

A Composite Semisupervised SVM for Classification of Hyperspectral Images

Mattia Marconcini, *Member, IEEE*, Gustavo Camps-Valls, *Senior Member, IEEE*, and Lorenzo Bruzzone, *Senior Member, IEEE*

Abstract—This letter presents a novel composite semisupervised support vector machine (SVM) for the spectral–spatial classification of hyperspectral images. In particular, the proposed technique exploits the following: 1) unlabeled data for increasing the reliability of the training phase when few training samples are available and 2) composite kernel functions for simultaneously taking into account spectral and spatial information included in the considered image. Experiments carried out on a hyperspectral image pointed out the effectiveness of the presented technique, which resulted in a significant increase of the classification accuracy with respect to both supervised SVMs and progressive semisupervised SVMs with single kernels, as well as supervised SVMs with composite kernels.

Index Terms—Composite kernels, kernel methods, remote-sensing hyperspectral image classification, semisupervised classification, support vector machines (SVMs).

I. INTRODUCTION

IN THE LAST decade, hyperspectral imaging has rapidly become an effective remote-sensing technology for many different applications. In particular, hyperspectral scanners, with respect to earlier multispectral sensors, allow one to discriminate species with very similar spectral signatures. However, while, on the one hand, this increased spectral resolution makes possible an accurate detection and identification, on the other hand, the high dimensionality of data significantly increases the complexity of the analysis [1], [2].

In this framework, kernel-based methods (KMs) have proven to be an effective tool for addressing hyperspectral image classification [3], [4]. *Supervised* kernel classifiers (e.g., support vector machines (SVMs) [5], [6] and kernel Fisher discriminant analysis [7]) are generally preferred to *unsupervised* kernel classifiers (e.g., support vector clustering [8] and support vector data description [9]). However, since gathering reliable prior information is often expensive (both in terms of time and economic costs), in most real applications, the amount of available training data is relatively small compared to the number of features (and thus of classifier parameters). This affects

the learning of supervised systems, resulting in the Hughes phenomenon [10].

Semisupervised approaches recently presented in the literature proved capable of mitigating the aforementioned problem [11]–[13]. In particular, the exploitation of both training data and unlabeled patterns allows semisupervised approaches to outperform standard supervised techniques when few training samples are available. The attention has been mainly focused on the development of semisupervised techniques based on SVMs, which demonstrated to be particularly effective in classification of hyperspectral images. The authors defined a progressive semisupervised SVM (PS³VM) technique, which exhibited very good discrimination capabilities even in critical situations [13]. The PS³VM technique is based on an iterative self-labeling strategy: the most significant unlabeled samples with the highest probability of being correctly classified are selected and labeled according to the discriminant function computed at the considered iteration. Then, these samples are included into the training set with a proper regularization parameter in order to gradually drive the system toward a correct solution.

As concerns the definition of feature mappings for handling hyperspectral images with KMs, only the spectral information is usually considered in the literature. However, recent works have proven that accounting also for the texture (or the local spatial information) permits one to increase the discrimination capability. An effective strategy for combining spectral and spatial information sources proved to be the use of kernel composition [14]. In [15], a new family of composite spectral–spatial kernels in the framework of supervised SVMs was presented, which resulted in a significant increase of the classification accuracy with respect to the standard approach where only spectral features are considered.

According to the previous observations, this letter presents a novel composite classifier based on PS³VM for addressing spectral–spatial categorization of hyperspectral images. In particular, the proposed technique properly integrates the following: 1) PS³VMs, which exploit unlabeled data for increasing the reliability of the training phase when few training samples are available, and 2) composite kernel functions, which allow one to effectively take into account spectral and spatial information included in the considered image. Experiments carried out on a hyperspectral image acquired by the Airborne Visible/InfraRed Imaging Spectrometer (AVIRIS) sensor pointed out the effectiveness of the proposed technique.

This letter is organized as follows. Section II presents the formulation of the proposed composite PS³VM. Section III reports and analyzes experimental results. Section IV draws the conclusion of the work.

Manuscript received July 19, 2008; revised October 3, 2008. First published January 27, 2009; current version published April 17, 2009. This work was supported in part by the Italian Ministry of Education, University and Research and in part by the Spanish Ministry of Education and Science.

M. Marconcini and L. Bruzzone are with the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, 38050 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it).

G. Camps-Valls is with the Departament d'Enginyeria Electrònica, Escola Tècnica Superior d'Enginyeria, Universitat de València, 46100 València, Spain (e-mail: gustavo.camps@uv.es).

Digital Object Identifier 10.1109/LGRS.2008.2009324

II. COMPOSITE PS³VMs

In this section, we describe the proposed composite PS³VM classifier by presenting first the rationale and then the mathematical formulation.

A. Rationale of the Proposed Technique

The rationale of the proposed technique is based on the following observations.

- 1) In hyperspectral data sets with a small ratio between the number of training samples and the number of features, PS³VMs resulted effective in obtaining good classification accuracies by jointly exploiting both labeled and unlabeled patterns for the learning of the algorithm [13]. However, PS³VMs have been studied by exploiting spectral but not contextual information.
- 2) The good performances exhibited by KMs using the original spectral bands as input features can be further increased by including spatial (contextual) information according to the use of kernel composition [15]. However, while composite kernels have already proven to be particularly useful with supervised SVMs, their effectiveness has not been demonstrated with semisupervised SVMs.

According to the two aforementioned observations, we aim at defining a novel composite PS³VM, which integrates the properties of PS³VMs and kernel composition for mitigating the Hughes phenomenon and accurately modeling the presence of different (spectral and spatial) information sources.

B. Formulation of the Proposed Technique

In the following, for the sake of simplicity, we describe the proposed composite PS³VM in the case of a two-class problem. Let us consider a hyperspectral image \mathcal{I}_ω made up of B_ω spectral bands. According to specific feature extraction applied to the neighborhood of each sample of the image (e.g., the local mean or standard deviation per spectral band), from the set of B_ω available “spectral” features, we define a new set of B_s “spatial” contextual features. Accordingly, let $\mathcal{I}_{\omega s} \in \mathbb{R}^{(B_\omega + B_s)}$ represent the resulting new image obtained by stacking both spectral and spatial features. Each pattern $\mathbf{x}_n = \{\mathbf{x}_n^\omega, \mathbf{x}_n^s\} \in \mathcal{I}_{\omega s}$ is given by the concatenation of \mathbf{x}_n^ω and \mathbf{x}_n^s , which represent the spectral and spatial components, respectively.

Let us now consider the training set $\mathcal{T} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X} = \{\mathbf{x}_l\}_{l=1}^N \in \mathcal{I}_{\omega s}$ is a subset of $\mathcal{I}_{\omega s}$ composed of N patterns for which true labels are available, i.e., $\mathcal{Y} = \{y_l\}_{l=1}^N$, $y_l \in \{-1, +1\}$. Moreover, let $\mathcal{X}' = \{\mathbf{x}'_u\}_{u=1}^M \in \mathcal{I}_{\omega s}$ be another subset of M unlabeled samples drawn from $\mathcal{I}_{\omega s}$, such that $\mathcal{X} \cap \mathcal{X}' = \emptyset$. Finally, let us define the transformation $\Phi(\mathbf{x}_n) \triangleq \{\varphi_\omega(\mathbf{x}_n^\omega), \varphi_s(\mathbf{x}_n^s)\}$ as the concatenation of nonlinear mappings $\varphi_\omega(\cdot)$ and $\varphi_s(\cdot)$ into Hilbert spaces for spectral and spatial features alone, respectively.

The proposed algorithm is made up of three main phases [13]: 1) initialization (only training samples in \mathcal{T} are used for initializing the discriminant function); 2) iterative semisupervised learning (training samples in \mathcal{T} and unlabeled samples in \mathcal{X}' are used for gradually adapting the discriminant function); and 3) convergence (all the unlabeled samples in \mathcal{X}' are labeled according to the final discriminant function). In the following, $\mathcal{T}^{(i)}$ and $\mathcal{X}'^{(i)}$ will denote the current training and unlabeled

set at the generic iteration i , respectively, whereas both the subscripts and superscripts ω and s will refer to spectral and spatial components, respectively.

1) *Phase 1—Initialization*: In the first phase, as for supervised SVMs, input data are transformed by $\Phi(\cdot)$, and an initial separation hyperplane $h^{(0)} : f^{(0)}(\mathbf{x}_n) = \mathbf{w}^{(0)} \cdot \Phi(\mathbf{x}_n) + b^{(0)} = 0$ (\mathbf{x}_n represents a generic sample of $\mathcal{I}_{\omega s}$) is determined on the basis of training patterns alone in the Hilbert space $\mathcal{H}_{\omega s} = \mathcal{H}_\omega \oplus \mathcal{H}_s$, where \oplus represents the direct sum operation in $\mathcal{H}_{\omega s}$. We have that $\mathcal{T}^{(0)} = \{(\mathbf{x}_l, y_l)\}_{l=1}^N$ and $\mathcal{X}'^{(0)} = \{\mathbf{x}'_u\}_{u=1}^M$. Accordingly, the function to minimize is

$$\begin{cases} \min_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}^{(0)}\|^2 + C \sum_{l=1}^N \xi_l \right\} \\ y_l [\mathbf{w}^{(0)} \cdot \Phi(\mathbf{x}_l) + b^{(0)}] \geq 1 - \xi_l \\ \xi_l \geq 0 \end{cases} \quad \forall l = 1, \dots, N \quad (1)$$

where \mathbf{w} is a vector normal to h , b is a constant such that $b/\|\mathbf{w}\|^2$ represents the distance of h from the origin, ξ_n denotes the slack variables allowing for (permitted) errors, and C is the associated *penalization parameter*, which permits to tune the generalization capability. The resulting dual Lagrange function to maximize is defined as

$$L(\boldsymbol{\alpha}) = \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n,m=1}^N y_n y_m \alpha_n \alpha_m \Phi(\mathbf{x}_n) \cdot \Phi(\mathbf{x}_m). \quad (2)$$

To avoid consideration of the $\Phi(\cdot)$ mapping explicitly, it is possible to exploit Mercer's theorem [6]. As all mappings in (2) occur in the form of an inner product, we can replace them with a proper kernel function $K_{\omega s}(\mathbf{x}_n, \mathbf{x}_m) = \Phi(\mathbf{x}_n) \cdot \Phi(\mathbf{x}_m)$ which ensures that the Lagrangian function is convex

$$L(\boldsymbol{\alpha}) = \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n,m=1}^N y_n y_m \alpha_n \alpha_m K_{\omega s}(\mathbf{x}_n, \mathbf{x}_m). \quad (3)$$

Any function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a valid kernel if (and only if), for any finite set of N samples, it produces kernel matrices $\mathbf{K} = [K_{nm}]_{n,m=1}^N = [K(\mathbf{x}_n, \mathbf{x}_m)]_{n,m=1}^N$ that are both symmetric [i.e., $K_{nm} = K_{mn}$] and positive definite [i.e., $\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \geq 0, \forall \boldsymbol{\alpha} \in \mathbb{R}^N$]. With respect to the standard approach with single kernels, we can exploit the significant advantage of modeling the solution as the sum of positive-definite matrices accounting for both the spatial and spectral counterparts

$$\begin{aligned} K_{\omega s}(\mathbf{x}_n, \mathbf{x}_m) &= \Phi(\mathbf{x}_n) \cdot \Phi(\mathbf{x}_m) \\ &= \{\varphi_\omega(\mathbf{x}_n^\omega), \varphi_s(\mathbf{x}_n^s)\} \cdot \{\varphi_\omega(\mathbf{x}_m^\omega), \varphi_s(\mathbf{x}_m^s)\} \\ &= K_\omega(\mathbf{x}_n^\omega, \mathbf{x}_m^\omega) + K_s(\mathbf{x}_n^s, \mathbf{x}_m^s) \end{aligned} \quad (4)$$

where $K_\omega(\cdot, \cdot)$ and $K_s(\cdot, \cdot)$ are kernel functions associated with spectral and spatial components, respectively. Note that $\boldsymbol{\alpha}^T (\mathbf{K}_\omega + \mathbf{K}_s) \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \mathbf{K}_\omega \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \mathbf{K}_s \boldsymbol{\alpha} \geq 0$, so $(\mathbf{K}_\omega + \mathbf{K}_s)$

is positive semidefinite and $K_{\omega_s}(\cdot, \cdot)$ is a valid kernel function. The dual problem can be formulated as follows:

$$\left\{ \begin{array}{l} \max_{\alpha} \left\{ \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n,m=1}^N y_n y_m \alpha_n \alpha_m \right. \\ \left. [K_{\omega}(\mathbf{x}_n^{\omega}, \mathbf{x}_m^{\omega}) + K_s(\mathbf{x}_n^s, \mathbf{x}_m^s)] \right\} \\ \sum_{n=1}^N y_n \alpha_n = 0 \\ 0 \leq \alpha_n \leq C, \quad \forall n = 1, \dots, N \end{array} \right\} \quad (5)$$

where the coefficients $\alpha_{n=1}^N$ represent the Lagrange multipliers. According to the Karush–Kuhn–Tucker conditions [6] (which are necessary and sufficient conditions for solving (5) with respect to α), the solution is a linear combination of the only training patterns associated with nonzero multipliers (i.e., either mislabeled training samples or correctly labeled training samples falling into the margin band $\mathcal{M} = \{\mathbf{x}_n | -1 \leq f(\mathbf{x}_n) \leq 1\}$), denoted as *support vectors*.

2) *Phase 2—Iterative Semisupervised Learning*: At the i th iteration, according to the current decision function $f^{(i)}(\mathbf{x}_n) = \mathbf{w}^{(i)} \cdot \Phi(\mathbf{x}_n) + b^{(i)}$, the estimated labels $\hat{y}_u^{(i)} = \text{sgn}[f^{(i)}(\mathbf{x}'_u)]$ are given to all the originally unlabeled samples $\mathbf{x}'_u \in \mathcal{X}'^{(i)}$. Then, a subset of the remaining unlabeled samples is iteratively selected and moved (together with the corresponding estimated labels) into the training set $\mathcal{T}^{(i+1)}$. On the one hand, the higher the distance from the separation hyperplane $h^{(i)} : \mathbf{w}^{(i)} \cdot \Phi(\mathbf{x}_n) + b^{(i)} = 0$, the higher the chance for an unlabeled sample to be correctly classified. On the other hand, the current unlabeled samples falling into the margin band are those with the highest probability to be associated with nonzero Lagrange multipliers once inserted into the training set (and thus can affect the position of $h^{(i+1)}$). According to these two observations, at each iteration, we consider from both sides of the margin the ρ unlabeled patterns (where the parameter $\rho \geq 1$ is defined *a priori* by the user) lying further from the decision boundary $h^{(i)}$. Such samples are called *semilabeled* samples, and the set of semilabeled patterns selected at the generic i th iteration is denoted as $\mathcal{H}^{(i)}$.

For $i \geq 1$, the bound minimization problem can be written as

$$\left\{ \begin{array}{l} \min_{\mathbf{w}, b, \xi, \xi'} \left\{ \frac{1}{2} \|\mathbf{w}^{(i)}\|^2 + C \sum_{l=1}^N \xi_l + \sum_{u=1}^{\mu^{(i)}} C_u^* \xi'_u \right\} \\ y_l \cdot [\mathbf{w}^{(i)} \cdot \Phi(\mathbf{x}_l) + b^{(i)}] \geq 1 - \xi_l, \quad \forall l = 1, \dots, N \\ \hat{y}_u^{(i-1)} \cdot [\mathbf{w}^{(i)} \cdot \Phi(\mathbf{x}'_u) + b^{(i)}] \geq 1 - \xi'_u, \quad \forall u = 1, \dots, \mu^{(i)} \\ \xi_l, \xi'_u \geq 0 \end{array} \right\} \quad (6)$$

where $\mu^{(i)}$ represents the number of semilabeled samples into the training set. The semilabeled samples $(\mathbf{x}'_u, \hat{y}_u^{(i-1)}) \in \mathcal{T}^{(i)}$ are associated with a regularization parameter $C_u^* \in \mathbb{R}^+$ that grows, depending on the number of iterations for which they have been assigned the same estimated label until iteration $i - 1$. The longer a semilabeled samples is associated with the same information class, the higher is expected to be the confidence of the system on that pattern. In fact, on increasing the value of C_u^* , the influence of the associated sample on the definition of the separation hyperplane increases. In order to avoid instabilities, the regularization parameter for the semilabeled

patterns increases gradually in a quadratic way. In particular, for the u th semilabeled sample, we have $\forall k = 1, \dots, \gamma$

$$C_u^* = \frac{C^{* \max} - C^*}{(\gamma - 1)^2} (k - 1)^2 + C^* \Leftrightarrow (\mathbf{x}'_u, \hat{y}_u^{(i-1)}) \in \mathcal{J}_k^{(i-1)} \quad (7)$$

where $\mathcal{J}_k^{(i-1)}$ includes all the current semilabeled patterns in the training set that have been assigned the same estimated label for k successive iterations, C^* is the initial regularization value for semilabeled samples (this is a user-defined parameter), and $C^{* \max}$ is the maximum regularization value of semilabeled samples and is related to C (i.e., $C^{* \max} = \tau \cdot C$, with $0 < \tau \leq 1$ being a constant). In (7), γ is defined as the maximum number of iterations for which the user allows the regularization parameter for semilabeled samples to increase.

In this case, the dual Lagrange functional becomes

$$\begin{aligned} L(\alpha, \alpha') = & \sum_{n=1}^N \alpha_n + \sum_{m=1}^{\mu^{(i)}} \alpha'_m - \frac{1}{2} \sum_{n,m=1}^N \alpha_n \alpha_m y_n y_m \Phi(\mathbf{x}_n) \\ & \cdot \Phi(\mathbf{x}_m) - \frac{1}{2} \sum_{n,m=1}^{\mu^{(i)}} \alpha'_n \alpha'_m \hat{y}_n^{(i-1)} \hat{y}_m^{(i-1)} \Phi(\mathbf{x}'_n) \\ & \cdot \Phi(\mathbf{x}'_m) - \sum_{n=1}^N \sum_{m=1}^{\mu^{(i)}} \alpha_n \alpha'_m y_n \hat{y}_m^{(i-1)} \Phi(\mathbf{x}_n) \cdot \Phi(\mathbf{x}'_m). \end{aligned} \quad (8)$$

We can notice that, as in (2), the $\Phi(\cdot)$ mappings only occur as inner products. Accordingly, the dual problem can be written as

$$\left\{ \begin{array}{l} \max_{\alpha, \alpha'} \left\{ \sum_{n=1}^N \alpha_n + \sum_{m=1}^{\mu^{(i)}} \alpha'_m - \frac{1}{2} \sum_{n,m=1}^N \alpha_n \alpha_m y_n y_m \right. \\ \times [K_{\omega}(\mathbf{x}_n^{\omega}, \mathbf{x}_m^{\omega}) + K_s(\mathbf{x}_n^s, \mathbf{x}_m^s)] \\ - \frac{1}{2} \sum_{n,m=1}^{\mu^{(i)}} \alpha'_n \alpha'_m \hat{y}_n^{(i-1)} \hat{y}_m^{(i-1)} \\ \times [K_{\omega}(\mathbf{x}_n^{\omega}, \mathbf{x}_m^{\omega}) + K_s(\mathbf{x}_n^s, \mathbf{x}_m^s)] \\ \left. - \sum_{n=1}^N \sum_{m=1}^{\mu^{(i)}} \alpha_n \alpha'_m y_n \hat{y}_m^{(i-1)} [K_{\omega}(\mathbf{x}_n^{\omega}, \mathbf{x}_m^{\omega}) + K_s(\mathbf{x}_n^s, \mathbf{x}_m^s)] \right\} \\ \sum_{n=1}^N \alpha_n y_n + \sum_{m=1}^{\mu^{(i)}} \alpha'_m \hat{y}_m^{(i-1)} = 0 \\ 0 \leq \alpha_n \leq C, \quad \forall n = 1, \dots, N \\ 0 \leq \alpha'_m \leq C^*_m, \quad \forall m = 1, \dots, \mu^{(i)}. \end{array} \right\} \quad (9)$$

Note that since the position of the separation hyperplane may change at each iteration, a proper dynamical adjustment is necessary. Accordingly, the semilabeled samples whose estimated labels at iteration i are different than those at iteration $i - 1$ (i.e., $\mathcal{S}^{(i)} = \{(\mathbf{x}'_u, \hat{y}_u^{(i-1)}) \in \mathcal{T}^{(i)} | \hat{y}_u^{(i)} \neq \hat{y}_u^{(i-1)}\}$) are reset to the unlabeled state and moved back to $\mathcal{X}'^{(i+1)}$ in order to be reconsidered at the following iterations.

3) *Phase 3—Convergence*: The algorithm stops when the following empirical criterion is satisfied:

$$\left\{ \begin{array}{l} |\mathcal{H}^{(i)}| \leq \lceil \beta \cdot M \rceil \\ |\mathcal{S}^{(i)}| \leq \lceil \beta \cdot M \rceil \end{array} \right\} \quad (10)$$

where M is the original number of unlabeled samples and β is a constant fixed *a priori* that tunes the sensitivity of the learning process. This means that convergence is reached if both the

TABLE I
NUMBER OF TRAINING AND TEST PATTERNS

	Information Classes	Training Set	Test Set
1	Scrub	68 (10.95%)	693 (15.10%)
2	Willow Swamp	39 (6.28%)	204 (4.44%)
3	Cabbage Palm Hammock	51 (8.21%)	205 (4.47%)
4	Cabbage Palm / Oak Hammock	62 (9.98%)	190 (4.14%)
5	Slash Pine	23 (3.70%)	138 (3.00%)
6	Oak / Broadleaf Hammock	55 (8.86%)	174 (3.79%)
7	Hardwood Swamp	27 (4.35%)	78 (1.70%)
8	Graminoid Marsh	34 (5.48%)	397 (8.65%)
9	Spartina Marsh	33 (5.31%)	487 (10.61%)
10	Cattail Marsh	57 (9.18%)	347 (7.56%)
11	Salt Marsh	73 (11.76%)	346 (7.54%)
12	Mud Flats	61 (9.82%)	442 (9.63%)
13	Water	38 (6.12%)	889 (19.37%)
	Overall	621	4590

number of mislabeled patterns and the number of remaining unlabeled patterns lying into the margin band at the current iteration are lower than or equal to $\lceil \beta \cdot M \rceil$.

At the end of the learning process, for any given input pattern $\mathbf{x}_n = \{\mathbf{x}_n^\omega, \mathbf{x}_n^s\} \in \mathcal{I}_{\omega s}$, the corresponding predicted label is $\hat{y}_n = \text{sgn}[f(\mathbf{x}_n)]$.

Given a kernel function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and a scalar $\delta \in \mathbb{R}_0^+$, it holds that $K_\delta(\cdot, \cdot) = \delta K(\cdot, \cdot)$ is a valid kernel, as $\boldsymbol{\alpha}^T \mathbf{K}_\delta \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \delta \mathbf{K} \boldsymbol{\alpha} = \delta \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \geq 0$. This important property lets us also define another kernel composition rule that permits one to balance the spatial and spectral contents of the image. In particular, in place of a direct summation of kernel functions, we can consider also a weighted concatenation of nonlinear transformations of \mathbf{x}_n^ω and \mathbf{x}_n^s :

$$K_\delta(\mathbf{x}_n, \mathbf{x}_m) = \mu \cdot K_\omega(\mathbf{x}_n^\omega, \mathbf{x}_m^\omega) + (1 - \mu) \cdot K_s(\mathbf{x}_n^s, \mathbf{x}_m^s) \quad (11)$$

where $0 < \mu < 1$ tunes spectral and spatial information.

III. EXPERIMENTAL RESULTS

The data set considered in the experimental phase consists of a 460×512 hyperspectral image acquired by the AVIRIS sensor over the KSC area (Florida, USA) on March 1996 [16], [17]. The data, have a spatial resolution of 18 m. After removing water absorption and low SNR bands, 176 of the original 224 bands were used for the experimental analysis.

The investigated site represents a series of impounded estuarine wetlands of the northern Indian River Lagoon [16]. For classification purposes, 13 land-cover classes were defined (see Table I). On the basis of *in situ* observations, a training and a test set (drawn from different homogenous areas) made up of 621 and 4590 patterns, respectively, were defined (see Table I).

In all the experiments, we took into account spatial information by simply computing the mean in a 5×5 window for each spectral channel (i.e., \mathbf{x}_n^s is the vector of mean values computed for a 5×5 neighborhood of \mathbf{x}_n^ω in each spectral band), so $B_s = B_\omega = 176$.

For comparisons, we analyzed the accuracies obtained by supervised SVMs and PS³VMs with single kernels, as well as by SVMs with composite kernels. Moreover, we compared the results with the ones obtained by employing the *stacked approach* for both SVMs and PS³VMs. This approach represents the simplest (and widely used) way for integrating spectral and spatial information by building feature vectors from the concatenation of spectral and spatial features while employing a single basic kernel function.

In general, RBF Gaussian K^{RBF} and polynomial K^p kernels proved to be particularly effective in addressing classification

of hyperspectral images [4]. Thus, in the experimental phase, we used such types of functions also when defining composite kernels. All the classifiers were trained according to a fivefold cross-validation (CV) strategy [18]. However, while, for training supervised SVMs, we used only labeled training samples, for training both the PS³VM and the proposed composite PS³VM, we also considered test samples (modeled as unlabeled). In the model selection phase, a grid search strategy was used for both “supervised” parameters (i.e., the spread σ of the Gaussian function, the degree d of the polynomial, and the regularization parameter C) and “semisupervised” parameters (i.e., ρ , γ , and C^*). As concerns the weighted summation kernel, we evaluated combinations with both $\mu = 0.25$ and $\mu = 0.75$. Taking into account that $0 < \mu < 1$ and that fixing $\mu = 0.5$ is equivalent to considering the direct summation case, these two values make it possible to understand if the weighted approach can lead to a further improvement.

The sequential minimal optimization algorithm [19] was employed for training supervised SVMs as well as, with proper modifications, both the PS³VM and the proposed composite PS³VM classifiers. The maximum possible value for the regularization parameter associated with semilabeled samples was fixed to $(C/2)$ (i.e., $\tau = 0.5$). A reasonable empirical choice for the convergence criterion proved to be $\beta = 3 \cdot 10^{-2}$.

Table II(a) reports the percentage of overall accuracy ($OA\%$) and kappa coefficient of accuracy obtained on the available test samples by both SVMs and PS³VMs with single kernels. Table II(b) compares the accuracies exhibited by the proposed composite PS³VMs with those obtained by supervised SVMs with composite kernels. The results confirm the effectiveness of the proposed technique, which outperformed both supervised SVMs and PS³VMs with single kernels, as well as SVMs with composite kernels. On the one hand, this means that, also when composite kernels are considered, the use of semisupervised SVMs involves higher accuracies than standard supervised SVMs. On the other hand, it comes out that, besides supervised SVMs, also the presented approach largely benefit from the employment of kernel composition.

The proposed composite PS³VMs resulted in a sharp increase of accuracy with respect to supervised SVMs (i.e., on average of approximately 6.5% in terms of kappa). In the best case (i.e., for the weighted summation kernel $0.75 \cdot K_\omega^{\text{RBF}} + 0.25 \cdot K_s^p$), the kappa coefficient increased by more than 9%. However, even when SVMs exhibited their best performances, composite PS³VMs provided a gain in kappa higher than 3%.

Despite the fact that the stacked approach only allowed PS³VMs to slightly improve the performances with respect to the single kernel case, the employment of kernel composition with semisupervised SVMs resulted in a sharp increase of accuracy. In particular, when the summation of RBF Gaussian kernels for spectral features and polynomial kernels for spatial features was considered, the kappa improvement was around 10%.

As expected, supervised SVMs exhibited fair performances when only spectral information was employed. With the stacked approach, better accuracies could be obtained; nevertheless, the best results occurred when composite kernels were considered. In particular, the highest accuracies have been obtained with combinations involving polynomial kernels both for spectral and spatial features (the increase with respect to

TABLE II
PERCENTAGE OF OVERALL ACCURACY (OA%) AND KAPPA COEFFICIENT OF ACCURACY OBTAINED ON TEST DATA BY THE FOLLOWING:
(a) SUPERVISED SVMs AND PS³VMS WITH SINGLE KERNEL FUNCTIONS TRAINED ACCORDING TO A FIVEFOLD CV STRATEGY AND
(b) SUPERVISED SVMs WITH COMPOSITE KERNELS AND THE PROPOSED COMPOSITE PS³VMS TRAINED ACCORDING TO A FIVEFOLD CV STRATEGY. SUPERSCRIPTS REFER TO THE TYPE OF KERNEL [I.E., RBF GAUSSIAN (RBF) OR POLYNOMIAL (p)]. SUBSCRIPTS REFER TO THE TYPE OF COMPONENTS CONSIDERED [I.E., SPECTRAL COMPONENTS ALONE (ω), SPATIAL COMPONENTS ALONE (s), OR CONCATENATION OF SPECTRAL AND SPATIAL COMPONENTS ($\{\omega, s\}$)]

Single Kernel function	OA%			Kappa		
	SVM	PS ³ VM	% Δ	SVM	PS ³ VM	% Δ
K_{ω}^{RBF}	74.62	80.02	+5.40	0.717	0.776	+5.9
K_{ω}^p	76.43	83.18	+6.75	0.737	0.812	+7.5
$K_{\{\omega,s\}}^{RBF}$	79.59	83.18	+3.59	0.771	0.812	+4.1
$K_{\{\omega,s\}}^p$	80.35	84.84	+4.49	0.780	0.823	+4.3

(a)

Composite Kernel function	OA%			Kappa		
	SVM	proposed composite PS ³ VM	% Δ	SVM	proposed composite PS ³ VM	% Δ
$K_{\omega}^{RBF} + K_s^{RBF}$	79.30	83.75	+4.45	0.768	0.818	+5.0
$0.25 \cdot K_{\omega}^{RBF} + 0.75 \cdot K_s^{RBF}$	79.98	84.07	+4.09	0.776	0.822	+4.6
$0.75 \cdot K_{\omega}^{RBF} + 0.25 \cdot K_s^{RBF}$	79.48	83.79	+4.31	0.770	0.819	+4.9
$K_{\omega}^{RBF} + K_s^p$	81.57	89.24	+7.67	0.795	0.879	+8.4
$0.25 \cdot K_{\omega}^{RBF} + 0.75 \cdot K_s^p$	79.15	88.56	+9.41	0.772	0.872	+10.0
$0.75 \cdot K_{\omega}^{RBF} + 0.25 \cdot K_s^p$	80.09	89.67	+9.58	0.779	0.884	+10.5
$K_{\omega}^p + K_s^{RBF}$	79.85	88.58	+8.73	0.774	0.872	+9.8
$0.25 \cdot K_{\omega}^p + 0.75 \cdot K_s^{RBF}$	81.26	88.69	+7.43	0.790	0.873	+8.3
$0.75 \cdot K_{\omega}^p + 0.25 \cdot K_s^{RBF}$	80.44	87.95	+7.51	0.781	0.864	+8.3
$K_{\omega}^p + K_s^p$	81.81	88.87	+7.06	0.798	0.875	+7.7
$0.25 \cdot K_{\omega}^p + 0.75 \cdot K_s^p$	86.14	89.37	+3.23	0.845	0.881	+3.6
$0.75 \cdot K_{\omega}^p + 0.25 \cdot K_s^p$	85.19	88.85	+3.66	0.835	0.874	+3.9

(b)

SVMs with single kernels is slightly lower than 10% in terms of kappa).

A qualitative analysis of the classification maps (not reported due to space constraints) confirmed the improvement given by the proposed composite PS³VMS technique, which resulted in a better discrimination among different information classes, especially in the most critical regions of the study area.

IV. CONCLUSION

In this letter, we defined a novel composite semisupervised classifier based on SVMs specifically designed for addressing spectral-spatial categorization of hyperspectral images. In particular, the proposed technique exploits the following: 1) unlabeled data for better constraining the learning of the classifier when only few labeled samples are available in the training phase and 2) composite kernel functions for taking into account effectively the spectral information and the local spatial content of the considered image. In this way, we can handle the small ratio between the number of available training patterns and the number of features, as well as the presence of different information sources.

Experiments carried out on an AVIRIS data set confirmed the effectiveness of the proposed technique, which resulted in a very high classification accuracy. In particular, the proposed method outperformed both SVMs and PS³VMS with single kernels, as well as supervised SVMs with composite kernels.

It should be pointed out that when different kernels are associated with spectral and spatial components, the model selection takes a longer time since the optimization of kernel parameter values specific for both functions should be considered, as well as the definition of weights associated with kernels. Nevertheless, taking into account the very good performances exhibited in the experimental phase, this seems a minor drawback for the presented technique.

As a future development of the proposed work, we are planning to extend the experimental analysis to other data sets, using also texture metrics for extracting spatial components. Moreover, we are also investigating the possibility of employing other rules for the kernel combination.

REFERENCES

- [1] L. Bruzzone, L. Gómez-Chova, M. Marconcini, and G. Camps-Valls, "Hyperspectral image classification with kernels," in *Kernel Methods in Bioengineering, Signal and Image Processing*, G. Camps-Valls, J. L. Rojo-Álvarez, and M. Martínez-Ramón, Eds. Hershey, PA: Idea Group Inc., 2007, ch. 17, pp. 374–398.
- [2] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
- [3] K. L. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–202, Mar. 2001.
- [4] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.
- [5] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
- [6] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
- [7] S. Mika, G. Rätsch, and K.-R. Müller, "A mathematical programming approach to the Kernel Fisher algorithm," in *Advances in Neural Information Processing Systems*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2000, pp. 591–597.
- [8] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, pp. 125–137, 2001.
- [9] D. M. Tax and R. P. Duin, "Support vector data description," *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, Jan. 2004.
- [10] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
- [11] O. Chapelle and B. Schölkopf, *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.
- [12] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006.
- [13] L. Bruzzone, M. Chi, and M. Marconcini, "Semisupervised support vector machines for classification of hyperspectral remote sensing images," in *Hyperspectral Data Exploitation: Theory and Applications*, C.-I Chang, Ed. New York: Wiley, 2007, ch. 11, pp. 275–311.
- [14] T. Joachims, N. Cristianini, and J. Shawe-Taylor, "Composite kernels for hypertext categorization," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 250–257.
- [15] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.
- [16] J. Ham, Y. Chen, M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.
- [17] [Online]. Available: <http://www.csr.utexas.edu/hyperspectral/codes.html>
- [18] C. R. Rao and Y. Wu, "On model selection," in *Model Selection*, vol. 38, P. Lahiri, Ed. Beachwood, OH: Inst. Math. Statist., 2001.
- [19] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1998, pp. 185–208.