Composite Kernels for Hyperspectral Image Classification

Gustavo Camps-Valls, *Member, IEEE*, Luis Gomez-Chova, Jordi Muñoz-Marí, Joan Vila-Francés, and Javier Calpe-Maravilla, *Member, IEEE*

Abstract—This letter presents a framework of composite kernel machines for enhanced classification of hyperspectral images. This novel method exploits the properties of Mercer's kernels to construct a family of composite kernels that easily combine spatial and spectral information. This framework of composite kernels demonstrates: 1) enhanced classification accuracy as compared to traditional approaches that take into account the spectral information only: 2) flexibility to balance between the spatial and spectral information in the classifier; and 3) computational efficiency. In addition, the proposed family of kernel classifiers opens a wide field for future developments in which spatial and spectral information can be easily integrated.

Index Terms—Composite kernels, contextual, hyperspectral, image classification, kernel, spectral, support vector machine (SVM), texture.

I. INTRODUCTION

THE information contained in hyperspectral data allows the characterization, identification, and classification of landcovers with improved accuracy and robustness [1]. In the remote sensing literature, many supervised and unsupervised methods have been developed for multispectral and hyperspectral image classification (e.g., maximum-likelihood classifiers, neural networks, neurofuzzy models, etc.) [2]–[4]. However, an important problem in the context of hyperspectral data is the high number of spectral bands and relatively low number of labeled training samples, which poses the well-known Hughes phenomenon [5]. This problem is usually reduced by introducing a feature selection/extraction step before training the hyperspectral classifier with the basic objective of reducing the high input dimensionality. However, including such a step is time-consuming, scenario-dependent, and sometimes requires *a priori* knowledge.

In recent years, *kernel methods* [6], such as support vector machines (SVMs) or kernel Fisher discriminant analysis, have demonstrated excellent performance in hyperspectral data classification in terms of accuracy and robustness [7]–[12]. The properties of kernel methods make them well-suited to tackle the problem of hyperspectral image classification since they can handle large input spaces efficiently, work with a relatively low

number of labeled training samples and deal with noisy samples in a robust way [9], [12], [13].

The good classification performance demonstrated by kernel methods using the spectral signature as input features could be further increased by including contextual (or even textural) information in the classifier, something that has been successfully illustrated in other classification algorithms (expectation–maximization, k-nearest neighbor classifiers, neural networks, etc.) [14]–[16]. However, to the authors' knowledge, kernel methods have so far taken into account the spectral information to develop the classifier [7], [9]–[12], and thus, the spatial variability of the spectral signature has not been considered.

In this letter, we explicitly formulate a full family of kernel-based classifiers that simultaneously take into account spectral, spatial, and local cross-information in a hyperspectral image. For this purpose, we take advantage of two especially interesting properties of kernel methods: 1) their good performance when working with high input dimensional spaces [9], [12] and 2) the properties derived from Mercer's conditions by which a scaled summation of (positive definite) kernel matrices are valid kernels, which have provided good results in other domains [17], [18]. Among all the available kernel machines, we focus on SVMs, which have recently demonstrated superior performance in the context of hyperspectral image classification [12], [19]. In any case, the formulations proposed in this letter are valid for any kernel classifier.

The letter is outlined as follows. Section II briefly reviews the formulation of SVM classifiers. Section III discusses the concept and properties of Mercer's kernels. Section IV presents the formulation of composite kernels for the versatile combination of spatial and spectral information for hyperspectral image classification. Section V presents the experimental results. In Section VI, we conclude this letter with further work, research opportunities, and final remarks.

II. SUPPORT VECTOR CLASSIFIERS

Given a labeled training dataset $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^N$ and $y_i \in \{-1, +1\}$, and a nonlinear mapping $\boldsymbol{\phi}(\cdot)$, usually to a higher (possibly infinite) dimensional (Hilbert) space, $\boldsymbol{\phi} : \mathbb{R}^N \longrightarrow \mathcal{H}$, the SVM method solves

$$\min_{\mathbf{w}, \xi_i, b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \right\} \tag{1}$$

constrained to

$$y_i \left(\langle \boldsymbol{\phi}(\mathbf{x}_i), \mathbf{w} \rangle + b \right) \geq 1 - \xi_i, \qquad \forall i = 1, \ldots, n \tag{2}$$

$$\xi_i \geq 0, \qquad \forall i = 1, \ldots, n \tag{3}$$

where $\mathbf{w}$ and b define a linear classifier in the feature space. The nonlinear mapping function $\boldsymbol{\phi}$ is performed in accordance with Cover's theorem [20], which guarantees that the transformed samples are more likely to be linearly separable in the resulting feature space. The regularization parameter C controls the generalization capabilities of the classifier and it must be selected by the user, and ξ_i are positive slack variables enabling to deal with permitted errors.

Due to the high dimensionality of vector variable $\mathbf{w}$, primal function (1) is usually solved through its Lagrangian dual problem, which consists of solving

$$\max_{\alpha_i} \left\{ \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle \right\} \quad (4)$$

constrained to $0 \le \alpha_i \le C$ and $\sum_i \alpha_i y_i = 0$, $i = 1, \ldots, n$, where auxiliary variables α_i are Lagrange multipliers corresponding to constraints in (2). It is worth noting that all $\boldsymbol{\phi}$ mappings used in the SVM learning occur in the form of inner products. This allows us to define a kernel function K

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle \quad (5)$$

and then a nonlinear SVM can be constructed using only the kernel function, without having to consider the mapping $\boldsymbol{\phi}$ explicitly. Then, by introducing (5) into (4), the dual problem is obtained. After solving this dual problem, $\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \boldsymbol{\phi}(\mathbf{x}_i)$, and the decision function implemented by the classifier for any test vector $\mathbf{x}$ is given by

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \right) \quad (6)$$

where b can be easily computed from the α_i that are neither 0 nor C, as explained in [6].

III. PROPERTIES OF MERCER'S KERNELS

In the context of SVMs in particular and kernel methods in general, one can use any kernel function $K(\cdot, \cdot)$ that fullfils Mercer's condition, which can be stated formally in the following theorem.

Theorem 1: Mercer's Kernel. Let $\mathcal{X}$ be any input space and $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathcal{R}$ a symmetric function, K is a Mercer's kernel if and only if the kernel matrix formed by restricting K to any finite subset of $\mathcal{X}$ is *positive semidefinite*, i.e., having no negative eigenvalues.

The Mercer condition constitutes the key requirement to obtain a unique global solution when developing kernel-based classifiers (e.g., SVMs) since they reduce to solving a convex optimization problem [13]. In addition, important properties for Mercer's kernels can be derived from the fact that they are positive-definite (affinity) matrices, as follows.

Proposition 1: Properties of Mercer's Kernels. Let K_1, K_2, and K_3 be valid Mercer's kernels over $\mathcal{X} \times \mathcal{X}$, with $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^N$, with $\mathbf{A}$ being a symmetric positive semidefinite $N \times N$ matrix, and $\alpha > 0$. Then the following functions are valid kernels: 1) $K(\mathbf{x}_i, \mathbf{x}_j) = K_1(\mathbf{x}_i, \mathbf{x}_j) + K_2(\mathbf{x}_i, \mathbf{x}_j)$; 2) $K(\mathbf{x}_i, \mathbf{x}_j) = \alpha K_1(\mathbf{x}_i, \mathbf{x}_j)$; and 3) $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{A} \mathbf{x}_j$.

It is worth noting that the size of the training kernel matrix is $n \times n$ and each position (i, j) of matrix $(K)_{ij}$ contains the similarity among all possible pairs of training samples ($\mathbf{x}_i$ and $\mathbf{x}_j$) measured with a suitable kernel function K fulfilling Mercer's conditions. Some popular kernels are: linear $(K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle)$, polynomial $(K(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1)^d, d \in \mathbb{Z}^+)$, or radial basis function (RBF) $(K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2), \sigma \in \mathbb{R}^+)$. This (distance or similarity) matrix is precomputed at the very beginning of the minimization procedure, and thus, one usually works with the transformed input data K rather than the original input space samples $\mathbf{x}_i$. This fact allows us to easily combine positive definite kernel matrices taking advantage of the properties in Proposition 1, as will be shown in the next section.

IV. COMPOSITE KERNELS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

A full family of composite kernels for the combination of spectral and contextual information is presented in this section. For this purpose, three steps are followed.

1) *Pixel definition:* A pixel entity $\mathbf{x}_i$ is redefined simultaneously both in the spectral domain using its spectral content $\mathbf{x}_i^\omega \in \mathbb{R}^{N_\omega}$ and in the spatial domain by applying some feature extraction to its surrounding area $\mathbf{x}_i^s \in \mathbb{R}^{N_s}$, which yields N_s spatial (contextual) features, e.g., the mean or standard deviation *per* spectral band.

2) *Kernel computation:* Once the spatial and spectral feature vectors $\mathbf{x}_i^s$ and $\mathbf{x}_i^\omega$ are constructed, different kernel matrices can be easily computed using any suitable kernel function that fulfills Mercer's conditions.

3) *Kernel combination.* At this point, we take advantage of the *direct sum* of Hilbert Spaces by which two (or more) Hilbert spaces $\mathcal{H}_k$ can combined into a larger Hilbert space. This well-known result from functional analysis theory [21] allows us to sum spectral and textural dedicated kernel matrices (K_ω and K_s, respectively), and introduce the cross-information between textural and spectral features ($K_{\omega s}$ and $K_{s\omega}$) in the formulation.

In the following, we present four different kernel approaches for the joint consideration of spectral and textural information in a unified framework for hyperspectral image classification.

A. Stacked Features Approach

The most commonly adopted approach in hyperspectral image classification is to exploit the spectral content of a pixel $\mathbf{x}_i \equiv \mathbf{x}_i^\omega$. However, performance can be improved by including both spectral and textural information in the classifier. This is usually done by means of the "stacked" approach, in which feature vectors are built from the concatenation of spectral and spatial features. Note that if the chosen mapping $\boldsymbol{\phi}$ is a transformation of the concatenation $\mathbf{x}_i \equiv \{\mathbf{x}_i^s, \mathbf{x}_i^\omega\}$, then the corresponding "stacked" kernel matrix is

$$K_{\{s,\omega\}} \equiv K(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle \quad (7)$$

which does not include explicit cross relations between $\mathbf{x}_i^s$ and $\mathbf{x}_j^\omega$.

B. Direct Summation Kernel

A simple composite kernel combining spectral and textural information naturally comes from the concatenation of non-

linear transformations of $\mathbf{x}_i^s$ and $\mathbf{x}_i^\omega$. Let us assume two non-linear transformations $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$ into Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. Then, the following transformation can be constructed:

$$\boldsymbol{\phi}(\mathbf{x}_i) = \{\varphi_1(\mathbf{x}_i^s), \varphi_2(\mathbf{x}_i^\omega)\} \tag{8}$$

and the corresponding dot product can be easily computed as follows:

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &= \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle \\ &= \langle \{\varphi_1(\mathbf{x}_i^s), \varphi_2(\mathbf{x}_i^\omega)\}, \{\varphi_1(\mathbf{x}_j^s), \varphi_2(\mathbf{x}_j^\omega)\} \rangle \\ &= K_s(\mathbf{x}_i^s, \mathbf{x}_j^s) + K_\omega(\mathbf{x}_i^\omega, \mathbf{x}_j^\omega). \end{aligned} \tag{9}$$

Note that the solution is expressed as the sum of positive definite matrices accounting for the textural and spectral counterparts, independently. Note that $dim(\mathbf{x}_i^\omega) = N_\omega$, $dim(\mathbf{x}_i^s) = N_s$, and $dim(K) = dim(K_s) = dim(K_\omega) = n \times n$.

C. Weighted Summation Kernel

By exploiting Property 2) in Proposition 1, a composite kernel that balances the spatial and spectral content in (10) can also be created, as follows:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mu K_s(\mathbf{x}_i^s, \mathbf{x}_j^s) + (1 - \mu) K_\omega(\mathbf{x}_i^\omega, \mathbf{x}_j^\omega) \tag{10}$$

where μ is a positive real-valued free parameter ($0 < \mu < 1$), which is tuned in the training process and constitutes a tradeoff between the spatial and spectral information to classify a given pixel. This composite kernel allows us to introduce *a priori* knowledge in the classifier by designing specific μ profiles *per* class, and also allows us to extract some information from the best tuned μ parameter.

D. Cross-Information Kernel

The preceding kernel classifiers can be conveniently modified to account for the cross relationship between the spatial and spectral information. Assume a nonlinear mapping $\varphi(\cdot)$ to a Hilbert space $\mathcal{H}$ and three linear transformations $\mathbf{A}_k$ from $\mathcal{H}$ to $\mathcal{H}_k$, for $k = 1, 2, 3$. Let us construct the following composite vector:

$$\boldsymbol{\phi}(\mathbf{x}_i) = \{\mathbf{A}_1 \varphi(\mathbf{x}_i^s), \mathbf{A}_2 \varphi(\mathbf{x}_i^\omega), \mathbf{A}_3(\varphi(\mathbf{x}_i^s) + \varphi(\mathbf{x}_i^\omega))\} \tag{11}$$

and compute the dot product

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &= \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle \\ &= \boldsymbol{\phi}(\mathbf{x}_i^s)^\top \mathbf{R}_1 \boldsymbol{\phi}(\mathbf{x}_j^s) + \boldsymbol{\phi}(\mathbf{x}_i^\omega)^\top \mathbf{R}_2 \boldsymbol{\phi}(\mathbf{x}_j^\omega) \\ &\quad + \boldsymbol{\phi}(\mathbf{x}_i^s)^\top \mathbf{R}_3 \boldsymbol{\phi}(\mathbf{x}_j^\omega) + \boldsymbol{\phi}(\mathbf{x}_i^\omega)^\top \mathbf{R}_3 \boldsymbol{\phi}(\mathbf{x}_j^s) \end{aligned} \tag{12}$$

where $\mathbf{R}_1 = \mathbf{A}_1^\top \mathbf{A}_1 + \mathbf{A}_3^\top \mathbf{A}_3$, $\mathbf{R}_2 = \mathbf{A}_2^\top \mathbf{A}_2 + \mathbf{A}_3^\top \mathbf{A}_3$, and $\mathbf{R}_3 = \mathbf{A}_3^\top \mathbf{A}_3$ are three independent positive definite matrices. Similarly to the direct summation kernel, it can be demonstrated that (12) can be expressed as the sum of positive definite matrices, accounting for the textural, spectral, and cross-terms between textural and spectral counterparts

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &= K_s(\mathbf{x}_i^s, \mathbf{x}_j^s) + K_\omega(\mathbf{x}_i^\omega, \mathbf{x}_j^\omega) \\ &\quad + K_{s\omega}(\mathbf{x}_i^s, \mathbf{x}_j^\omega) + K_{\omega s}(\mathbf{x}_i^\omega, \mathbf{x}_j^s). \end{aligned} \tag{13}$$

The only restriction for this formulation to be valid is that $\mathbf{x}_i^s$ and $\mathbf{x}_j^\omega$ need to have the same dimension ($N_\omega = N_s$).

An intuitive example of this composite kernel would be as follows. Let the spatial features $\mathbf{x}_i^s$ be the average of the reflectance values in a given window around pixel $\mathbf{x}_i$ for each band, and let the spectral features $\mathbf{x}_i^\omega$ be the actual spectral signature ($\mathbf{x}_i = \mathbf{x}_i^\omega$). Then, $K_s(K_\omega)$ represents the distance matrix among all spatial (spectral) features, and $K_{\omega s}$ represents the similarity matrix formed by the distances among the spectra and the averaged neighborhoods.

Note that solving the minimization problem in all kinds of composite kernels requires the same number of constraints as in the conventional SVM algorithm, and thus no additional computational efforts are induced in the presented approaches.

V. EXPERIMENTAL RESULTS

A. Model Development

Experiments were carried out using the familiar AVIRIS image taken over northwest Indiana's Indian Pine test site in June 1992.[1] Following [7], we first used a part of the 145 × 145 scene, called the *subset scene*, consisting of pixels [27–94] × [31–116] for a size of 68 × 86, which contains four labeled classes (the background pixels were not considered for classification purposes). Second, we used the *whole scene*, consisting of the full 145 × 145 pixels, which contains 16 classes, ranging in size from 20–2468 pixels. We removed 20 noisy bands covering the region of water absorption, and finally worked with 200 spectral bands. In both datasets, we used 20% of the labeled samples for training and the rest for validation. In all cases, we used the polynomial kernel ($d = \{1, \ldots, 10\}$) for the spectral features according to previous results [7], [12] and used the RBF kernel ($\sigma = \{10^{-1}, \ldots, 10^3\}$) for the spatial features according to the locality assumption in the spatial domain. In the case of the weighted summation kernel, μ was varied in steps of 0.1 in the range [0,1]. For simplicity and for illustrative purposes, μ was the same for all labeled classes in our experiments. For the "stacked" ($K_{\{s,\omega\}}$) and cross-information ($K_{s\omega}$, $K_{\omega s}$) approaches, we used the polynomial kernel. The penalization factor in the SVM was tuned in the range $C = \{10^{-1}, \ldots, 10^7\}$. A *one-against-one* multiclassification scheme was adopted in both cases.

The most simple but powerful spatial features $\mathbf{x}_i^s$ that can be extracted from a given region are based on moment criteria. In this letter, we take into account the first two momenta to build the spatial kernels. Two situations were considered: 1) using the mean of the neighborhood pixels in a *window* ($dim(\mathbf{x}_i^s) = 200$) *per* spectral channel or 2) using the mean and standard deviation of the neighborhood pixels in a *window* per spectral channel ($dim(\mathbf{x}_i^s) = 400$). Inclusion of higher order momenta or cumulants did not improve the results in our case study. The *window* size was varied between 3 × 3 and 9 × 9 pixels in the training set.

B. Model Comparison

Table I shows the validation results of several classifiers for both images. We include results from six kernel classifiers: spec-

TABLE I
OVERALL ACCURACY, OA[%], AND KAPPA STATISTIC, κ, ON THE VALIDATION
SETS OF THE SUBSET AND WHOLE SCENES FOR DIFFERENT SPATIAL AND
SPECTRAL CLASSIFIERS. THE BEST SCORES FOR EACH CLASS ARE
HIGHLIGHTED IN BOLDFACE FONT. THE OA[%] THAT ARE STATISTICALLY
DIFFERENT (AT 95% CONFIDENCE LEVEL, AS TESTED THROUGH PAIRED
WILCOXON RANK SUM TEST) FROM THE BEST MODEL ARE UNDERLINED

	SUBSET SCENE		WHOLE SCENE	
	OA[%]	κ	OA[%]	κ
Spectral classifiers[†]				
Euclidean [15]	67.43	—	48.23	—
bLOOC+DAFE+ECHO [15]	93.50	—	82.91	—
K_ω [7]	95.90	—	87.30	—
K_ω developed in this paper	95.10	0.94	88.55	0.87
Spatial-spectral classifiers				
Mean				
K_s	93.44	0.92	84.55	0.82
$K_{\{s,\omega\}}$	96.84	0.97	94.21	0.93
$K_s + K_\omega$	97.12	0.97	92.61	0.91
$\mu K_s + (1-\mu)K_\omega$	97.43	0.97	**95.97**	**0.94**
$K_s + K_\omega + K_{s\omega} + K_{\omega s}$	**97.44**	**0.97**	94.80	0.94
Mean and standard deviation [‡]				
K_s	94.86	0.94	88.00	0.86
$K_{\{s,\omega\}}$	98.23	0.97	94.21	0.93
$K_s + K_\omega$	98.26	0.98	95.45	0.95
$\mu K_s + (1-\mu)K_\omega$	**98.86**	**0.98**	**96.53**	**0.96**

[†] One difference with the data and results reported in [15] is that they studied the scene using 17 classes (Soybeans-notill was split into two classes) whereas we used 16 classes. Also note that the use of the LOOC algorithm instead of the bLOOC algorithm could improve performance, as proposed in [23], [24]. Differences between the obtained accuracies reported in [7] and the presented here could be due to the random sample selection, however they are not statistically significant. [‡] Note that by using mean and standard deviation features, $N_\omega \neq N_s$ and thus no cross kernels ($K_{s\omega}$ or $K_{\omega s}$) can be constructed.

Fig. 1. Classification results in the *subset image*. (a) Labeled scene and classification maps using (b) the contextual kernel K_s (window size: 7×7), (c) spectral kernel K_ω, and (d) weighted summation kernel ($\mu K_s + (1-\mu)K_\omega$, $\mu = 0.2$, window size: 7×7).

Fig. 2. Classification results in the *whole image*. (a) Labeled scene and classification maps using (b) the contextual kernel K_s (window size: 5×5), (c) spectral kernel K_ω, and (d) weighted summation kernel ($\mu K_s + (1-\mu)K_\omega$, $\mu = 0.4$, window size: 5×5).

tral (K_ω), contextual (K_s), the stacked approach ($K_{\{s,\omega\}}$), and the three presented composite kernels. In addition, two standard methods are included for baseline comparison: bLOOC + DAFE + ECHO, which uses contextual and spectral information to classify homogeneous objects, and the Euclidean classifier [15], which only uses the spectral information. All models are compared numerically (overall accuracy, OA[%]) and statistically (kappa test and Wilcoxon rank sum test). Table I shows the results averaged over ten random realizations that were obtained to avoid skewed conclusions.

Several conclusions can be obtained from Table I. First, all kernel-based methods produce better (and statistically significant) classification results than previous methods (simple Euclidean and LOOC-based method), as previously illustrated in [7]. It is also worth noting that the contextual kernel classifier K_s alone produces good results in both images, mainly due to the presence of large homogeneous classes and the high spatial resolution of the sensor. Note that the extracted textural features $\mathbf{x}_i^s$ contain spectral information to some extent as we computed them *per* spectral channel, thus they can be regarded as contextual or local spectral features. However, the accuracy is inferior to the best spectral kernel classifiers (both K_ω implemented here and in [7]), which demonstrates the relevance of the spectral information for hyperspectral image classification. Furthermore, it is worth mentioning that all composite kernel classifiers improved the results obtained by the usual spectral kernel, which confirms the validity of the presented framework. This improvement was higher in the most difficult case of the whole scene
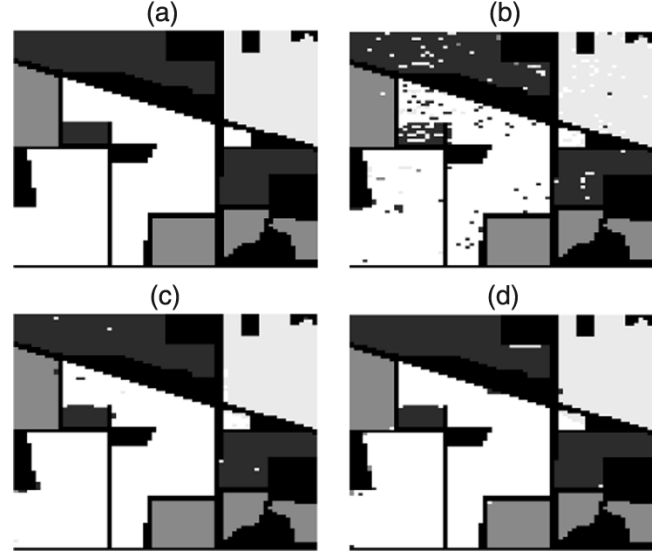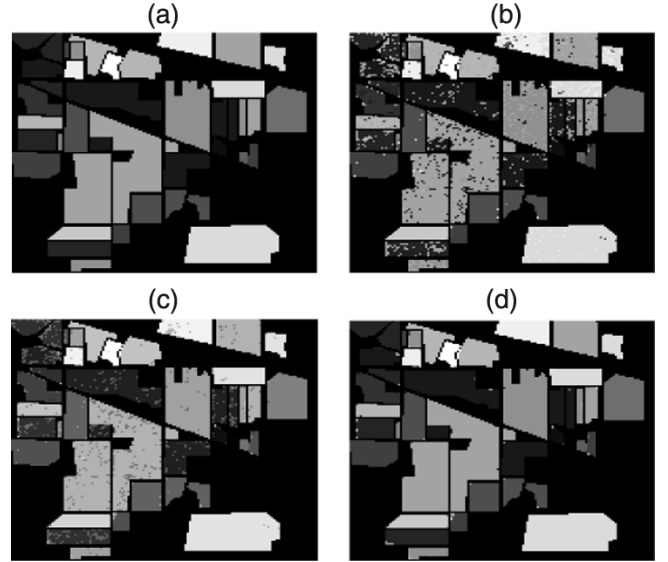
(11% increase versus 4% in the subset image) since the spatial variability of the spectral signature was reduced, and classifiers take advantage of the spatial correlation to enhance their accuracy by correctly identifying neighboring classes.

Additionally, as can be observed, there is superior performance of cross-information and weighted summation kernels with respect to the usual stacked approach. This behavior is more noticeable in the case of the whole scene and high input space dimension (using the first two momenta). The latter is a clear shortcoming of the stacked kernel approach since the risk of overfitting arises as the number of extracted features (input dimension) increases. Finally, it is also worth noting that, as the textural extraction method is refined (extracting the first two

momenta), the classification accuracy increases, which, in turn, demonstrates the robustness of kernel classifiers to high input space dimension. This property of kernel machines could be exploited to develop stacked-based classifiers that are constrained to a moderate number of extracted spatial features.

The good numerical and statistical results obtained can be assessed by showing the best classified images in Figs. 1 and 2. It is worth noting that narrow interclass boundaries are smoothed and better discerned with the inclusion of composite kernels. Finally, two relevant issues should be highlighted from the obtained results: 1) optimal μ and window size seem to act as efficient alternative tradeoff parameters to account for the textural information ($\mu = 0.2$ and 7×7 for the subset image, $\mu = 0.4$ and 5×5 for the whole image), and 2) results have been significantly improved without considering any feature selection step previous to model development. These findings should be further explored in more applications and scenarios. In conclusion, composite kernels offer excellent performance for the classification of hyperspectral images by simultaneously exploiting both the spatial and spectral information.

VI. CONCLUSION

We have presented a full framework of composite kernels for hyperspectral image classification, which efficiently combines contextual and spectral information. This approach opens a wide range of further developments in the context of Mercer's kernels for hyperspectral image classification. For instance, tuning the μ parameter as a function of prior knowledge on class distribution could be considered.

Our immediate future work is tied to the use of other kernel distances, such as the *Spectral Angle Mapper* [24], and more sophisticated texture techniques for describing the spatial structure of the classes, such as Gabor filters, Markov random fields, and cooccurrence matrices [25].

ACKNOWLEDGMENT

REFERENCES

[1] P. Swain, "Fundamentals of pattern recognition in remote sensing," in *Remote Sensing: The Quantitative Approach*. New York: McGraw-Hill, 1978, pp. 136–188.

[2] F. Melgani and S. R. Serpico, "A statistical approach to the fusion of spectral and spatio-temporal contextual information for the classification of remote-sensing images," *Pattern Recognit. Lett.*, vol. 23, pp. 1053–1061, 2002.

[3] A. Bardossy and L. Samaniego, "Fuzzy rule-based classification of remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 2, pp. 362–374, Feb. 2002.

[4] L. Bruzzone and R. Cossu, "A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 9, pp. 1984–1996, Sep. 2002.

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

[6] B. Schölkopf and A. Smola, *Learning With Kernels—Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press, 2002.

[7] J. A. Gualtieri, S. R. Chettri, R. F. Cromp, and L. F. Johnson, "Support vector machine classifiers as applied to AVIRIS data," presented at the *1999 Airborne Geoscience Workshop*, Feb. 1999.

[8] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification," *Int. J. Remote Sens.*, vol. 23, no. 4, pp. 725–749, 2002.

[9] G. Camps-Valls, L. Gómez-Chova, J. Calpe, E. Soria, J. D. Martín, L. Alonso, and J. Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1530–1542, Jul. 2004.

[10] M. Dundar and A. Langrebe, "A cost-effective semisupervised classifier approach with kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 1, pp. 264–270, Jan. 2004.

[11] 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.

[12] 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.

[13] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[14] T. Yamasaki and D. Gingras, "Image classification using spectral and spatial information based on MRF models," *IEEE Trans. Image Process.*, vol. 4, no. 9, pp. 1333–1339, Sep. 1995.

[15] S. Tadjudin and D. Landgrebe, "Classification of high dimensional data with limited training samples," Ph.D. dissertation, School of Elect. Eng. Comput.Sci., Purdue Univ., West Lafayette, IN, May 1998. Tech. Rep. TR-ECE-98-9.

[16] C. M. Bachmann, T. F. Donato, G. M. Lamela, W. J. Rhea, M. H. Bettenhausen, R. A. Fusina, K. R. du Bois, J. H. Porter, and B. R. Truitt, "Automatic classification of land cover on Smith Island, VA, using HyMAP imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2313–2330, Oct. 2002.

[17] B. Mak, J. Kwok, and S. Ho, "A study of various composite kernels for kernel eigenvoice speaker adaptation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, May 2004, pp. 325–328.

[18] J.-T. Sun, B.-Y. Zhang, Z. Chen, Y.-C. Lu, C.-Y. Shi, and W. Ma, "GE-CKO: A method to optimize composite kernels for web page classification," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intelligence*, vol. 1, Sep. 2004, pp. 299–305.

[19] G. Camps-Valls and L. Bruzzone, "Regularized methods for hyperspectral image classification," in *SPIE Int. Symp. Remote Sensing*, Gran Canaria, Spain, Sep. 2004.

[20] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with application in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, pp. 326–334, Jun. 1965.

[21] M. C. Reed and B. Simon, *Functional Analysis*, ser. Methods of Modern Mathematical Physics. New York: Academic, 1980, vol. I.

[22] Q. Jackson and D. A. Landgrebe, "An adaptive method for combined covariance estimation and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 5, pp. 1082–1087, May 2002.

[23] B.-C. Kuo and D. A. Landgrebe, "A covariance estimator for small sample size classification problems and its application to feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 4, pp. 814–819, Apr. 2002.

[24] G. Mercier and M. Lennon, "Support vector machines for hyperspectral image classification with spectral-based kernels," in *Proc. IGARSS*, vol. 1, Toulouse, France, Sep. 2003, pp. 288–290.

[25] D. Clausi and B. Yue, "Comparing co-occurrence probabilities and Markov random fields for texture analysis of SAR sea ice imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 1, pp. 215–228, Jan. 2004.