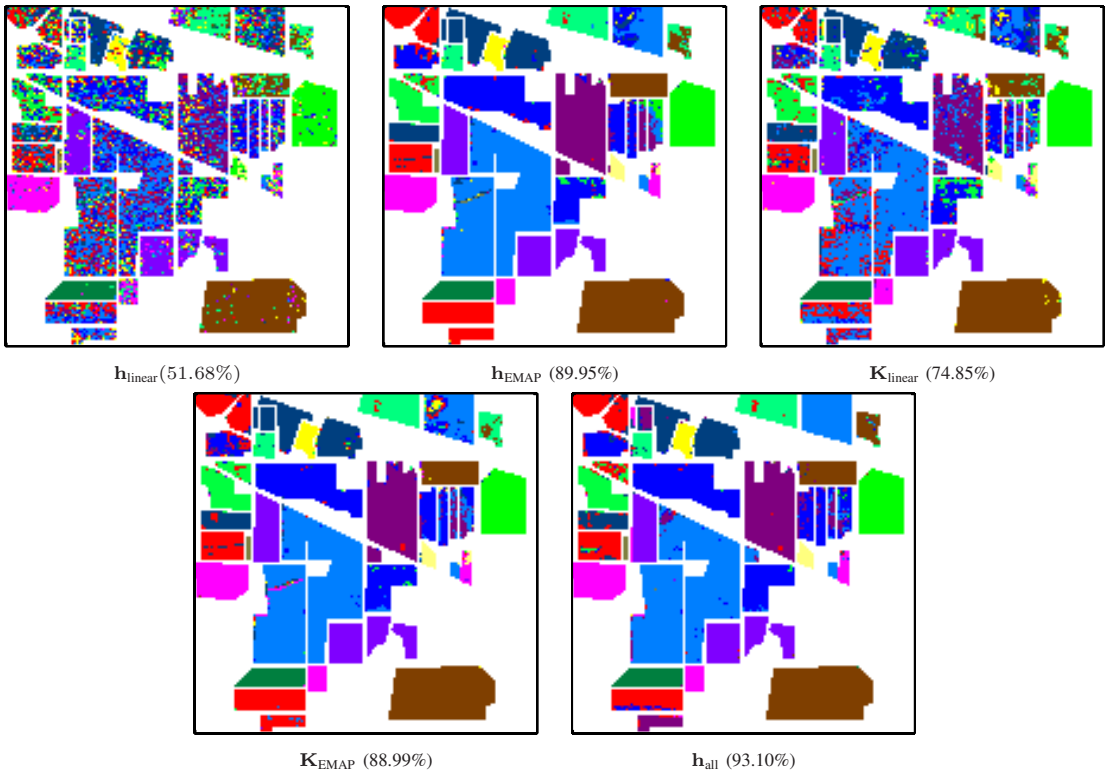


Fig. 4. Classification maps (along with the overall accuracies) obtained by the proposed framework for the AVIRIS Indian Pines dataset, using approximately 30 training samples per class.

training and test sets. Similar observations can be reported for the ROSIS Pavia University scene as the case alredy shown in the previous section with the AVIRIS Indian Pines data, i.e., the proposed framework with $\mathbf{h}_{\text{all}}$ (which learns all the available linear and nonlinear features) obtained very competitive results with minimum computational cost.

3) Experiment 3: Since the accuracy values obtained by $\mathbf{h}_{\text{EMAP}}$, $\mathbf{K}_{\text{EMAP}}$, $\mathbf{h}_{\text{all}}$ and $\mathbf{h}_{\text{subset}}$ are apparently similar, in our third experiment with the ROSIS Pavia University scene we analyze the statistical differences among all the considered features using the McNemar's test [37]. In this test, a value of $|Z| > 1.96$ indicates that there is a significant difference in accuracy between two classification methods. The sign of Z is also a criterion to indicate whether a first classifier is more accurate than a second one ($Z > 0$) or vice-versa ($Z < 0$). Table VII provides the results obtained for all the considered types of features with the ROSIS Pavia University data set. As it can be seen from Table VII, the performance of EMAP features ($\mathbf{h}_{\text{EMAP}}$) and their kernel transformation ($\mathbf{K}_{\text{EMAP}}$) is very similar in statistical sense. Therefore, instead of using $\mathbf{K}_{\text{EMAP}}$, we can simply resort to $\mathbf{h}_{\text{EMAP}}$, which provides similar accuracies with lower computational cost. Furthermore, it is noticeable that the performance of the original spectral information ($\mathbf{h}_{\text{linear}}$) is significantly different from that achieved by the nonlinear transformations. As a result, this experiment reveals that it is very important to combine both linear and nonlinear features for classification. This is successfully achieved by the presented method using all the features ($\mathbf{h}_{\text{all}}$) and a carefully selected subset ($\mathbf{h}_{\text{subset}}$),

TABLE V

OVERALL, AVERAGE AND INDIVIDUAL CLASS ACCURACIES [%] OBTAINED BY THE PROPOSED FRAMEWORK (WITH DIFFERENT TYPES OF FEATURES) WHEN APPLIED TO THE ROSIS PAVIA UNIVERSITY HYPERSPECTRAL DATA SET USING THE FIXED TRAINING SET IN FIG. 2(C). THE PROCESSING TIME (IN SECONDS) IS ALSO REPORTED IN EACH CASE.

Class	# Samples		Features					
	Train	Test	$\mathbf{h}_{\text{linear}}$	$\mathbf{h}_{\text{EMAP}}$	$\mathbf{K}_{\text{linear}}$	$\mathbf{K}_{\text{EMAP}}$	$\mathbf{h}_{\text{all}}$	$\mathbf{h}_{\text{subset}}$
Asphalt	548	6631	70.92	97.56	82.55	98.16	98.82	97.63
Bare soil	540	18649	53.23	99.12	67.44	98.76	98.43	98.91
Bitumen	392	2099	70.89	93.79	74.37	91.47	89.47	92.14
Bricks	524	3064	72.91	98.92	94.45	98.99	98.40	98.73
Gravel	265	1345	97.77	99.85	99.18	99.93	99.93	99.85
Meadows	532	5029	86.56	89.32	93.32	90.59	94.69	91.39
Metal sheets	375	1330	74.29	99.92	90.53	100.00	99.85	10.00
Shadows	514	3682	75.29	99.40	90.52	99.16	99.62	99.48
Trees	231	947	95.67	92.19	96.83	96.73	98.20	95.99
Overall accuracy			67.06	97.37	79.50	97.43	97.80	97.53
Average accuracy			77.50	96.67	87.72	97.09	97.49	97.12
κ			59.74	96.50	74.40	96.58	97.08	96.72
Time (seconds)			0.92	3.56	156.08	166.50	2082.3	5.00

TABLE VI

COMPARISON BETWEEN THE PROPOSED FRAMEWORK AND COMPOSITE KERNEL (CK) [25] AND GENERALIZED COMPOSITE KERNEL (GCK) [26] USING THE ROSIS PAVIA UNIVERSITY SCENE. THE PROCESSING TIME (IN SECONDS) IS ALSO REPORTED IN EACH CASE.

Accuracies	Proposed framework					GCK	SVM			
	$\mathbf{h}_{\text{EMAP}}$	$\mathbf{K}_{\text{linear}}$	$\mathbf{K}_{\text{EMAP}}$	$\mathbf{h}_{\text{all}}$	$\mathbf{h}_{\text{subset}}$	$[\mathbf{K}_{\text{linear}}, \mathbf{K}_{\text{EMAP}}]$	$\mathbf{h}_{\text{EMAP}}$	$\mathbf{K}_{\text{linear}}$	$\mathbf{K}_{\text{EMAP}}$	CK$[\mathbf{K}_{\text{linear}}, \mathbf{K}_{\text{EMAP}}]$
Overall accuracy	97.37	79.50	97.43	97.80	97.53	98.05	93.03	80.89	90.80	92.97
Average accuracy	96.67	87.72	97.09	97.49	97.12	97.73	94.04	89.09	94.08	94.92
κ statistic	96.50	74.40	96.58	97.08	96.72	97.42	90.91	76.12	88.13	90.86
Time (seconds)	3.56	156.08	166.50	2082.3	5.00	944.75	7.77	121.89	148.80	307.85

providing very competitive results in the considered analysis scenario.

For illustrative purposes, Fig. 5 shows some of the classification maps obtained after applying the proposed framework to the ROSIS Pavia University scene using the fixed training set depicted in Fig. 2(c). As we can observe in Fig. 5, a very good delineation of complex urban structures can be observed in the results obtained using any of the features including EMAPs, such as $\mathbf{h}_{\text{EMAP}}$. Quite opposite, the accuracies obtained using the original spectral information only ($\mathbf{h}_{\text{linear}}$) are low in comparison with the other approaches. In this particular case, as it was already observed in the experiments with the AVIRIS Indian Pines scene, the introduction of the kernel version

TABLE VII

STATISTICAL SIGNIFICANCE OF THE DIFFERENCES IN CLASSIFICATION ACCURACIES (MEASURED USING THE MCNEMAR'S TEST IN [37]) FOR THE PROPOSED FRAMEWORK, USING DIFFERENT TYPES OF FEATURES EXTRACTED FROM THE ROSIS PAVIA UNIVERSITY SCENE.

	Value of Z calculated by the McNemar's test					
	$\mathbf{h}_{\text{linear}}$	$\mathbf{h}_{\text{EMAP}}$	$\mathbf{K}_{\text{linear}}$	$\mathbf{K}_{\text{EMAP}}$	$\mathbf{h}_{\text{all}}$	$\mathbf{h}_{\text{subset}}$
$\mathbf{h}_{\text{linear}}$	-	-108.5776	-56.1976	-109.4933	-111.6610	-109.9032
$\mathbf{h}_{\text{EMAP}}$	108.5776	-	79.7141	-0.8906	-5.5907	-3.8908
$\mathbf{K}_{\text{linear}}$	56.1976	-79.7141	-	-80.8939	-83.7806	-81.2676
$\mathbf{K}_{\text{EMAP}}$	109.4933	0.8906	80.7806	-	-5.7926	-1.4969
$\mathbf{h}_{\text{all}}$	111.6610	5.5907	83.7806	5.7926	-	3.6986
$\mathbf{h}_{\text{subset}}$	109.9032	3.8908	81.2676	1.4969	-3.6986	-

$\mathbf{K}_{\text{linear}}$ improves the obtained results, but not to the levels observed when EMAP features are used in the proposed framework.

D. *Other Experiments*

In this section, we conduct an evaluation of the proposed approach using the ROSIS Pavia Centre and HYDICE Washington DC data sets. In the previously conducted experiments, we observed that the proposed framework with $\mathbf{h}_{\text{subset}}$ (which integrates both linear and nonlinear features) could obtain very good performance with minimum computational cost. Therefore, in this section we only evaluate the proposed framework by using $\mathbf{h}_{\text{subset}}$. Table VIII shows the obtained classification accuracies (as a function of the number of training samples) for these two data sets in this particular case. From Table VIII, it can be concluded that the proposed approach achieved very good performance, even with very limited training sets. Also, since the proposed approach does not require kernel transformations, it exhibits low computational cost. The low standard deviation values reported on Table VIII also indicate that the proposed framework is quite robust.

IV. CONCLUSIONS AND FUTURE RESEARCH LINES

In this paper, we have developed a new framework for multiple feature learning which is based on the integration of different types of (linear and nonlinear) features. A main contribution of the presented approach is the joint consideration of both linear and nonlinear features without any regularization parameters to control the weight of each feature, so that different types of available features can be jointly exploited (in a collaborative and flexible way) for hyperspectral image classification. Our main goal is to address a common situation in practice, in which some classes may be separated using linearly derived features while others may require nonlinearly derived features. Until now, a main trend when using multiple feature learning relies on the use of kernels, *i.e.*, multiple kernel learning (MKL). However, very few techniques have been explored in order to adaptively select the most useful type of feature for different classes in the scene. In this work, we give a first step in this direction and contribute a

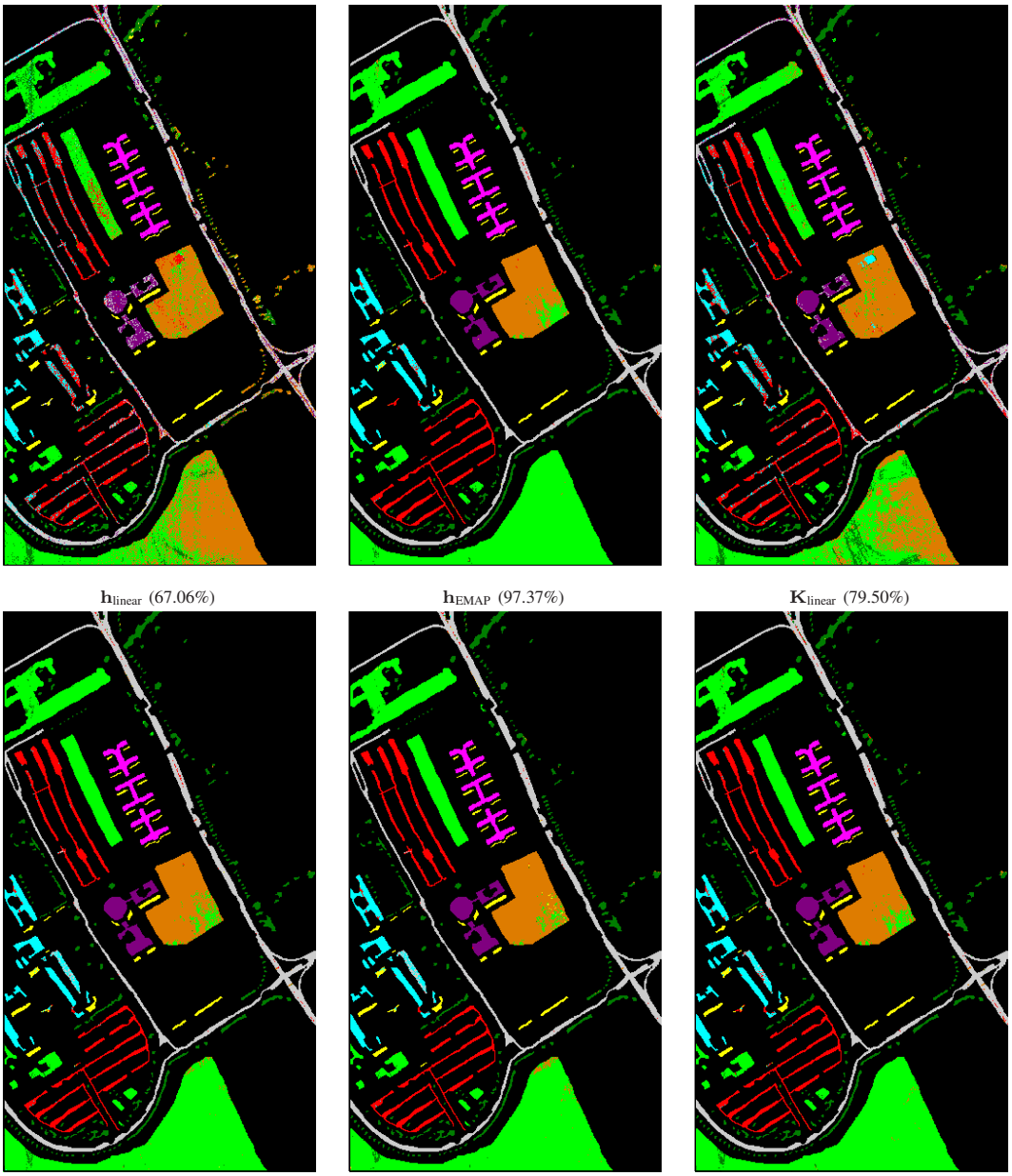


Fig. 5. Classification maps (along with the overall accuracies) obtained by the proposed framework for the ROSIS Pavia University dataset, using the fixed training set in Fig. 2(c).

framework which is flexible and able to deal with both linear and nonlinear class boundaries. A main innovation of our proposed approach is that it is more flexible than MKL, in the sense that it can consider linear and nonlinear features and not only kernel features. As a result, MKL can be considered as a special case of the proposed framework. Although the presented framework is general and suitable to incorporate any kind of input features, in this work we have considered a set of highly representative features such as the original (spectral) information

TABLE VIII

OVERALL (OA), AVERAGE (AA) ACCURACY AND κ STATISTIC –PLUS/MINUS THE STANDARD DEVIATION– AS A FUNCTION OF THE NUMBER OF LABELED SAMPLES PER CLASS (WITH THE TOTAL NUMBER OF LABELED SAMPLES IN THE PARENTHESES) OBTAINED BY THE PROPOSED METHOD FOR THE ROSIS PAVIA CENTRE AND HYDICE WASHINGTON DC DATA SETS.

ROSIS Pavia Centre data					
Accuracies	Number of labeled samples per class (total labeled samples)				
	10 (50)	20 (100)	30 (150)	40 (200)	50 (250)
Overall accuracy	92.25±2.17	93.05±1.83	94.26±0.95	94.70±0.76	95.39±0.77
Average accuracy	94.02±1.97	95.49±0.89	96.03±0.92	96.52±0.35	96.72±0.29
κ statistic	89.43±2.85	90.54±2.40	92.13±1.27	92.74±1.01	93.66±1.03
Time (seconds)	0.1473	0.1549	0.1585	0.1665	0.1765
HYDICE Washington DC data					
Accuracies	Number of labeled samples per class (total labeled samples)				
	10 (70)	20 (140)	30 (210)	40 (280)	50 (350)
Overall accuracy	91.01±2.93	92.29±1.73	95.54±1.33	96.01±0.60	97.57±0.47
Average accuracy	94.16±1.85	94.74± 1.32	96.84±0.64	97.03±0.33	98.00±0.26
κ statistic	89.19±3.43	90.66±2.06	94.56±1.59	95.13±0.73	97.02±0.57
Time (seconds)	0.2722	0.2930	0.3093	0.3556	0.4267

contained in the scene, a set of (spatial) morphological features extracted using different attributes, as well as kernel-based transformations of the aforementioned features. The framework therefore permits great flexibility in the exploitation of the advantages of each type of feature, as well as the incorporation of additional features in future developments.

Our experimental results, conducted with four widely used hyperspectral scenes, indicate that spatial-based features are very important for classification, while there is no significant difference between the original (spectral and spatial-based) features and their kernel-based transformations. However, the joint consideration of a pool of linear and nonlinear features allowed us to approach the classification problem in a way that is more general and flexible. In addition, our proposed strategy allowed us to reduce the computational complexity of the framework by selecting the most relevant features *a priori*, although the proposed framework can naturally select the most useful out of a large pool of input features for classification, without any requirement in terms of setting of regularization parameters or *a priori* information to control the weight of each feature. It should also be noted that the classification accuracies reported for the four considered hyperspectral scenes rank among the most accurate ones ever reported for these scenes. An important observation from our experiments is that, under the proposed multiple feature learning framework, kernel transformations may not be able to improve class separability (in particular, for nonlinear features). Since in this context kernel transformations increase computational complexity, our proposed framework allows excluding such kernel features and using the original features instead for specific applications.

As future work, we will conduct a more detailed investigation of other possible (linear and nonlinear) features that can be integrated in the proposed framework. Based on the observation that kernel-based features may not be as important as other features in our presented framework, the computational complexity can be further reduced by adaptively selecting the most relevant features for classification. We are also developing parallel versions of the proposed framework in a variety of architectures, such as commodity graphics processing units (GPUs) or multi-GPU platforms.

ACKNOWLEDGMENT

The authors would like to gratefully thank Prof. David Landgrebe for providing the hyperspectral data set used in our experiments. Last but not least, the authors gratefully thank the Associate Editor who handled the manuscript and the anonymous reviewers for their outstanding comments and suggestions, which greatly helped to improve the technical quality and presentation of the manuscript.

REFERENCES

[1] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, pp. 110–122, September 2009.

[2] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, 2013.

[3] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 17–28, 2002.

[4] A. Green, M. Berman, P. Switzer, and M. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 26, pp. 65–74, 1988.

[5] J. Bayliss, J. A. Gualtieri, and R. Cromp, "Analysing hyperspectral data with independent component analysis," in *Proceedings of SPIE*, vol. 3240, 1997, pp. 133–143.

[6] I. Dopido, M. Zortea, A. Villa, A. Plaza, and P. Gamba, "Unmixing prior to supervised classification of remotely sensed hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, pp. 760–764, 2011.

[7] L. O. Jimenez-Rodriguez, E. Arzuaga-Cruz, and M. Velez-Reyes, "Unsupervised linear feature-extraction methods and their effects in the classification of high-dimensional data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 2, pp. 469 –483, 2007.

[8] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis: An Introduction*. Springer, 2006.

[9] B. Scholkopf and A. Smola, *Learning With Kernels? Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press Series, 2002.

[10] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 1351–1362, 2005.

[11] B. Du, L. Zhang, L. Zhang, T. Chen, and K. Wu, "A discriminative manifold learning based dimension reduction method," *Int. J. Fuzzy Syst.*, vol. 14, no. 2, pp. 272–277, 2012.

[12] W. Kim and M. Crawford, "Adaptive classification for hyperspectral image data using manifold regularization kernel machines," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 11, pp. 4110–4121, 2010.

[13] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 2, pp. 309–320, 2001.

[14] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, pp. 480–491, 2005.

[15] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 3, pp. 862–873, 2009.

[16] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[17] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote-sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[18] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 1351–1362, 2005.

[19] M. Fauvel, J. Benediktsson, J. Chanussot, and J. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 11, pp. 3804–3814, 2008.

[20] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[21] D. Tuia, G. Matasci, G. Camps-Valls, and M. Kanevski, "Learning relevant image features with multiple kernel classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, pp. 3780–3791, 2010.

[22] C. Wang, D. You, Y. Y. Zhang, S. Wang, and Y. Zhang, "Representative multiple kernel learning for classification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, pp. 2852–2865, 2012.

[23] K. Bakos and P. Gamba, "Hierarchical hybrid decision tree fusion of multiple hyperspectral data processing chains," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 1, pp. 388–394, Jan 2011.

[24] E. . J. X. . G.-R. M. Tuia, D. ; Merenyi, "Foreword to the special issue on machine learning for remote sensing data processing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 4, pp. 1007–1011, Jul 2014.

[25] G. Camps-Valls, L. Gomez-Chova, J. Muz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, pp. 93–97, 2006.

[26] J. Li, P. Marpu, A. Plaza, J. Bioucas-Dias, and J. A. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 9, pp. 4816–4829, 2013.

[27] D. Böhning, "Multinomial logistic regression algorithm," *Ann. Inst. Statist. Math.*, vol. 44, pp. 197–200, 1992.

[28] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 10, pp. 3747–3762, 2010.

[29] J. Li, J. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, pp. 4085–4098, 2010.

[30] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 6, pp. 957–968, 2005.

[31] J. Li, J. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 19, pp. 3947–3960, 2011.

[32] J. S. Borges, J. M. Bioucas-Dias, and A. R. S. Marcal, "Bayesian hyperspectral image segmentation with discriminative class learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 2151–2164, 2011.

[33] J. Bioucas-Dias and M. Figueiredo, "Logistic regression via variable splitting and augmented Lagrangian tools," Instituto Superior Técnico, TULisbon, Tech. Rep., 2009.

[34] P. Marpu, M. Pedergnana, M. Dalla Mura, J. Benediktsson, and L. Bruzzone, "Automatic generation of standard deviation attribute profiles for spectral spatial classification of remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 2, pp. 293–297, March 2013.

[35] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *International Journal of Remote Sensing*, vol. 31, no. 22, pp. 5975–5991, 2010.

[36] G. Licciardi, F. Pacifici, D. Tuia, S. Prasad, T. West, F. Giacco, C. Thiel, J. Inglada, E. Christophe, J. Chanussot, and P. Gamba, "Decision fusion for the classification of hyperspectral data: Outcome of the 2008 grs-s data fusion contest," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 11, pp. 3857–3865, Nov 2009.

[37] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogrammetric engineering and remote sensing*, vol. 70, no. 5, pp. 627–633, 2004.