Combining Support Vector Machines and Markov Random Fields in an Integrated Framework for Contextual Image Classification

Gabriele Moser, *Member, IEEE*, and Sebastiano B. Serpico, *Fellow, IEEE*

***Abstract*—In the framework of remote-sensing image classification, support vector machines (SVMs) have lately been receiving substantial attention due to their accurate results in many applications as well as their remarkable generalization capability even with high-dimensional input data. However, SVM classifiers are intrinsically noncontextual, which represents an important limitation in image classification. In this paper, a novel and rigorous framework, which integrates SVMs and Markov random field models in a unique formulation for spatial contextual classification, is proposed. The developed contextual generalization of SVMs, is obtained by analytically relating the Markovian minimum-energy criterion to the application of an SVM in a suitably transformed space. Furthermore, as a second contribution, a novel contextual classifier is developed in the proposed general framework. Two specific algorithms, based on the Ho–Kashyap and Powell numerical procedures, are combined with this classifier to automate the estimation of its parameters. Experiments are carried out with hyperspectral, multichannel synthetic aperture radar, and multispectral high-resolution images and the behavior of the method as a function of the training-set size is assessed.**

***Index Terms*—Contextual image classification, Ho–Kashyap algorithm, Markov random fields, Powell algorithm, span bound, support vector machines (SVMs).**

I. Introduction

SUPERVISED classification plays a central role in remote sensing data analysis, for instance, for land-use or land-cover mapping, forest inventory, or urban-area monitoring [1]. Classifiers based on support vector machines (SVMs) and on the maximum-margin principle have lately been receiving substantial attention due both to their remarkable generalization capability [2] even with high-dimensional input data and to their accurate results in many applications [3], [4]. However, SVM theory has been developed by focusing on independent and identically distributed samples and class labels [2]. In terms of image classification, this focus results in an intrinsically non-contextual approach, i.e., each pixel is classified *per se*, regardless of the neighboring pixels [1]. This approach discards the information associated with the correlations among distinct pixels in the image and represents an important limitation on the use of SVM-based classifiers for image-analysis purposes.

In the present paper, this problem is addressed by proposing a novel integrated framework that combines the SVM and Markov random field (MRF) [5] approaches to classification and consequently makes it possible to introduce a rigorous contextual generalization of SVMs. MRFs are a very general family of probabilistic models that allow spatial information to be included in Bayesian image-analysis schemes in terms of the minimization of suitable "energy functions" [5], [6]. Accordingly, combining Bayesian processing schemes with MRF models is rather straightforward, but SVMs are non-Bayesian classification techniques, a property that makes the problem of integrating the SVM and MRF approaches nontrivial.

This problem can be cast in the general framework of image processing and pattern recognition but plays a particularly relevant role in remote sensing, especially in relation to the dimensionality issue and the peculiar characteristics of the spatial context in Earth observation images. Classifying such images often requires dealing with high-dimensional data [1]. Typical examples include the classification tasks that involve hyperspectral images and/or the extraction of several additional features (e.g., textural descriptors). On the contrary, dealing with high-dimensional data is usually not needed in other image-processing areas (e.g., biomedical imaging or computer vision). Data dimensionality issues critically affect the accuracy of supervised classifiers due to the Hughes phenomenon. This phenomenon causes a severe loss of accuracy when the number of training samples is insufficient to compute reliable estimates of the classifier parameters in a high-dimensional feature space [1], [7]. To avoid such problems, a feature reduction is often applied in remote sensing prior to classification [8], [9]. However, this preliminary step is nontrivial and possibly time-consuming. Support vector methods do not require preliminary feature reduction because they have proven to be robust against the Hughes effect (unless the size of the training set is very small). Therefore, they are currently acknowledged as powerful tools in addressing dimensionality issues and have attracted a strong interest in remote-sensing image classification.

Similarly, the contextual information in remote sensing images is difficult to model in terms of object shape (unlike other

Manuscript received May 13, 2011; revised January 29, 2012 and June 21, 2012; accepted July 10, 2012. Date of publication September 18, 2012; date of current version April 18, 2013. This work was supported by the Italian Space Agency (ASI) within the framework of the project entitled "OPERA—Civil protection from floods."

The authors are with the Department of Telecommunications, Electronic, Electrical, and Naval Engineering (DITEN), University of Genoa, 16145 Genova, Italy (e-mail: gabriele.moser@unige.it; sebastiano.serpico@unige.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2012.2211882

applications of image recognition) because of the presence of natural covers and of the great variability in the shapes and sizes of artificial land covers on the ground [10]. On the contrary, MRF models, which characterize local and global spatial image statistics on a per-pixel basis, are usually flexible enough to capture the contextual information associated with remote-sensing imagery. These properties explain the success of MRF models in many classification problems, including ones involving multisensor [11], multitemporal [12]–[14], or multiresolution [15] satellite images.

The above considerations motivate the usefulness of the integration of SVMs and MRFs in a unique framework to jointly exploit both the spectral and the spatial information associated with the remotely sensed image to be classified. Here, such a novel framework is developed by defining an analytical relationship between the Markovian minimum-energy decision rule and the application of an SVM in a suitably transformed space induced by a specific kernel (which will be referred to as the "Markovian kernel"). To our knowledge, this kind of formulation has not been previously proposed in the literature. Previous techniques have been developed to incorporate spatial information in kernel functions for remote sensing image classification (see, for instance, [16] or [17]). However, unlike these methods, the proposed strategy aims at fully integrating the SVM and MRF approaches in a unique framework. This formulation represents a rigorous and flexible tool to incorporate contextual information, as modeled by MRFs, into support vector classification and encompasses both kernels corresponding to finite-dimensional inner product spaces (e.g., linear or polynomial kernels) and kernels corresponding to infinite-dimensional inner product spaces (e.g., the Gaussian kernel). In the latter case, the SVM-MRF integration is obtained by formalizing the classification task in terms of suitable continuous- and discrete-time random processes and by extending methodological arguments, which were originally introduced in digital communications to design optimum receivers for communication over channels with additive white Gaussian noise (AWGN) [18].

To further investigate the potential of this approach, a novel contextual classifier is also developed that suitably formalizes the iterated conditional mode (ICM) technique in the proposed framework. ICM is an iterative MRF energy-minimization method that converges to (at least) a local minimum and usually represents a good tradeoff between classification accuracy and computational burden [5], [11], [19]. Both the kernel and regularization parameters of the SVM and the spatial smoothing parameter of the MRF are automatically optimized by extending the techniques developed in [20] and [21] and based on the Ho–Kashyap and Powell numerical algorithms, respectively. These techniques, as compared to classical grid searches (e.g., cross validation or leave-one-out [1]), usually allow similar accuracies to be reached in considerably shorter times [20], [21], thus favoring their practical applicability even to large data sets.

The main novel contributions of the paper are the following: the definition of a rigorous framework for integrating the SVM and MRF approaches to remote sensing image classification; the development of a contextual kernel-based automatic classifier in this framework; and the extension of the parameter-estimation techniques in [20] and [21] to this classifier. Experiments on diverse types of remote-sensing data, including hyperspectral, multichannel synthetic aperture radar (SAR), and multispectral high-resolution images, are presented, demonstrating the accuracy and effectiveness of the proposed approach. The behavior of the method as a function of the training-set size is also discussed.

The paper is organized as follows. Previous work on the incorporation of spatial information into SVMs is summarized in Section II. The proposed integrated framework is described in Section III, and the developed contextual classifier is presented in Section IV. Experimental results are discussed in Section V, and finally conclusions are drawn in Section VI. Analytical details and mathematical proofs concerning the proposed framework are reported in the Appendix.

II. Previous Work

Previous techniques aimed at incorporating spatial information into SVM-based classifiers for remote-sensing images were mainly based on two approaches. First, Bayesian-like processing results were derived from SVM-based discriminant functions and combined with MRF models of the contextual information [22]. Methods to derive approximated posterior probabilities from an SVM discriminant function were proposed in [23] and [24] and were used in remote sensing in [25]–[28]. In [29], SVMs were directly used as probability density estimators and were combined with MRFs to perform image segmentation. A second approach to incorporating spatial information into SVMs was based on the application of non-contextual SVMs after contextual (e.g., textural or morphological) feature extraction [30], possibly fusing spectral and spatial features through dedicated composite kernels [16], [17]. The approach of composite kernels has also been applied within graph-based [31], semi-supervised [32], multisource [33], and multitemporal [34] classification schemes. In [35], spatial information extracted from the neighborhood of each pixel is incorporated into an SVM by suitably modifying the related primal optimization problem. In [36], a context-dependent kernel is developed in the framework of object recognition. In [37], a recursive graph-kernel approach is proposed to incorporate spatial dependencies into the kernel function.

Several approaches used to extend SVMs to the case of mutually dependent samples have also been proposed in machine learning and computer vision, although they have not been applied in remote sensing. Hidden Markov models, SVMs, and Viterbi's decoding are combined in [38] to classify 1-D sequences of correlated samples and are applied to named-entity classification and speech analysis. A maximum-margin formulation is proposed in [39] to generate structured output classification results that exhibit internal dependencies (e.g., interdependent samples in a sequence or interdependent pixels in an image). This method has been applied to taxonomic text classification, named-entity recognition, sequence alignment, and natural language parsing. Support vector or, more generally, maximum-margin approaches are incorporated into Markov networks in [40]–[42] for applications in handwriting,

hypertext, document recognition, and protein analysis. The problem of the approximate training of a family of generalized SVM-based classifiers, including the combination of the maximum-margin and Markov network approaches, is addressed in [43]. In [44], conditional random fields, which represent a generalization of MRFs, are combined with SVMs in a hierarchical technique applied to computer-vision tasks (e.g., object detection). In [45], the Fisher kernel is proposed, which is expressed in terms of the log-likelihood function and may be used to characterize dependencies among data. Unlike most kernel-based methods, the application of the Fisher kernel often requires a parametric model for the data statistics [46].

III. Proposed SVM-MRF Framework

A. Preliminary Comments

Let a remote-sensing image be composed of d channels and let $\mathcal{I}$ be the related pixel lattice, $\boldsymbol{x}_i (\boldsymbol{x}_i \in \mathbb{R}^d)$ be the feature vector of the ith pixel, and $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$ containing all image samples $\boldsymbol{x}_i (i \in \mathcal{I})$. We focus on supervised binary classification, i.e., we assume that two thematic classes are present in the imaged scene, and we denote by $\mathcal{L} \subset \mathcal{I}$ the set of training pixels for such classes. The extension to the multiclass case will be presented in Section IV-D. A binary label y_i is attached to each ith pixel, such that $y_i = 1$ or $y_i = -1$ if the pixel is associated with either class $(i \in \mathcal{I})$. The true class label is known for each training sample (i.e., if $i \in \mathcal{L}$). Hence, a continuous-valued random field $\{\boldsymbol{x}_i\}_{i \in \mathcal{I}}$ of feature vectors and a discrete-valued random field $\{y_i\}_{i \in \mathcal{I}}$ of class labels are defined on the lattice $\mathcal{I}$.

If $\{f_i\}_{i \in \mathcal{I}}$ is an arbitrary field on $\mathcal{I}$ and $\mathcal{A} \subset \mathcal{I}$ is a subset of pixels, we will denote by $\boldsymbol{f}_{\mathcal{A}}$ the vector collecting all field samples f_i with $i \in \mathcal{A}$. We shall denote by $P(\cdot)$ and $p(\cdot)$ the probability mass functions (pmfs) of discrete random variables and the probability density functions (pdfs) of continuous random variables, respectively.

In Section III-B and C, we briefly review the key ideas of SVM- and MRF-based classification. Then, the proposed integrated framework is detailed, focusing first on the (analytically simpler) case of kernels associated with finite-dimensional transformed spaces (see Section III-D) and generalizing in Section III-E to kernels related to infinite-dimensional spaces.

B. Non-Contextual SVM-Based Image Classification

The SVM approach computes a linear discriminant function f between the two classes in a nonlinearly transformed space by minimizing an upper bound on the generalization error of the classifier and by adopting a kernel-based formulation [2]. Let K be a kernel function, i.e., let a real vector space $\mathcal{F}$, endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$, and a mapping $\Phi : \mathcal{X} \to \mathcal{F}$ exist such that $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle_{\mathcal{F}}$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$[2]. The SVM-based discriminant function can be equivalently expressed as a linear function in $\mathcal{F}$, i.e., $(\boldsymbol{x} \in \mathbb{R}^d)$

$$f(\boldsymbol{x}) = \langle w, \Phi(\boldsymbol{x}) \rangle_{\mathcal{F}} + b \tag{1}$$

or as a weighted sum of nonlinear kernels centered on a subset of the training samples, i.e.,

$$f(\boldsymbol{x}) = \sum_{j \in \mathcal{S}} \alpha_j y_j K(\boldsymbol{x}, \boldsymbol{x}_j) + b \tag{2}$$

where w and $b (w \in \mathcal{F}, b \in \mathbb{R})$ are the weight and bias of the linear discriminant function in $\mathcal{F}$. $\mathcal{S} \subset \mathcal{L}$ is a subset of the training samples, whose elements are named support vectors. An unknown sample $\boldsymbol{x}$ is assigned the estimated class label $\hat{y} = \operatorname{sgn} f(\boldsymbol{x})$ by the classifier. Both the selection of the samples in $\mathcal{S}$ and the computation of the set of coefficients $\{\alpha_j\}_{j \in \mathcal{S}}$ and b are obtained by solving the following quadratic programming (QP) problem[1][2]:

$$\begin{cases} \min_{\boldsymbol{\alpha} \in \mathbb{R}^\ell} \left(\frac{1}{2} \boldsymbol{\alpha}^\top Q \boldsymbol{\alpha} - \mathbf{1}^\top \boldsymbol{\alpha} \right) \\ \boldsymbol{\alpha} \in [0, C]^\ell, \quad \boldsymbol{y}_{\mathcal{L}}^\top \boldsymbol{\alpha} = 0 \end{cases} \tag{3}$$

where ℓ is the number of training samples (i.e., the cardinality of $\mathcal{L}$); $\mathbf{1}$ is an ℓ-dimensional vector with unitary components; Q is the $\ell \times \ell$ symmetric matrix whose (i, j)th entry is $Q_{ij} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) (i, j \in \mathcal{L})$; and C is a regularization parameter of the method. In particular, $\mathcal{S} = \{i \in \mathcal{L} : \alpha_i > 0\}$. Further parameters may be included in the kernel K (e.g., the variance of a Gaussian kernel or the order of a polynomial kernel). Necessary and sufficient conditions (namely, Mercer's conditions) are known for a given function K to be a kernel. For a detailed discussion, see [2].

C. Contextual MRF-Based Image Classification

MRFs represent a wide family of stochastic models for the contextual information associated with image data. They offer a powerful and computationally efficient approach to formalizing this information in terms of prior-probability distributions in Bayesian image analysis [5]. Let $\{\partial i\}_{i \in \mathcal{I}}$ be a neighborhood system on $\mathcal{I}$, where $\partial i \subset \mathcal{I}$ is the set of the neighbors of each ith pixel. For instance, a first- and a second-order neighborhood of the ith pixel consist of the 4 and 8 surrounding pixels, respectively (see Fig. 1). The label field $\{y_i\}_{i \in \mathcal{I}}$ is an MRF with respect to this neighborhood system if its joint pmf $P(\boldsymbol{y}_{\mathcal{I}})$ is strictly positive and if the following Markovianity property holds [5]:

$$P\left(y_i | \boldsymbol{y}_{\mathcal{I} - \{i\}}\right) = P(y_i | \boldsymbol{y}_{\partial i}) \qquad \forall i \in \mathcal{I}. \tag{4}$$

Let us assume that the joint pdf of $\boldsymbol{x}_{\mathcal{I}}$ given $\boldsymbol{y}_{\mathcal{I}}$ can be factorized as $p(\boldsymbol{x}_{\mathcal{I}} | \boldsymbol{y}_{\mathcal{I}}) = \prod_{i \in \mathcal{I}} p(\boldsymbol{x}_i | y_i)$, where $p(\boldsymbol{x}_i | y_i)$ is the pdf of $\boldsymbol{x}_i$ given y_i (conditional-independence hypothesis), and that $\{y_i\}_{i \in \mathcal{I}}$ is an MRF. Then, through the Hammersley–Clifford theorem, the maximization of the joint posterior distribution $P(\boldsymbol{y}_{\mathcal{I}} | \boldsymbol{x}_{\mathcal{I}})$ of all labels given all feature vectors (i.e., the Bayesian "maximum *a-posteriori*" (MAP) decision rule) is equivalent to the minimization of a global energy function defined according to the neighborhood system [5], [6]. Moreover, the posterior distribution $P(\boldsymbol{y}_{\mathcal{I}} | \boldsymbol{x}_{\mathcal{I}})$ is uniquely determined by the collection of all marginal posterior distributions

[1]All finite-dimensional vectors in the paper are assumed to be column vectors, and the superscript "$\top$" denotes the transpose operator.

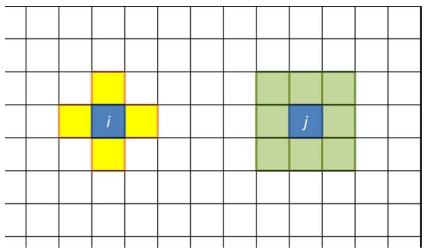

Fig. 1. Examples of neighborhoods: the black lines denote the pixel lattice, the yellow pixels constitute the first-order neighborhood of the ith pixel, and the green pixels constitute the second-order neighborhood of the jth pixel.

$P(y_i|\boldsymbol{x}_i, \boldsymbol{y}_{\partial i})(i \in \mathcal{I})$ of the label of each pixel, conditioned to its feature vector and neighboring labels [6]. This marginal distribution can be written (up to a multiplicative factor) as $\exp[-U_i(y_i|\boldsymbol{x}_i, \boldsymbol{y}_{\partial i})]$, where [5]

$$U_i(y_i|\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}) = -\ln p(\boldsymbol{x}_i|y_i) + \beta \mathcal{E}_i(y_i|\boldsymbol{y}_{\partial i}) \tag{5}$$

is a local posterior energy function, $\mathcal{E}_i$ is a local prior energy function characterizing the MRF model chosen for $\{y_i\}_{i\in\mathcal{I}}$ and β is a positive parameter. Examples of prior energies $\mathcal{E}_i$ may include the popular Potts model (or multilevel logistic model) [47] as well as more sophisticated anisotropic or discontinuity-adaptive models [5]. From a data-fusion perspective, (5) additively combines two energy contributions related to non-contextual class-conditional statistics and to contextual information, respectively. The parameter β tunes the reciprocal weights of these two terms and consequently affects the spatial smoothing properties of the MRF [11].

For a detailed discussion of MRFs, see [5]. Here, we only recall that many widely used techniques for the minimization of the aforementioned global energy (i.e., for MAP classification) can be expressed in terms of the following energy-difference function $(i \in \mathcal{I})$:

$$\Delta U_i(\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}) = U_i(-1|\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}) - U_i(1|\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}). \tag{6}$$

For instance, this property holds for both the simulated-annealing (SA) global energy-minimization method (because the related Gibbs or Metropolis sampler can be formalized in terms of ΔU_i) [6] and for the ICM [19], modified Metropolis dynamics [48], and highest-confidence-first [5] local-energy-minimization techniques. Also, the maximizer of posterior marginals method, which does not use the MAP criterion but follows a segmentation-error cost function [49], can be formalized in terms of ΔU_i because, like SA, it is based on Gibbs or Metropolis sampling.

D. Finite-Dimensional Case

To integrate the MRF and SVM approaches, we shall prove that, under proper assumptions, the energy-difference function in (6) can be expressed as an SVM kernel expansion associated with a suitable kernel. With the same notations as in Section III-B and C, we accept the following assumptions.

Assumption 1: $\{y_i\}_{i\in\mathcal{I}}$ is an MRF with respect to the neighborhood system $\{\partial i\}_{i\in\mathcal{I}}$. $\mathcal{E}_i$ and β denote the related prior local energy and smoothing parameter, respectively. The conditional-independence hypothesis holds with regard to the pdf of $\{\boldsymbol{x}_i\}_{i\in\mathcal{I}}$ given $\{y_i\}_{i\in\mathcal{I}}$.

Assumption 2: The dimension $\dim \mathcal{F} = D$ of the transformed space $\mathcal{F}$ induced by the kernel K is finite.

By invoking Assumption 2, the inner product space $\mathcal{F}$ can be identified with the $\mathbb{R}^D$ space, endowed with the usual dot product. More precisely, under Assumption 2, $\mathcal{F}$ and $\mathbb{R}^D$ are isometrically isomorphic,[2] i.e., a bijective mapping $\boldsymbol{A} : \mathcal{F} \to \mathbb{R}^D$ exists, such that $\langle u, v\rangle_{\mathcal{F}} = \boldsymbol{A}(u)^\top \boldsymbol{A}(v)$ for all $u, v \in \mathcal{F}$.

Assumption 3: For each ith pixel, the pdf of the D-dimensional transformed vector $\tilde{\boldsymbol{x}}_i = \boldsymbol{A}[\Phi(\boldsymbol{x}_i)]$, conditioned to either $y_i = 1$ or $y_i = -1$, is a Gaussian with a mean $\boldsymbol{m}^+$ or $\boldsymbol{m}^-$, respectively, with the same covariance matrix Σ for both classes $(i \in \mathcal{I})$.

As mentioned above, the finite-dimension condition is accepted here as a working hypothesis, but it will be removed in Section III-E, so it does not represent a restriction. The Gaussianity and equal-covariance statements are analytical restrictions. However, as discussed in [52], these restrictions are expected to be relatively mild, because they are formulated in the transformed space $\mathbb{R}^D$, which may be very high dimensional and is obtained through a potentially very general and flexible nonlinear mapping [2], [52]. In particular, due to this nonlinearity, the assumption that the transformed vector $\tilde{\boldsymbol{x}}_i$ has the same covariance matrix for both classes does not imply that the original feature vector $\boldsymbol{x}_i$ has the same covariance matrix for both classes.

Note that $\boldsymbol{m}^+$, $\boldsymbol{m}^-$, and Σ do not depend on the location i, which means that a stationary behavior is implicitly assumed for the pixelwise class-conditional statistics.[3] Both this stationarity and the conditional-independence hypothesis (which is widely accepted when using MRFs for classification) are stated to favor analytical tractability. However, as reflected in the possible dependence of $\mathcal{E}_i$ on i, no stationarity is involved with regard to the MRF model for the label field. This generality allows an arbitrary MRF to be plugged into the proposed framework.

Theorem 1: Under Assumptions 1, 2, and 3, the energy-difference function associated with the application of the Markovian minimum-energy rule in the transformed space $\mathcal{F}$ can be expressed as a kernel expansion, i.e.,

$$\Delta U_i(\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}) = \sum_{j\in\mathcal{S}} \alpha_j y_j K_{\mathsf{MRF}}(\boldsymbol{x}_i, \varepsilon_i; \boldsymbol{x}_j, \varepsilon_j) + b \tag{7}$$

where the kernel K_{MRF} and the additional feature ε_i are defined as follows:

$$\varepsilon_i = \mathcal{E}_i(-1|\boldsymbol{y}_{\partial i}) - \mathcal{E}_i(1|\boldsymbol{y}_{\partial i}) \qquad \forall i \in \mathcal{I} \tag{8}$$

$$K_{\mathsf{MRF}}(\boldsymbol{x}, \varepsilon; \boldsymbol{x}', \varepsilon') = K(\boldsymbol{x}, \boldsymbol{x}') + \beta\varepsilon\varepsilon' \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}, \ \forall \varepsilon, \varepsilon' \in \mathbb{R}. \tag{9}$$

[2]We recall that two inner product spaces are isometrically isomorphic when a bijective isometry exists between them, i.e., when they can be put in one-to-one correspondence with each other by a mapping that also preserves inner products and consequently norms and distances [50], [51].

[3]This means that the distribution of $\tilde{\boldsymbol{x}}_i$, conditioned to y_i, is the same for all pixels $i \in \mathcal{I}$ [1], [5].

Furthermore, $\mathcal{S}, \{\alpha_j\}_{j\in\mathcal{S}}$ and b can be obtained by solving the QP problem in (3) with the matrix Q redefined as $Q_{ij} = y_i y_j K_{\mathsf{MRF}}(\boldsymbol{x}_i, \varepsilon_i, \boldsymbol{x}_j, \varepsilon_j)(i, j \in \mathcal{L})$.

The proof is reported in Appendix I. Therefore, the minimum-energy rule, formulated in terms of the energy-difference function (e.g., by SA or ICM) in the transformed space induced by the kernel K, is equivalent to a kernel expansion. The original feature vector $\boldsymbol{x}_i$, an additional feature ε_i, and a further kernel K_{MRF}, which fuses the two types of features, are involved in this expansion. Moreover, ε_i is defined through the prior energy $\mathcal{E}_i$ of the MRF model and is consequently related to contextual information. K_{MRF} is a linear combination of two contributions related to pixelwise spectral information (through K) and to the adopted contextual MRF model (through ε_i) and will be referred to as the Markovian kernel; β acts as a smoothing parameter that tunes the tradeoff between the two terms. In particular, the kernel expansion in (7) is formally identical to the basic SVM formulation in (2) up to the inclusion of the Markovian feature ε_i and the replacement of K by K_{MRF}. This identity allows contextual information modeled by an MRF to be incorporated into support vector classification, thus rigorously combining the MRF and SVM approaches for classification into a unique framework.

Note also that deriving $\mathcal{S}$ and the α_i and b coefficients of (7) through a QP problem prevents the need to estimate $\boldsymbol{m}^+$, $\boldsymbol{m}^-$, and Σ. This estimation would be a critical (if not unfeasible) task because it would explicitly involve transformed samples in the D-dimensional space $\mathcal{F}$, where D might be extremely large.

We recall that a kernel formally similar to K_{MRF} (namely, the "direct-summation kernel") was introduced in [16] as the sum of two separate kernels whose input vectors are composed of spectral channels and spatial features, respectively. However, this previous approach is based on the general inner space properties associated with kernel functions and is not related to a specific model for the spatial context. Here, we prove that an SVM with the Markovian kernel K_{MRF}, whose inputs include a spatial feature that is explicitly related with an MRF, is equivalent to a Markovian minimum-energy classification in a transformed space.

E. Infinite-Dimensional Case

In this section, we prove that the proposed SVM-MRF integration is also valid when $\dim\mathcal{F} = +\infty$. This is accomplished by removing Assumptions 2 and 3, accepting Assumptions 4 and 5 (see below), and retaining Assumption 1. Such a generalization is not straightforward because the discussion presented in the previous section and the proof of Theorem 1 (see Appendix I) make use of linear-algebra arguments that cannot be directly extended to infinite-dimensional vector spaces. Indeed, this generalization is needed to make the proposed framework practically useful, as many widely accepted kernels (e.g., the Gaussian kernel) induce infinite-dimensional transformed spaces [2].

Assumption 4: The dimension of the transformed space $\mathcal{F}$ induced by the kernel K is infinite.

In the previous section, Assumption 2 allowed $\mathcal{F}$ to be identified with $\mathbb{R}^D$. Here, analogously, Assumption 4 allows $\mathcal{F}$ to be identified with the inner product space $L^2[0,1]$ of the Lebesgue square-integrable real-valued signals defined on the interval [0, 1]. More precisely, there exists a bijective mapping $A : \mathcal{F} \to L^2[0,1]$ such that $\langle u, v\rangle_{\mathcal{F}} = \langle A(u), A(v)\rangle_{L^2}$ for all $u, v \in \mathcal{F}$, where $\langle\cdot,\cdot\rangle_{L^2}$ is the usual inner product in $L^2[0,1]$ [2], [50]. Details can be found in Appendix II.

Therefore, in the present infinite-dimensional setting, $\tilde{x}_i = A[\Phi(\boldsymbol{x}_i)](i \in \mathcal{I})$ takes on values in $L^2[0,1]$, i.e., it can be thought of as a continuous-time random process $\tilde{x}_i(t)(0 \le t \le 1)$ whose realizations are square-integrable signals. This auxiliary random signal $\tilde{x}_i(t)$ and the "time" variable t are introduced to formalize the point associated with each ith pixel through the mapping to an infinite-dimensional space induced by K. They will play the role of analytical tools in the proofs of the theorems reported below.

Assumption 5: For each ith pixel, $\tilde{x}_i(t)$, conditioned to either $y_i = 1$ or $y_i = -1$, is a Gaussian random process with mean $m^+(t) = E\{\tilde{x}_i(t)|y_i = 1\}$ or $m^-(t) = E\{\tilde{x}_i(t)|y_i = -1\}$, respectively, and the same autocovariance function for both classes. Furthermore, this autocovariance function is continuous and positive definite.

Assumption 5 plays the same role in the infinite-dimensional setting as Assumption 3 in the previous finite-dimensional case. Hence, the same comments reported in Section III-D on the possible restrictive role of Assumption 3 also hold for Assumption 5. The key result under the stated hypotheses is that $\mathcal{F}$ can be identified not only with $L^2[0,1]$ but also with a further inner product space $\mathcal{H}$, in which a very convenient class-conditional probabilistic model holds. This result is summarized by the following proposition.

Proposition 1: Under Assumptions 4 and 5, a further bijective isometry $B : \mathcal{F} \to \mathcal{H}$ exists from $\mathcal{F}$ to a real vector space $\mathcal{H}$, endowed with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, such that

- The elements of $\mathcal{H}$ are real-valued sequences.
- Consequently, for each ith pixel, $\bar{x}_i = B[\Phi(\boldsymbol{x}_i)]$ is a random sequence whose realizations are in $\mathcal{H}(i \in \mathcal{I})$.
- This sequence $\bar{x}_i(n)$ can be decomposed as follows:

$$\bar{x}_i(n) = \begin{cases} \mu^+(n) + \nu(n) & \text{for } y_i = 1 \\ \mu^-(n) + \nu(n) & \text{for } y_i = -1 \end{cases} \tag{10}$$

  where $\nu(n)$ is a discrete-time white Gaussian process independent from y_i, with zero mean and unitary variance, and $\mu^+(n)$ and $\mu^-(n)$ are the means of $\bar{x}_i(n)$ when conditioned to the two classes $(n = 0, 1, 2, \ldots; i \in \mathcal{I})$.

The proof is based on the Karhunen–Loeve transformation [53] and is reported in Appendix II. Hence, when $\dim\mathcal{F} = +\infty$, each pixel can be associated with a discrete-time random process that can be expressed as the related class-conditional mean with AWGN. Therefore, the theory of Bayesian optimum receiver design under AWGN conditions can be generalized to the present problem. As discussed in [18], this theory allows for the identification of a "relevant" subspace of the considered space of random sequences such that 1) a projection onto this subspace yields no loss of information with respect to Bayesian decisions and 2) this subspace is always one- or 2-D [18]. We exploit this idea to bring the current infinite-dimensional setting

back to the previous finite-dimensional one and consequently to prove the following theorem.

Theorem 2: Under Assumptions 1, 4, and 5, the same conclusions as in Theorem 1 hold.

The proof is detailed in Appendix III. Theorems 1 and 2 jointly make the proposed integrated SVM-MRF framework very general and feasible for an arbitrary kernel function, associated with either a finite- or an infinite-dimensional transformed inner product space.

IV. Contextual Classifier in the Proposed Framework

A. Proposed Classifier: The Binary Case

We also propose here a contextual classifier that formalizes ICM in the described SVM-MRF framework. Here, ICM has been chosen as a conceptually simple energy-minimization approach to demonstrate the power of the proposed framework through a practical contextual classifier. However, any other energy-minimization technique, which can be expressed in terms of the energy-difference function of (6), could be formulated in the proposed framework (see Section III-C). As usual in support vector classification, we focus first on the case of binary classification and generalize subsequently to the multiclass case (see Section IV-D).

According to the ICM approach, the proposed method is iterative and is initialized with a preliminary classification map, generated here by a non-contextual SVM. To make the technique automatic, two dedicated algorithms are integrated into the initialization phase to optimize both the SVM kernel and regularization parameters and the MRF smoothing parameter β. We collect all SVM parameters in a q-dimensional vector $\boldsymbol{\theta}$ (including C and the possible kernel parameters), and we stress the dependence on β and $\boldsymbol{\theta}$ by explicitly denoting as $K(\cdot|\boldsymbol{\theta}), K_{\mathsf{MRF}}(\cdot|\beta,\boldsymbol{\theta})$ and $\Delta U_i(\cdot|\beta,\boldsymbol{\theta})$ the original (non-contextual) kernel, the Markovian kernel, and the energy-difference function. With the quantities computed at the rth iteration marked by a superscript $r (r = 0, 1, 2, \ldots)$, the method performs the following processing steps.

- Initialization phase—Set $r = 0$ and perform the following operations.
  a) Compute an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ by the method described in Section IV-B.
  b) Generate an initial classification map $\{\hat{y}_i^0\}_{i\in\mathcal{I}}$ by running a non-contextual SVM with the original kernel $K(\cdot|\hat{\boldsymbol{\theta}})$ and the estimate $\hat{\boldsymbol{\theta}}$ obtained in step a).
  c) Compute an estimate $\hat{\beta}$ of β by applying the method described in Section IV-C to the initial classification obtained in step b).
- Iterative phase—Iterate the following steps until convergence.
  1) Compute the updated contextual feature ε_i^r of each ith pixel $(i \in \mathcal{I})$ by applying (8) to the current classification map $\{\hat{y}_i^r\}_{i\in\mathcal{I}}$.
  2) Compute the energy-difference function $\Delta U_i^r(\cdot|\hat{\beta},\hat{\boldsymbol{\theta}})$ in (7) by training an SVM (i.e., by solving the related QP problem) with the kernel $K_{\mathsf{MRF}}(\cdot|\hat{\beta},\hat{\boldsymbol{\theta}})$, the original features, the contextual feature updated in step 1, and the estimates $\hat{\beta}$ and $\hat{\boldsymbol{\theta}}$ resulting from the initialization phase.
  3) Update the label of each pixel by the minimum-energy rule, i.e., set $\hat{y}_i^{r+1} = 1$ if $\Delta U_i^r(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_{\partial i}^r|\hat{\beta},\hat{\boldsymbol{\theta}}) > 0$ and $\hat{y}_i^{r+1} = -1$ otherwise $(i \in \mathcal{I})$.
  4) If the percentage of pixels in which the current and the updated label configurations (i.e., $\{\hat{y}_i^r\}_{i\in\mathcal{I}}$ and $\{\hat{y}_i^{r+1}\}_{i\in\mathcal{I}}$) differ is larger than a predefined threshold (here, equal to 0.01%), go back to step 1; otherwise stop.

Convergence to a local-energy minimum is ensured for the ICM approach. Note that there is no restriction on the prior energy $\mathcal{E}_i$, i.e., the MRF model used in the proposed classifier can be chosen arbitrarily. The contextual feature ε_i is updated at each iteration of the algorithm according to the current classification map. Because K_{MRF} depends on this feature, the corresponding kernel matrix (i.e., the matrix Q defined in Theorem 1) also needs to be recomputed at each iteration.

B. Automatic Estimation of SVM Parameters

To automatically optimize $\boldsymbol{\theta}$, the algorithm proposed in [21] for support vector regression is extended to classification. This algorithm uses the Powell procedure to numerically maximize an analytical upper bound on the leave-one-out error. The bound can be defined both for SVM regression [54] and for SVM classification [55]; therefore, the algorithm in [21] can be generalized to the proposed classifier. A traditional grid-search procedure, based, for instance, on k-fold cross validation, could also be used to optimize $\boldsymbol{\theta}$. However, such a procedure is generally time-consuming because it repeats the training process k times for each node in the grid. Its result is also affected by the user-defined trial-and-error choices of the grid structure and size. However, the strategy adopted here neither requires user intervention nor relies on the user's expertize and, as discussed in detail in [21] with regard to the case of regression, usually allows accuracies comparable to those of cross validation to be obtained in much shorter times.

We denote explicitly as $\boldsymbol{\alpha}(\boldsymbol{\theta}), b(\boldsymbol{\theta}), f(\cdot|\boldsymbol{\theta})$, and $\mathcal{S}(\boldsymbol{\theta})$ the solution of the QP problem in (3), the related bias coefficient, the discriminant function in (2), and the set of support vectors obtained when the parameter vector $\boldsymbol{\theta}$ is used. We also suppose that $\boldsymbol{\theta}$ can take values in the whole space $\mathbb{R}^q$. As in [21], this assumption is aimed at formalizing parameter estimation as an unconstrained optimization. For instance, when a Gaussian kernel with variance σ^2 is used, the assumption is met by setting $\boldsymbol{\theta} = (\ln C, \ln \sigma)$ so that no positivity constraint on C and σ needs to be explicitly formulated.

As proven in [55] under mild assumptions, the leave-one-out error rate of a non-contextual SVM classifier is upper bounded by the fraction $J(\boldsymbol{\theta})$ of the training samples $i \in \mathcal{L}$ that are support vectors (i.e., $i \in \mathcal{S}(\boldsymbol{\theta})$) and satisfy the following inequality:

$$\alpha_i(\boldsymbol{\theta}) S_i^2(\boldsymbol{\theta}) \geq y_i f(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{11}$$

where $S_i^2(\boldsymbol{\theta})$ is a real coefficient, named "span" and associated with the ith support vector. $J(\boldsymbol{\theta})$ is called the "span bound," and the span of each support vector can be computed either by a further QP problem [55], [56] or by a fast procedure based on linear-algebra arguments [57]. Computing the bound involves only training samples, with no need for additional validation sets or cross-validation routines.

The span bound is known to be very tight on the leave-one-out error rate and has also been experimentally found to be remarkably well correlated with hold-out (i.e., test-set) errors [56], provided that the training and test samples are identically distributed. This observation suggests that minimizing $J(\cdot)$ can be an effective strategy to search for parameter values that favor an accurate SVM solution, a conclusion also confirmed in [21] in the case of regression. However, $J(\cdot)$ is nondifferentiable [56], which prevents the application of numerical minimization techniques based on gradients or Hessian matrices. In [56], this drawback was overcome by introducing a regularized differentiable version of the bound. However, this approach required the introduction of an additional regularization parameter that had to be manually set.

Here, we extend to the proposed classifier the approach adopted in [21] for regression, which numerically minimizes $J(\cdot)$ through the Powell algorithm [58]. This algorithm does not involve derivatives and consequently avoids the need for additional regularization parameters. It performs a sequence of line searches along a set of search directions in the q-dimensional parameter space and iteratively updates the search directions so that they emulate, albeit without derivatives, the behavior of the conjugate-gradient technique [58]. Convergence is analytically ensured, at least, to a local minimum of the input functional. For further details on the Powell algorithm, we refer to [21] and [58]. While a cross-validation grid search aims at exhaustively exploring a predefined and suitably discretized portion of the parameter space, the proposed approach performs a sequence of moves in this space. This sequence is guided by an error bound and helps in approaching a locally optimum solution in a considerably shorter time. The proposed parameter-estimation algorithm will be referred to as the "Powell-span-bound" algorithm in the following.

C. Automatic Estimation of the MRF Smoothing Parameter

To estimate β, the technique developed in [20] for MRF models whose energy functions linearly depend on the unknown parameters is extended to the SVM-MRF formulation. Here, such a technique is feasible due to the linear relationship between β and the energy-difference function [see (7)]. Given the initial map $\{\hat{y}_i^0\}_{i\in\mathcal{I}}$, the method formalizes a condition of correct classification of the training set as a system of linear inequalities and numerically solves this system by the Ho–Kashyap algorithm [20], [59]. Given the estimate $\hat{\boldsymbol{\theta}}$ provided by the Powell-span-bound algorithm, the training set is correctly classified at the first ICM iteration if and only if

$$y_i \Delta U_i^0 \left(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_{\partial i}^0 | \beta, \hat{\boldsymbol{\theta}}\right) \geq 0 \qquad \forall i \in \mathcal{L}. \tag{12}$$

Note that, in these inequalities, the label y_i of each training pixel $i \in \mathcal{L}$ is known. Simple algebraic manipulations allow this condition to be equivalently stated as

$$E\boldsymbol{\zeta} \geq \mathbf{0} \tag{13}$$

where $\boldsymbol{\zeta} = (\zeta_1, \zeta_2)^\top$ is a 2-D vector such that $\beta = \zeta_2/\zeta_1$ and $\zeta_1 > 0$, $\mathbf{0}$ is an ℓ-dimensional vector with zero components, and E is an $\ell \times 2$ matrix whose (i,h)th entry E_{ih} is ($i \in \mathcal{L}$; $h = 1, 2$)

$$\begin{aligned} E_{i1} &= \sum_{j\in\mathcal{S}(\hat{\boldsymbol{\theta}})} \alpha_j(\hat{\boldsymbol{\theta}}) y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j | \hat{\boldsymbol{\theta}}) + y_i b(\hat{\boldsymbol{\theta}}) \\ E_{i2} &= \sum_{j\in\mathcal{S}(\hat{\boldsymbol{\theta}})} \alpha_j(\hat{\boldsymbol{\theta}}) y_i y_j \varepsilon_i^0 \varepsilon_j^0. \end{aligned} \tag{14}$$

According to [20], we formulate the estimation of β as the numerical solution of the linear system of inequalities in (13) to identify a vector $\boldsymbol{\zeta}$ that favors correct classification of the training samples. The solution of the system is numerically addressed by the Ho–Kashyap procedure, which is based on a mean-square criterion and is well known in pattern recognition as a linear binary-classification tool [59]. The Ho–Kashyap technique is iterative and is proven to exhibit good convergence properties. In particular, convergence in a finite number of iterations to a solution of the system is ensured, when such a solution exists [59]. Further details on the Ho–Kashyap technique can be found in [20] and [59].

This estimation algorithm for β involves an input classification map. In principle, the algorithm could be separately run at each ICM iteration to update $\hat{\beta}$ according to the current map. Here, as in [20], we choose to run the algorithm only once in the initialization phase to minimize the overall computational burden. Therefore, the estimate $\hat{\beta}$ is fixed during all subsequent ICM iterations. Past experiments, mentioned in [20] with regard to several MRF models, suggested that the difference in accuracy between running the method only in the initialization and running it at each ICM iteration was very small and not worth the corresponding increase in execution time.

Note that many MRF models widely used in remote sensing involve one smoothing parameter, as in the described formulation [6], [11], [47]. However, in general, there might be further parameters inside the function $\mathcal{E}_i$ (e.g., to incorporate anisotropy information into the MRF model [5]). Here, we do not address the problem of estimating them, and we implicitly assume that they have been properly set using other techniques prior to the application of the proposed classifier.

D. Extension to Multiclass Classification

As usual in support vector classification, when M classes $\omega_1, \omega_2, \ldots, \omega_M$ ($M > 2$) are present in the imaged scene, the proposed classifier is generalized by decomposing the multiclass problem into a collection of binary subproblems. Here, the "one-against-one" approach is used because, compared to other feasible choices such as "one-against-all," it usually favors high accuracy and minimizes possible issues due to unbalanced classes [2]. Accordingly, for each pair of distinct classes

$(\omega_h, \omega_k)(h = 1, 2, \ldots, M-1; k = h+1, h+2, \ldots, M)$, the following operations are performed: 1) a binary contextual classifier is separately trained by running the procedure in Section IV-A until convergence; 2) the trained binary classifier is used to classify the whole image; and 3) for each ith pixel $(i \in \mathcal{I})$, a vote is cast for either ω_h or ω_k according to the class label assigned to the pixel in step 2). Then, a final M-class classification map is derived by the majority-vote rule.

To extend the Powell-span-bound algorithm, the span bound $J_{hk}(\boldsymbol{\theta})$ associated with each pair $(\omega_h, \omega_k)(h = 1, 2, \ldots, M-1; k = h+1, h+2, \ldots, M)$ is computed during the related initialization phase, and the Powell procedure is used to minimize the following average span-bound:

$$J(\boldsymbol{\theta}) = \sum_{h=1}^{M-1} \sum_{k=h+1}^{M} \hat{P}_h \hat{P}_k J_{hk}(\boldsymbol{\theta}) \tag{15}$$

where $\hat{P}_h$ is an estimate of the prior probability of ω_h, computed as the relative frequency of its training samples in the training set $(h = 1, 2, \ldots, M)$. With regard to the Ho–Kashyap-based algorithm, we stack together, in a unique linear system, all the inequalities resulting from separately applying the procedure in Section IV-C with each pair $(\omega_h, \omega_k)(h = 1, 2, \ldots, M-1; k = h+1, h+2, \ldots, M)$. Then, the Ho–Kashyap algorithm is applied to numerically solve the resulting linear system.

The proposed SVM-MRF classification method will be denoted as the "Markovian support vector classifier" (MSVC) in the following sections.

V. Experimental Results

A. Data Sets and Experimental Setup

Thanks to its nonparametric formulation, the proposed approach is feasible for the supervised classification of arbitrary types of remote-sensing images. Experiments were carried out with three types of satellite data to investigate the applicability of MSVC. The first data set was the well-known AVIRIS hyperspectral image [145 × 145 pixels; 202 channels, after discarding the bands affected by atmosphere absorption; Fig. 2(a)] acquired over NW Indian Pine in 1992 and presenting nine main classes, mostly related to vegetated land covers. The numbers of training and test samples available per class are listed in Table I. Hereafter, this data set will simply be called "Indian Pine." The second data set (named "Itaipu") was composed of a three-channel 4-m resolution IKONOS image [1999 × 1501 pixels; Fig. 3(a)] acquired over Itaipu (Brazil/Paraguay border). It presented seven classes, including vegetated and urban/built-up land covers (Table II). The third data set (named "Pavia") was composed of a multipolarization and multifrequency SIR-C/XSAR image (700 × 280 pixels) acquired near Pavia (Italy). It consisted of one X-band VV-polarized SAR amplitude [Fig. 4(a)] and three C-band amplitude channels (VV and VH polarizations and a total power channel) and included two main classes (i.e., "dry soil" and "wet soil"). Additional details on "Pavia" and "Indian Pine" can be found in [20] and [9], respectively. For all data sets, training and test fields were chosen to be spatially disjoint to minimize the correlation between training and test samples and the resulting possible bias in accuracy assessment. The total training- and test-set sizes were 3726 and 3728 for "Indian Pine," 9999 and 174 789 for "Itaipu," and 7551 and 7009 for "Pavia."

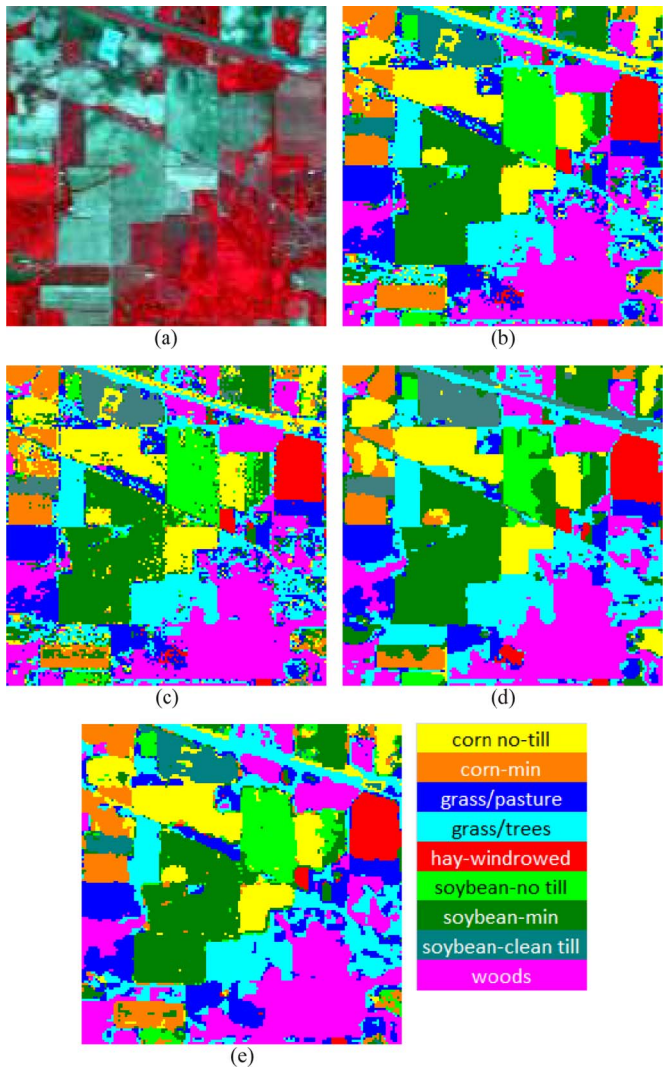

Fig. 2. "Indian Pine": (a) false-color composition of the AVIRIS channels with center wavelengths 841.04, 646.72, and 557.49 nm and classification maps generated by (b) MSVC, (c) non-contextual SVM, (d) MRF-Gauss, and (e) the graph-kernel method in [37].

A Gaussian radial basis function kernel [2] and a Potts MRF model [5] were used for K and $\mathcal{E}_i$, respectively, i.e., $(\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d; i \in \mathcal{I})$

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right)$$
$$\mathcal{E}_i(y_i|\boldsymbol{y}_{\partial i}) = -\sum_{j \in \partial i} \delta(y_i, y_j) \tag{16}$$

where $\sigma(\sigma > 0)$ is the standard deviation of the kernel (hence, $\boldsymbol{\theta} = (\ln C, \ln \sigma)$), and δ is the usual Kronecker symbol (i.e., $\delta(a, b) = 1$ for $a = b$ and $\delta(a, b) = 0$ otherwise).

The results of MSVC were compared with those given by the following: 1) a traditional non-contextual SVM (equivalent to the initialization of MSVC); 2) a classical MRF-based

TABLE I

"INDIAN PINE": TRAINING- AND TEST-SAMPLE SIZES; CLASSIFICATION ACCURACIES (ON THE TEST SET) AND COMPUTATION TIMES OF MSVC, A NON-CONTEXTUAL SVM, MRF-GAUSS, MRF-SVM-POST, AND THE COMPOSITE-KERNEL AND GRAPH-KERNEL METHODS IN [16] AND [37]; Z STATISTICS OF THE MCNEMAR'S TEST APPLIED TO VALIDATE WHETHER THE DIFFERENCE BETWEEN THE ACCURACY OF MSVC AND THE ACCURACY OF EACH OF THE OTHER ALGORITHMS IS STATISTICALLY SIGNIFICANT. MRF-GAUSS WAS APPLIED AFTER FEATURE REDUCTION BY SFBE. ALL OTHER METHODS WERE APPLIED TO THE ORIGINAL 202 HYPERSPECTRAL CHANNELS. FOR EACH METHOD, BOTH THE TIME TAKEN BY THE WHOLE CLASSIFICATION PROCEDURE (INCLUDING PARAMETER OPTIMIZATION, TRAINING, AND APPLICATION OF THE CLASSIFIER TO THE INPUT IMAGE) AND THE TIME WITHOUT PARAMETER OPTIMIZATION (I.E., THE TIME TAKEN ONLY TO TRAIN THE CLASSIFIER AND TO APPLY IT TO THE INPUT IMAGE ONCE THE PARAMETER VALUES ARE GIVEN) ARE REPORTED

class	training samples	test samples	MSVC	SVM	MRF-Gauss	MRF-SVM-Post	Composite kernel	Graph kernel
corn no-till	762	575	89.91%	80.17%	90.78%	86.09%	82.43%	86.09%
corn-min	435	326	95.40%	73.93%	69.33%	93.25%	76.99%	83.74%
grass/pasture	232	225	92.00%	90.67%	90.22%	94.22%	91.11%	90.67%
grass/trees	394	283	98.59%	98.59%	100%	100%	97.53%	97.53%
hay-windrowed	235	227	100%	100%	100%	100%	100%	100%
soybean-no till	470	443	92.10%	79.23%	44.24%	95.71%	90.29%	88.71%
soybean-min	142	936	90.71%	83.65%	95.30%	80.24%	88.25%	86.00%
soybean-clean till	328	226	82.30%	77.88%	86.28%	89.38%	66.37%	73.01%
woods	728	487	97.95%	97.54%	97.95%	96.71%	97.74%	95.48%
		overall accuracy	92.84%	85.76%	86.40%	90.37%	88.12%	88.60%
		average accuracy	93.22%	86.85%	86.01%	92.84%	87.86%	89.03%
		Z statistics (McNemar's test)	—	−14.58	−11.05	−5.99	−9.65	−9.48
time including parameter optimization			17 min	13 min	6 s	42 min	5 h 20 min	3 h 44 min
time without parameter optimization			1 min 58 s	25 s	1 s	27 s	36 s	50 s

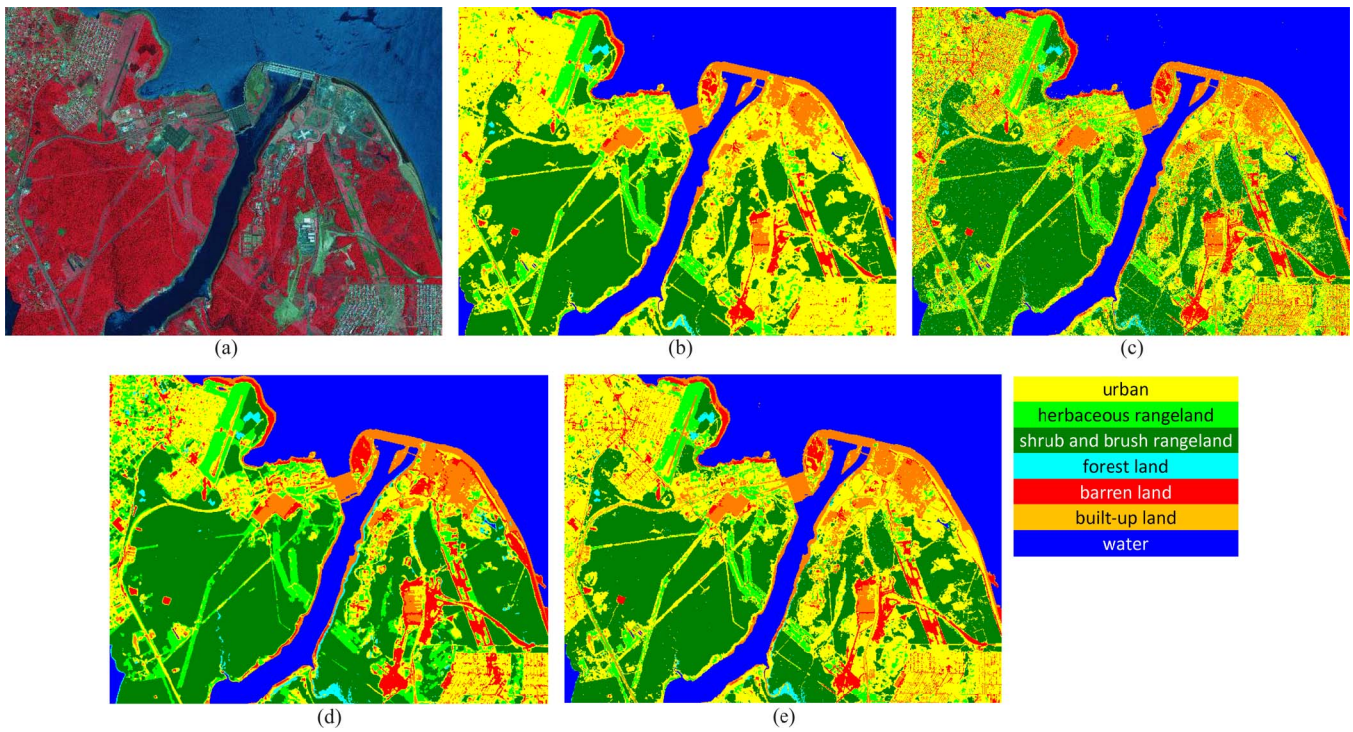

Fig. 3. "Itaipu": (a) RGB false-color composition of the near-infrared, red, and green channels and classification maps generated by (b) MSVC, (c) non-contextual SVM, (d) MRF-SVM-Post, and (e) the graph-kernel method in [37].

classifier, in which the Potts model was used and the pixelwise class posterior probabilities [11] were computed according to Gaussian pdfs (MRF-Gauss) for "Itaipu" and "Indian Pine" and by the "k-nearest neighbor" algorithm [1] (MRF-k-NN) for "Pavia"; 3) a previous approach to SVM-MRF combination (MRF-SVM-Post), consisting of an MRF-based classifier, in which the Potts model was used and the pixelwise class posterior probabilities were approximated by the method described in [24]; and 4) the two previous non-Markovian contextual extensions of SVMs proposed in [16] and [37] and based on composite and graph kernels, respectively.

The MRF-SVM-Post method applies Platt's algorithm [23] and pairwise class-coupling [24] to the discriminant function of the non-contextual SVM to derive posterior-probability approximations for all considered classes. The multivariate Gaussian model is often accepted for class-conditional statistics in multispectral data, making the application of MRF-Gauss to "Itaipu" and "Indian Pine" appropriate. On the other hand, class-conditional statistics in multichannel SAR are strongly non-Gaussian and no accurate parametric model is currently available for the joint statistics of four SAR amplitude or intensity channels, so the nonparametric MRF-k-NN approach was

TABLE II
"ITAIPU": TRAINING- AND TEST-SAMPLE SIZES; CLASSIFICATION ACCURACIES (ON THE TEST SET) AND COMPUTATION TIMES OF MSVC, A NON-CONTEXTUAL SVM, MRF-GAUSS, MRF-SVM-POST, AND THE COMPOSITE-KERNEL AND GRAPH-KERNEL METHODS IN [16] AND [37]; Z STATISTICS OF THE MCNEMAR'S TEST APPLIED TO VALIDATE WHETHER THE DIFFERENCE BETWEEN THE ACCURACY OF MSVC AND THE ACCURACY OF EACH OF THE OTHER ALGORITHMS IS STATISTICALLY SIGNIFICANT. FOR EACH METHOD, BOTH THE TIME TAKEN BY THE WHOLE CLASSIFICATION PROCEDURE (INCLUDING PARAMETER OPTIMIZATION, TRAINING, AND APPLICATION OF THE CLASSIFIER TO THE INPUT IMAGE) AND THE TIME WITHOUT PARAMETER OPTIMIZATION (I.E., THE TIME TAKEN ONLY TO TRAIN THE CLASSIFIER AND TO APPLY IT TO THE INPUT IMAGE, ONCE THE PARAMETER VALUES ARE GIVEN) ARE REPORTED

class	training samples	test samples	MSVC	SVM	MRF-Gauss	MRF-SVM-Post	Composite kernel	Graph kernel
urban	1891	18982	97.47%	58.27%	78.02%	84.76%	85.42%	87.35%
herbaceous rangeland	1098	5546	99.77%	92.28%	96.48%	99.40%	100%	100%
shrub and brush rangeland	1891	48919	100%	98.76%	99.81%	99.93%	99.99%	100%
forest land	189	674	100%	91.39%	99.41%	100%	100%	100%
barren land	1148	3855	72.37%	68.33%	75.88%	72.27%	73.41%	68.72%
built-up land	1891	5267	96.28%	93.45%	98.82%	99.85%	99.15%	95.94%
water	1891	91546	99.97%	99.95%	99.94%	99.95%	99.98%	100%
overall accuracy			98.98%	93.92%	96.85%	97.66%	97.79%	97.81%
average accuracy			95.12%	86.06%	92.62%	93.74%	93.99%	93.14%
Z statistics (McNemar's test)			—	−91.72	−55.59	−43.49	−37.66	−36.3
time including parameter optimization			18 min	6 min	37 s	22 min	12 h 40 min	6 h 58 min
time without parameter optimization			17 min	5 min	35 s	6 min	7 min	16 min

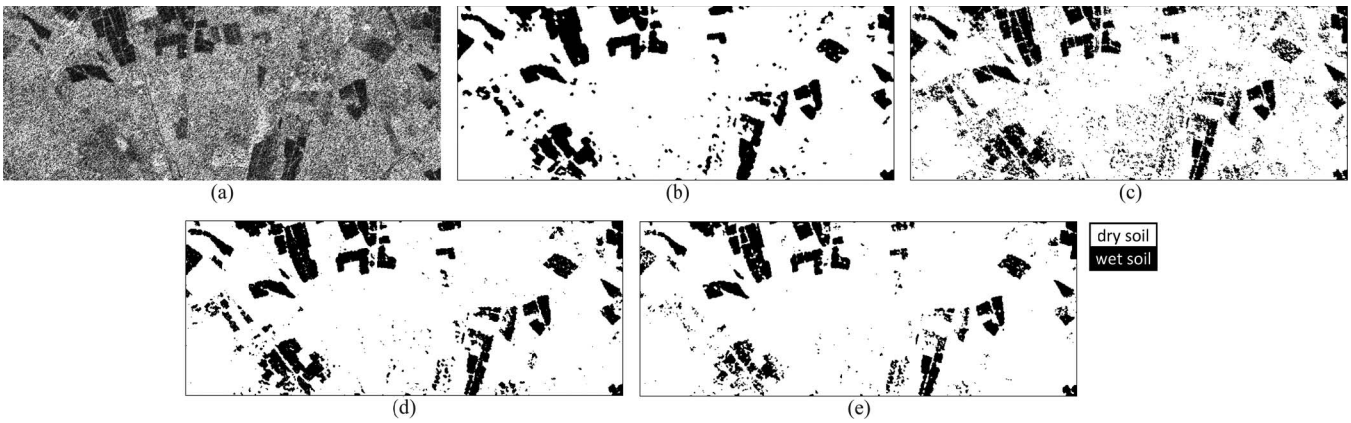

Fig. 4. "Pavia": (a) X-band channel after equalization and classification maps generated by (b) MSVC, (c) non-contextual SVM, (d) MRF-k-NN, and (e) the graph-kernel method in [37].

applied instead. The parameter k was optimized by threefold cross validation [1].

The Gaussian kernel was used, and C and σ were also optimized by the technique in Section IV-B for the non-contextual SVM and for MRF-SVM-Post. The smoothing parameters of the Potts models in the MRF-Gauss, MRF-k-NN, and MRF-SVM-Post methods were optimized by the aforementioned method in [20], which is based on the Ho–Kashyap algorithm. The energy functions of the related MRF models were minimized by ICM. In particular, in the case of the MRF-SVM-Post method, ICM was initialized with the multiclass classification map obtained by the non-contextual SVM.

Within the composite kernels introduced in [16], the "weighted-summation kernel" was used. It is a linear combination of two separate kernels, whose input vectors are the image channels and additional spatial features, respectively. As in [16], the employed spatial features were the local means and standard deviations of the image channels computed through a moving window. The graph kernel in [37] is obtained from a base kernel through a recursive moving-average procedure. For the base kernel in [37] and for the two separate kernels in [16], the Gaussian in (16) was used. With this choice, the graph-kernel and composite-kernel methods depend on four and five parameters, respectively (i.e., C, σ, the window size, and the recursion depth for [37]; C, the σ's of the two separate kernels, the window size, and the weight in the linear combination for [16]). As in [16] and [37], all such parameters were optimized by k-fold cross validation ($k = 3$ was used).

For each data set, the McNemar's test was separately applied to the map of each method used for comparison to validate whether the difference in accuracy between this map and the result of MSVC was statistically significant. The test computes the following standardized normal statistics:

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \tag{17}$$

where f_{12} is the number of test samples that are erroneously classified by MSVC and not by the comparison method, and f_{21} has a dual meaning [60], [61]. Accepting the common 5% level of significance, the difference between the results of MSVC and of each benchmark technique is statistically significant if $|Z| > 1.96$ [60], [61]. When this condition is met, a negative or positive value of Z indicates that MSVC or the compared method, respectively, is more accurate.

TABLE III
"PAVIA": TRAINING- AND TEST-SAMPLE SIZES; CLASSIFICATION ACCURACIES (ON THE TEST SET) AND COMPUTATION TIMES OF MSVC, A NON-CONTEXTUAL SVM, MRF-k-NN, MRF-SVM-POST, AND THE COMPOSITE-KERNEL AND GRAPH-KERNEL METHODS IN [16] AND [37]; Z STATISTICS OF THE MCNEMAR'S TEST APPLIED TO VALIDATE WHETHER THE DIFFERENCE BETWEEN THE ACCURACY OF MSVC AND THE ACCURACY OF EACH OF THE OTHER ALGORITHMS IS STATISTICALLY SIGNIFICANT. FOR EACH METHOD, BOTH THE TIME TAKEN BY THE WHOLE CLASSIFICATION PROCEDURE (INCLUDING PARAMETER OPTIMIZATION, TRAINING, AND APPLICATION OF THE CLASSIFIER TO THE INPUT IMAGE) AND THE TIME WITHOUT PARAMETER OPTIMIZATION (I.E., THE TIME TAKEN ONLY TO TRAIN THE CLASSIFIER AND TO APPLY IT TO THE INPUT IMAGE, ONCE THE PARAMETER VALUES ARE GIVEN) ARE REPORTED

class	training samples	test samples	MSVC	SVM	MRF-k-NN	MRF-SVM-Post	Composite kernel	Graph kernel
dry soil	6205	5082	95.43%	92.84%	94.79%	94.06%	97.28%	98.25%
wet soil	1346	1927	98.24%	91.13%	96.11%	95.23%	88.27%	90.45%
		overall accuracy	96.20%	92.37%	95.15%	94.38%	94.81%	96.11%
		average accuracy	96.84%	91.99%	95.45%	94.65%	92.78%	94.35%
		Z statistics (McNemar's test)	—	-13.71	-6.81	-9.19	-4.83	-0.39
time including parameter optimization			44 s	31 s	3 s	78 s	16 min	11 min
time without parameter optimization			12 s	1 s	1 s	2 s	2 s	4 s

Due to the impact of the Hughes phenomenon on Gaussian Bayesian classifiers in high-dimensional spaces [1], MRF-Gauss was applied to "Indian Pine" after performing a preliminary feature reduction. The sequential forward band extraction (SFBE) technique was used for this purpose [9]. SFBE is a supervised method that, given a set of hyperspectral channels, extracts a set of synthetic multispectral channels that are optimized to maximize accuracy in a given classification problem. Here, it was chosen as a feature-reduction tool due to its accurate performance in the application to "Indian Pine" (details can be found in [9], where the training and test sets were the same as in the present paper). The optimum reduced number of features for SFBE with "Indian Pine" was 26 [9].

To assess the possible sensitivity of MSVC to the Hughes phenomenon and the possible need for feature reduction, dedicated experiments were performed with "Indian Pine." The training set was progressively subsampled, and the performances of MSVC with and without feature reduction (performed again by SFBE; see Section V-C) were assessed. Finally, a dedicated analysis was conducted to investigate the automatic parameter-optimization algorithms plugged into MSVC and to compare them with benchmark grid-search techniques (see Section V-D).

B. Classification Results

Tables I–III report the classification accuracies obtained by MSVC and the aforementioned benchmark classifiers over the test sets of the three data sets. In the case of "Indian Pine," the results reported in Table I for the considered support vector classifiers were obtained with all available 202 hyperspectral channels, i.e., without feature reduction. For all data sets, fewer than ten iterations were needed by MSVC to reach convergence. Similarly, fewer than 5 and 60 000 iterations were needed by the Powell and Ho–Kashyap algorithms, respectively, to reach convergence for all data sets.

MSVC obtained very accurate classification results with overall accuracies (OAs) around 93%, 99%, and 96% for "Indian Pine," "Itaipu," and "Pavia," respectively. Moreover, MSVC allowed approximately 6.4%, 1%, and 2% increases in OA with respect to the traditional MRF-Gauss and MRF-k-NN classifiers for "Indian Pine," "Itaipu," and "Pavia," respectively, and $4 \div 7\%$ increases in OA, as compared to the non-contextual SVM for all three data sets. Similar comments hold with regard to average accuracies (Tables I–III). These results indicate the effectiveness of the developed SVM-MRF framework and the ability of MSVC to exploit this framework to outperform the results provided by separate applications of classical SVM- and MRF-based classifiers. Accuracy improvements were also obtained by MSVC in comparison to MRF-SVM-Post, which represents a previous attempt at combining SVMs and MRFs by plugging the approximated posterior probabilities given by [23], [24] into a traditional MRF classifier. Also, the composite and graph-kernel methods in [16] and [37] generated accurate results. However, MSVC obtained higher overall and average accuracies for all three data sets. In the case of "Indian Pine," a $4 \div 5\%$ difference in OA was observed between MSVC and these two techniques, while smaller differences could be observed for the other two data sets. These results confirm the effectiveness of the composite and graph-kernel approaches to the contextual generalization of SVMs and demonstrate the improved accuracy of the proposed method compared to these previous techniques. This result is interpreted as being due to the spatial-modeling capability of the Markovian approach, which is an intrinsic component of the proposed integrated framework. MSVC obtained high accuracies on "Indian Pine" even though all hyperspectral channels were used and strongly overlapping classes, related to similar vegetated covers, were involved. This result confirms that MSVC can also take advantage of the robustness of SVM-based techniques against the Hughes phenomenon.

According to McNemar's test, the differences between the accuracies of MSVC and of the previous methods were statistically (very) significant in all cases but one. $Z < -5$ and $Z < -30$ were obtained with "Indian Pine" and "Itaipu," respectively (Tables I and II). In the case of "Pavia," $Z < -4$ was obtained when comparing MSVC with each previous technique except for the graph kernel, while the difference in accuracy between MSVC and the graph kernel was not significant (Table III). In particular, very large values of $|Z|$ were obtained when comparing MSVC and the non-contextual SVM. Large values of $|Z|$, albeit smaller than in the case of the non-contextual SVM, were also achieved in the comparisons with the previous Markovian classifiers.

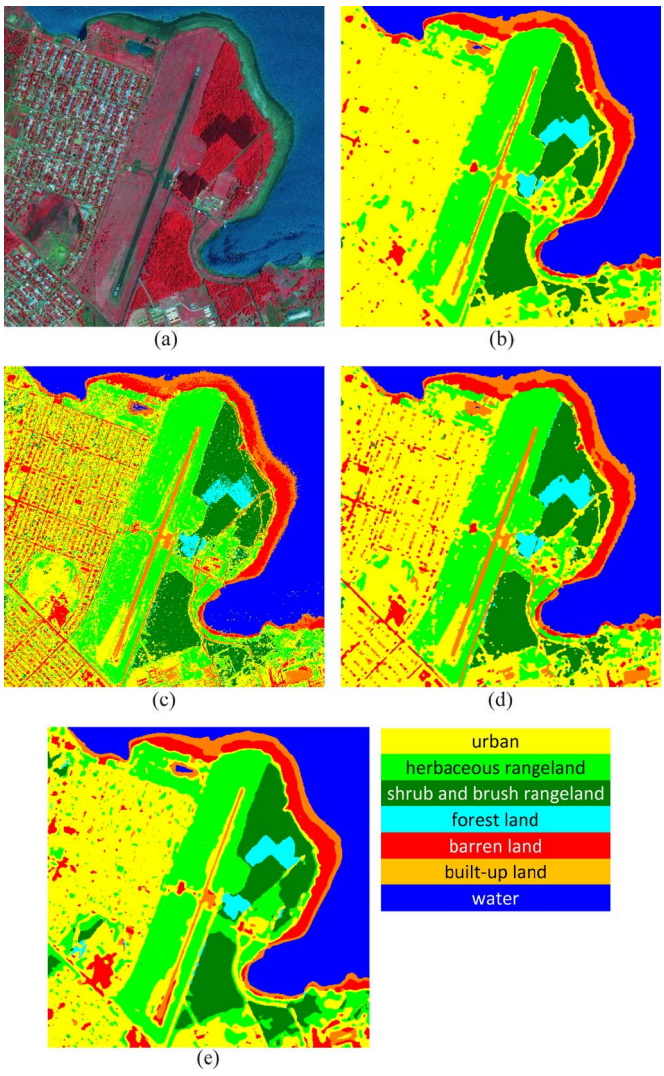

Fig. 5. "Itaipu": (a) details of the RGB false-color composition in Fig. 3(a) and of the maps generated by (b) MSVC, (c) non-contextual SVM, (d) MRF-SVM-Post, and (e) the graph-kernel method in [37].

Visual comparisons between the classification maps of MSVC [Figs. 2(b), 3(b), and 4(b)] and of the non-contextual SVM [Figs. 2(c), 3(c), and 4(c)] confirm the increased accuracy of the former. Incorporating contextual information into support vector classification through MSVC also makes it possible to significantly improve the spatial regularity of the classification results and reduce the sensitivity to noise or speckle. In the cases of "Pavia" and "Itaipu," improved spatial regularity of MSVC was also remarked in comparison to MRF-k-NN, MRF-SVM-Post, and the graph kernel [Figs. 4(d) and (e), 3(d) and (e)]. This aspect is particularly relevant for "Pavia" because the equivalent numbers of looks of all four SAR channels were $3 \div 4$, i.e., speckle was quite strong in this data set. This result confirms the importance of the contextual Markovian approach integrated into the proposed framework. In the case of "Indian Pine," the spatial regularities of the results from the MSVC, MRF-Gauss, and graph-kernel methods [Figs. 2(d) and (e)] were similar. Similar findings were obtained for the maps generated using the composite-kernel technique (which are omitted for brevity).

Fig. 5 focuses on an urban area in the "Itaipu" scene. MSVC [Fig. 5(b)] discriminates this area quite well almost as a whole, with a rather limited confusion with other classes, even though the "urban" class is spatially very heterogeneous in high-resolution images. In contrast, the non-contextual SVM, as expected, was severely affected by this heterogeneity [Fig. 5(c)]. Thanks to spatial modeling, the MRF-SVM-Post and graph-kernel techniques generated less noisy results than the non-contextual SVM, though with significantly more confusion between "urban," "barren land," and "herbaceous" than that observed using MSVC [Figs. 5(d) and (e)].

MSVC, applied with a quad-core 2.51-GHz CPU with 2-GB RAM, took approximately 17 min, 18 min, and 44 s with "Indian Pine," "Itaipu," and "Pavia," respectively (see Tables I–III), to complete the whole classification procedure, including parameter optimization, training, and application of the classifier to the whole image. These times were rather short, even though spatially large ("Itaipu") and spectrally large ("Indian Pine") data sets were involved, and are typically acceptable in land-cover or land-use mapping applications. Because the non-contextual SVM is the initialization of MSVC, it obviously took less time than MSVC. Very short times were required by MRF-Gauss and MRF-k-NN, mostly because the related Ho–Kashyap procedures needed only a few iterations to converge. However, the reduced computational burden of the non-contextual SVM and classical MRF classifiers was obtained at the cost of worse accuracy. Longer times were taken by MRF-SVM-Post than by MSVC. In the case of "Indian Pine," MRF-SVM-Post took 42 min due to the relatively long convergence time of the Ho–Kashyap procedure for this specific classifier and data set. The composite and graph-kernel classifiers took considerably longer times (a few hours with "Indian Pine" and "Itaipu" and more than 10 min with "Pavia") due to the grid search performed in the multidimensional space of their parameters to search for the minimum cross-validation error. Such long times were observed not because of the formulations of the composite and graph-kernel functions but because, unlike MSVC, automatic optimization algorithms for their parameters have not yet been proposed. If the time taken by parameter optimization is not accounted for (i.e., the parameters are fixed at their optimal values, and only the time required to train the classifier and to apply it to the input image is considered), then the composite-kernel classifier, the non-contextual SVM, and MRF-SVM-Post exhibited very similar computational burdens. The graph-kernel classifier was approximately twice as slow due to the recursive kernel computation, and MSVC was approximately two to six times slower due to the ICM iterations. Note that the cross-validation procedure was slower for the composite-kernel classifier than for the graph-kernel classifier because the former and the latter include five and four parameters, respectively. These results further confirm the importance of plugging efficient parameter optimization techniques into the contextual classification and favorably confirm that MSVC, endowed with the Powell-span-bound and Ho–Kashyap-based parameter-optimization algorithms, is an overall effective strategy, at least for the data sets considered, for incorporating contextual information into SVMs.

Finally, we recall that in the original version of "Indian Pine," 16 classes were included and endowed with ground truth.[4] In the present paper, nine classes were used to ensure that the results were comparable with those presented in some of our previous publications (e.g., [8], [9], and [62]), in which Gaussian Bayesian classifiers were used and the classes whose training set sizes were too small to compute non-singular sample covariance matrices were discarded. When the considered methods were applied to "Indian Pine" with all 16 classes, the resulting OAs were 87.86% for MSVC, 78.56% for the non-contextual SVM, 83.63% for MRF-Gauss, 80.60% for MRF-SVM-Post, 83.16% for the composite-kernel classifier, and 83.28% for the graph-kernel classifier. MRF-Gauss was applied after reducing to eight features through SFBE to allow all sample covariance matrices to be nonsingular [1]. Hence, the same comments made previously about the improved accuracy of MSVC compared with the previous benchmark classifiers and about the effectiveness of the MRF approach for contextual classification are further confirmed by this 16-class experiment. These accuracies were lower than those in the nine-class experiment because some of the additional seven classes were spectrally very overlapping (in the feature space) with one another and with the grass and corn classes listed in Table I. Note that, also in this case, spatially disjoint training and test areas were used to avoid a positive bias in the accuracies.

C. Experiments With Training-Set Reduction

Focusing on the hyperspectral data set ("Indian Pine," nine classes, see Table I), ten experiments were conducted by progressively removing, at each run and for each class, a randomly selected 20% of the training samples until approximately 15% of the original training-set size (which was 3726) was retained. To address the possible need for feature reduction, MSVC was separately applied to each subsampled training set, both with all 202 channels and after using SFBE to extract s features $(2 \leq s \leq 60)$. Fig. 6 shows the behavior of the OA of MSVC as a function of the removed training-set percentage, both with and without feature reduction. In the former case, the maximum OA obtained as s ranged in [2, 60] is plotted as a function of the removed training-set percentage.

When using the whole training set, no accuracy improvement was obtained by feature reduction compared to the case with all 202 channels. When applying MSVC to all channels after subsampling the training set, OA was always above 88% in all runs. No relevant accuracy gain was obtained by feature reduction compared to the use of all channels when up to approximately 65% of the training set was removed. Feature reduction resulted in a $1 \div 2\%$ gain in OA when $70 \div 85\%$ was removed. These results confirm that MSVC exhibits robustness against the Hughes phenomenon typical of SVM-based classifiers, and overall they do not suggest a need for a feature-reduction step. Such a step may yield a minor accuracy improvement only in cases of small sample sizes (at least for the considered hyperspectral data set).

[4]See ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C.tif.zip.

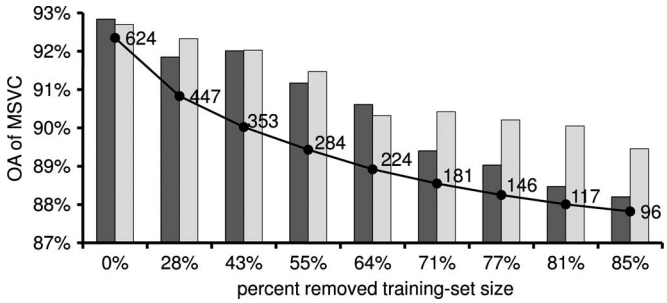

Fig. 6. "Indian Pine": Plot of the OA of MSVC as a function of the percentage of removed training samples without (dark gray) and with feature reduction (light gray). In the latter case, the maximum OA is plotted, as the reduced number of features ranges in [2, 60]. The black line plots the average number of training samples per class.

D. Experiments About Parameter Setting

Focusing first on (C, σ), the Powell-span-bound method was compared with an exhaustive grid search for the parameter values that yield the maximum OA on the test set for the non-contextual initialization phase of MSVC (named "hold-out grid search" [63]). This comparative approach was chosen because the span bound is related to the error rate of the non-contextual initialization. The hold-out grid search makes use not only of training samples but also of test samples, whereas the Powell-span-bound method only involves the training set. Thus, the comparison is partially unfair and puts the Powell-span-bound method at disadvantage. From this perspective, the hold-out grid search is used here as a benchmark. An alternate strategy would be to split the training set into two subsets, using the first subset to train the classifier and performing a grid search for the parameter values yielding the maximum accuracy over the second subset. This approach, named "validation-set grid search," could be operatively used to apply the grid search to the available set of input labeled samples. For the sake of comparison, we preferred the hold-out method over the validation-set strategy because the latter is affected by the specific subdivision of the training set into two subsets, while the Powell-span-bound method does not require this subdivision and exploits all available training samples.

Specifically, $\log_{10} C$ and $\log_{10} \sigma$ were varied in $[-2, 5]$ and $[-0.75, 1.25]$, respectively. These two ranges were discretized and a non-contextual SVM was separately trained for each node of the resulting rectangular grid. As (C, σ) varied over the grid, remarkable correlation coefficients, equal to 0.89, 0.87, and 0.95 for "Indian Pine," "Itaipu," and "Pavia," respectively, were obtained between the overall error rates (i.e., $1-$OA) of the non-contextual classification map on the test set and the span bounds. This result is visually consistent with the plots of the span bound and of the overall error rate as functions of (C, σ) (see Fig. 7 for such plots in the case of "Pavia"; similar plots, not shown for brevity, were obtained for the other data sets) and confirms that the choice of the span bound as the functional leading the optimization of the SVM parameters is appropriate.

An accuracy comparison was also performed in the proposed contextual framework between the Powell-span-bound and the hold-out grid search. For this purpose, the non-contextual map obtained by training an SVM with the parameter values

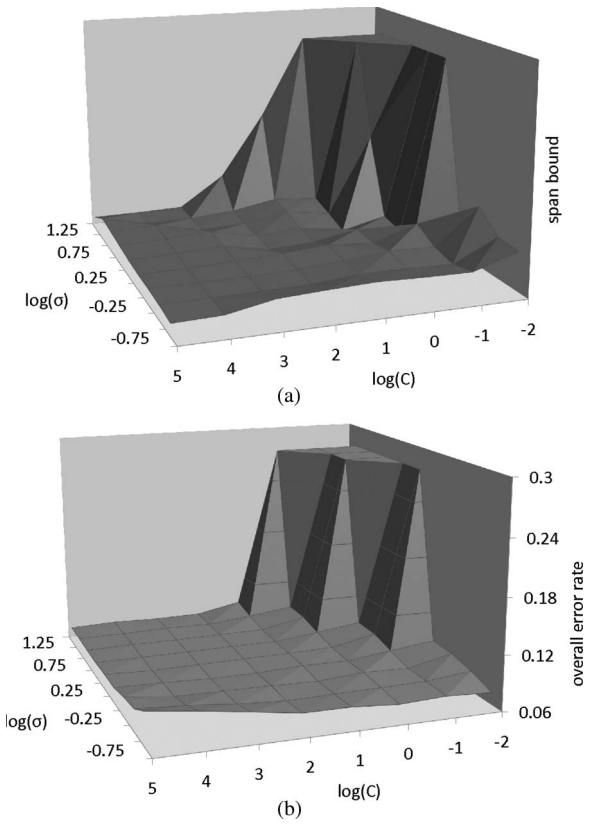

Fig. 7. "Pavia": 2-D plots of the (a) span bound and (b) of the overall error rate (on the test set) of the non-contextual initialization of MSVC, as functions of (C, σ).

determined by the grid search was used both to run the Ho–Kashyap algorithm for the estimation of β and to initialize ICM. The results of this procedure (Table IV, Case A) were compared with those of MSVC. A remarkable similarity can be noted between the accuracies of the two approaches. With "Pavia," the grid search allowed a 1.2% gain in OA compared to MSVC, which provided more accurate results for "Indian Pine" (1.9% difference in OA) and "Itaipu" (almost negligible difference in accuracy). This similarity in accuracy holds even though different parameter values were obtained by the two procedures, a result consistent with the plots in Fig. 7, which show large valley-like behaviors with many local minima and maxima. The solutions of the Powell-span-bound and grid-search approaches correspond to different minima related to distinct parameter values but to very similar accuracies, a behavior already noted in [21] in the case of regression. These results confirm the accuracy of the Powell-span-bound algorithm for support vector classification and the effectiveness of its integration into MSVC. In general, higher accuracies might be expected for the grid search than for this algorithm due to the exhaustive hold-out strategy. However, the grid search explores the parameter space up to a discretization step, which critically affects the overall execution time, whereas the proposed algorithm does not involve discretizing the parameters. This difference explains why this latter procedure was (slightly) more accurate for two data sets.

With regard to the optimization of β, the Ho–Kashyap-based algorithm was compared with a hold-out grid search performed as follows: 1) ICM was initialized with the non-contextual map obtained using the values of C and σ determined by the Powell-span-bound algorithm; 2) a discrete scale was defined on the range $\log_{10}\beta \in [-5, 1]$, and ICM was separately run until convergence with each value of β in this grid; and 3) the value of β corresponding to the highest test-set OA in the grid was selected. The comments above on the (un)fairness of this comparison and on the use of the hold-out grid search as a benchmark hold here as well. As expected, this search produced slightly more accurate results (Table IV, Case B) than the Ho–Kashyap-based algorithm, but the difference in OA was below 1% for all data sets. This result confirms the effectiveness of this extension of the technique in [20] to the estimation of the smoothing parameter β in the proposed integrated framework.

VI. Conclusion

A novel SVM-MRF integrated framework for remote-sensing image classification has been developed by establishing an analytical relationship between the Markovian minimum-energy decision rule and a suitable SVM-kernel expansion. An additional contextual feature and an appropriate kernel, both derived from the adopted MRF model, determine this expansion. A novel supervised contextual classifier has also been developed by specializing the ICM technique for the proposed framework. When applied to hyperspectral, multichannel SAR, and high-resolution multispectral images, the method produced very accurate results, outperforming both a non-contextual SVM and previous Markovian classifiers based on parametric and nonparametric class-statistics modeling. These results suggest a strong flexibility of the method in combining the generalization capability of SVMs and the spatial-modeling properties of MRFs to classify diverse types of remote-sensing data. The proposed classifier also favorably compared, in terms of both accuracy and computational burden, with a previous approach to the combination of SVMs and MRFs, based on the calculation of approximate posterior probabilities from a non-contextual SVM discriminant function [23], [24]. The developed method obtained improvements in accuracy also as compared to two previous non-Markovian contextual extensions of SVM, which were based on composite and graph kernels [16], [37]. The accuracy obtained in the experiments using hyperspectral data also highlighted a remarkable robustness of the method to the Hughes phenomenon, essentially suggesting the possible need for preliminary feature reduction only when few training samples were available.

Given a kernel function and a spatial MRF model (with respect to a predefined neighborhood system), the proposed classifier automatically optimizes both the SVM kernel and regularization parameters and the MRF spatial smoothing parameter. This is accomplished by extending previous techniques developed by the authors and based on the Powell and Ho–Kashyap algorithms. The experimental results produced by such automatic parameter optimization were very similar to those produced by the benchmark case of grid searches for the parameter values corresponding to the highest accuracy for the test set. These results confirm the accuracy of the parameter-optimization techniques incorporated into the proposed

TABLE IV
CLASSIFICATION ACCURACY OF MSVC, WHEN APPLIED WITH: (CASE A) (C, σ) OPTIMIZED BY HOLD-OUT GRID SEARCH AND β DERIVED BY APPLYING THE METHOD IN SECTION IV-C TO THE RESULTING NON-CONTEXTUAL INITIALIZATION MAP; (CASE B) (C, σ) OPTIMIZED BY THE METHOD IN SECTION IV-B AND β OPTIMIZED BY HOLD-OUT GRID SEARCH

"Indian Pine"			"Itaipu"			"Pavia"		
class	Case A	Case B	class	Case A	Case B	class	Case A	Case B
corn-no	88.00%	90.61%	urban	98.16%	97.49%	dry	96.62%	95.59%
corn-min	90.49%	96.32%	herbaceous	96.03%	99.66%	wet	99.27%	99.79%
grass/pasture	92.44%	93.33%	shrub-brush	99.99%	100%			
grass/trees	99.29%	98.23%	forest	100%	100%			
hay-windrowed	100%	100%	barren	77.51%	73.10%			
soybean-no	79.01%	90.97%	built-up	91.93%	96.30%			
soybean-min	91.99%	92.74%	water	99.95%	99.98%			
soybean-clean	82.30%	84.07%						
woods	97.95%	98.56%						
OA	90.96%	93.64%	OA	98.91%	99.00%	OA	97.35%	96.75%
AA	91.27%	93.87%	AA	94.80%	95.22%	AA	97.95%	97.69%

classifier and suggest that it can be used as a powerful contextual classification tool, able to jointly exploit the SVM and MRF approaches in a completely automatic way.

A limitation may be that the MRF-parameter optimization algorithm focuses on MRF models characterized by a single parameter that weights a spatial energy contribution. However, a straightforward extension can be formulated for the case of multiple weight parameters that tune the tradeoff among several energy contributions (e.g., related to distinct sources of contextual information [11], [20]) by introducing multiple additional contextual features and by accordingly updating the Markovian kernel. As in [20], however, we did not address the problem of estimating possible parameters inside each energy contribution. From this perspective, an interesting extension of the proposed classifier could be the integration of further MRF parameter-estimation algorithms (e.g., based on approximate log-likelihood functionals [64]) that also take into account these possible additional parameters.

Furthermore, it would be interesting to formulate SA in the proposed integrated framework to minimize dependence on the initialization map due to its better ability to approach global energy minima. Graph-cut methods have also recently attracted considerable attention in MRF-based classification. They can reach, in acceptable times, global energy minima in the case of binary classification and local minima with strong optimality properties in the multiclass case [65], [66]. The proposed SVM-MRF framework could be extended to support graph-cuts by suitably reformulating its energy-function formalization. Testing the proposed classifier with more sophisticated MRFs could also be a future extension of this work. This addition will be feasible because the method involves no restriction on the adopted MRF model. Combinations with anisotropic [5], edge-preserving [6], or multiresolution [15] MRFs and with region-based [67] or texture-based [68], [69] approaches for spatial image modeling would be particularly important to further optimize accuracy with high-resolution images.

APPENDIX I
PROOF OF THEOREM 1

With the same notations as in Section III-D, because $\tilde{\boldsymbol{x}}_i$ depends only on $\boldsymbol{x}_i$ for each pixel $i \in \mathcal{I}$, conditional independence also holds for the random field $\{\tilde{\boldsymbol{x}}_i\}_{i\in\mathcal{I}}$ of the transformed vectors, conditioned to the label field. Therefore, according to (5), the local posterior energy corresponding to the application of the MRF minimum-energy rule in $\mathcal{F}$ can be written (up to additive constants) as

$$U_i(y_i|\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}) = \beta\mathcal{E}_i(y_i|\boldsymbol{y}_{\partial i}) - \frac{1}{2}\ln\det\Sigma + \\ - \begin{cases} \frac{1}{2}(\tilde{\boldsymbol{x}}_i - \boldsymbol{m}^+)^\top\Sigma^{-1}(\tilde{\boldsymbol{x}}_i - \boldsymbol{m}^+) & \text{for } y_i = 1 \\ \frac{1}{2}(\tilde{\boldsymbol{x}}_i - \boldsymbol{m}^-)^\top\Sigma^{-1}(\tilde{\boldsymbol{x}}_i - \boldsymbol{m}^-) & \text{for } y_i = -1. \end{cases} \tag{18}$$

By invoking the equal-covariance assumption, the resulting energy-difference function is a linear function of $\tilde{\boldsymbol{x}}_i$, i.e.,

$$\Delta U_i(\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}) = \tilde{\boldsymbol{w}}^\top\tilde{\boldsymbol{x}}_i + b + \beta\varepsilon_i = \\ = \langle w, \Phi(\boldsymbol{x}_i)\rangle_{\mathcal{F}} + b + \beta\varepsilon_i \tag{19}$$

where $\tilde{\boldsymbol{w}}$ and $b(\tilde{\boldsymbol{w}} \in \mathbb{R}^D; b \in \mathbb{R})$ are suitable algebraic combinations of $\boldsymbol{m}^+$, $\boldsymbol{m}^-$ and Σ, $w = \boldsymbol{A}^{-1}(\tilde{\boldsymbol{w}})(w \in \mathcal{F})$, ε_i is given by (8), and the equality $\tilde{\boldsymbol{w}}^\top\tilde{\boldsymbol{x}}_i = \langle w, \Phi(\boldsymbol{x}_i)\rangle_{\mathcal{F}}$ holds due to the isometry between $\mathcal{F}$ and $\mathbb{R}^D$.

If $\mathcal{G}$ is the $(D+1)$-dimensional space of all pairs $(u, r)(u \in \mathcal{F}, r \in \mathbb{R})$, endowed with the inner product $\langle(u, r), (v, s)\rangle_{\mathcal{G}} = \langle u, v\rangle_{\mathcal{F}} + rs(u, v \in \mathcal{F}; r, s \in \mathbb{R})$ and often called the direct sum of the spaces $\mathcal{F}$ and $\mathbb{R}$ [51], then

$$\Delta U_i(\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}) = \langle W, \Psi(\boldsymbol{x}_i, \varepsilon_i)\rangle_{\mathcal{G}} + b \tag{20}$$

where $\Psi(\boldsymbol{x}, \varepsilon) = (\Phi(\boldsymbol{x}), \varepsilon\sqrt{\beta})(\boldsymbol{x} \in \mathcal{X}, \varepsilon \in \mathbb{R})$ and $W = (w, \sqrt{\beta}) \in \mathcal{G}$. Therefore, the energy-difference function associated with the minimum-energy rule in $\mathcal{F}$ turns out to be equivalent to a linear discriminant function in $\mathcal{G}$. This energy difference can be obtained by augmenting the original d-dimensional feature space with the additional feature ε_i and by mapping the resulting $(d+1)$-dimensional space in $\mathcal{G}$ through Ψ. As in Section III-B, such a linear discriminant function can be equivalently computed as the kernel expansion in (7), where the support vectors and their coefficients are determined by the QP problem in (3), and K is replaced by $(\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}; \varepsilon, \varepsilon' \in \mathbb{R})$

$$K_{\mathsf{MRF}}(\boldsymbol{x}, \varepsilon; \boldsymbol{x}', \varepsilon') = \langle\Psi(\boldsymbol{x}, \varepsilon), \Psi(\boldsymbol{x}', \varepsilon')\rangle_{\mathcal{G}} \\ = K(\boldsymbol{x}, \boldsymbol{x}') + \beta\varepsilon\varepsilon'. \tag{21}$$

This expression corresponds to the kernel in (9) and completes the proof.

APPENDIX II
PROOF OF PROPOSITION 1 AND ANALYTICAL COMMENTS

Let us first recall a few basic definitions and results from linear functional analysis [50], [51]: 1) an inner product space is a Hilbert space when it is complete with respect to the metric induced by its inner product; 2) a Hilbert space is separable when it admits a dense enumerable subset; 3) all infinite-dimensional separable Hilbert spaces are isometrically isomorphic to one another; 4) an example of an infinite-dimensional separable Hilbert space is the space $L^2[0,1]$ of the Lebesgue square-integrable real-valued signals, equipped with the usual integral inner product

$$\langle u,v\rangle_{L^2}=\int_0^1 u(t)v(t)dt \qquad \forall u,v\in L^2[0,1]. \tag{22}$$

The inner product space $\mathcal{F}$ induced by a kernel can always be proven to be Hilbert separable [2]. Thus, if $\dim\mathcal{F}=+\infty$, then $\mathcal{F}$ is isometrically isomorphic to $L^2[0,1]$.

The equal-autocovariance condition in Assumption 5 means that the following equations hold for each pair of times $t,\tau(0\le t,\tau\le 1)$ [53], [70]

$$\begin{aligned}\Sigma(t,\tau)&=E\left\{\left[\tilde{x}_i(t)-m^+(t)\right]\left[\tilde{x}_i(\tau)-m^+(\tau)\right]|y_i=1\right\}\\&=E\left\{\left[\tilde{x}_i(t)-m^-(t)\right]\left[\tilde{x}_i(\tau)-m^-(\tau)\right]|y_i=-1\right\}.\end{aligned} \tag{23}$$

The continuity and positive definiteness of this autocovariance function Σ are accepted to ensure analytical tractability. They basically prevent degenerate cases and are not severely restrictive. In particular, to prove Proposition 1, we note that, through Assumption 5, $\tilde{x}_i(t)(0\le t\le 1)$ can be decomposed as either $\tilde{x}_i(t)=m^+(t)+\eta(t)$ or $\tilde{x}_i(t)=m^-(t)+\eta(t)$ for $y_i=1$ or $y_i=-1(i\in\mathcal{I})$, respectively, where $\eta(t)$ is a zero-mean Gaussian noise, independent of y_i and with autocovariance function Σ.

Because Σ is a continuous and positive-definite function,[5] the Karhunen–Loeve expansion can be applied to expand $\eta(t)$ as the superposition of a sequence of orthogonal, uncorrelated Gaussian components [53]. Then, the whitening transformation, which divides each sample of this sequence by its standard deviation to obtain unitary variance, further allows $\eta(t)$ to be mapped onto a discrete-time unit-variance white Gaussian process.

More formally, there exists a positive sequence $\{\lambda_n\}_{n=0}^{\infty}$ of eigenvalues of Σ and a corresponding sequence $\{e_n(t)\}_{n=0}^{\infty}$ of real eigenfunctions such that $\{e_n(t)\}_{n=0}^{\infty}$ is a complete orthonormal basis of $L^2[0,1]$ and the following expansion holds:

$$\eta(t)=\sum_{n=0}^{\infty}\eta_n e_n(t) \tag{24}$$

where $\eta_n=\langle\eta,e_n\rangle_{L^2}$ is the projection coefficient of $\eta(t)$ onto the nth basis function $e_n(t)$. These coefficients form a zero-mean Gaussian sequence with uncorrelated samples and variances λ_n, i.e., $(n,n'=0,1,\ldots;n\ne n')$[53]

$$E\{\eta_n\}=0,\qquad E\left\{\eta_n^2\right\}=\lambda_n,\qquad E\{\eta_n\eta_{n'}\}=0. \tag{25}$$

Moreover, the positive-definiteness condition implies that $\lambda_n>0$ for all n and, because $\{e_n(t)\}_{n=0}^{\infty}$ is a complete orthonormal basis of $L^2[0,1]$, the following equality holds [50]:

$$\langle u,v\rangle_{L^2}=\sum_{n=0}^{\infty}\langle u,e_n\rangle_{L^2}\langle v,e_n\rangle_{L^2} \qquad \forall u,v\in L^2[0,1]. \tag{26}$$

Let $\mathcal{H}$ be the space of real-valued sequences $s(n)(n=0,1,2,\ldots)$ such that $\sum_{n=0}^{\infty}\lambda_n s(n)^2<+\infty$. $\mathcal{H}$ may be readily proven to be a Hilbert space when endowed with the following inner product:

$$\langle r,s\rangle_{\mathcal{H}}=\sum_{n=0}^{\infty}\lambda_n r(n)s(n), \qquad \forall r,s\in\mathcal{H}. \tag{27}$$

Due to (26) and (27), the linear transformation R that associates with the continuous-time signal $u(t)$ in $L^2[0,1]$ the following sequence $s(n)$ of normalized projection coefficients:

$$s(n)=\frac{\langle u,e_n\rangle_{L^2}}{\sqrt{\lambda_n}}, \qquad n=0,1,2,\ldots \tag{28}$$

is an isometry from $L^2[0,1]$ to $\mathcal{H}$. R is also bijective because $\{e_n(t)\}_{n=0}^{\infty}$ is a complete orthonormal basis. Hence, $L^2[0,1]$ and $\mathcal{H}$ are isometrically isomorphic. Therefore, the compound mapping $B(\cdot)=R[A(\cdot)]$ from $\mathcal{F}$ to $\mathcal{H}$ is a bijective isometry, i.e., $\mathcal{F}$ and $\mathcal{H}$ are isometrically isomorphic as well.

In particular, for each $i\in\mathcal{I}$, $\bar{x}_i=R(\tilde{x}_i)=B[\Phi(\boldsymbol{x}_i)]$ is a sequence $\bar{x}_i(n)(n=0,1,2,\ldots)$ in $\mathcal{H}$ and can be written as either $\bar{x}_i(n)=\mu^+(n)+\nu(n)$ or $\bar{x}_i(n)=\mu^-(n)+\nu(n)$ for $y_i=1$ or $y_i=-1$, respectively, where $\mu^+=R(m^+),\mu^-=R(m^-)$, and $\nu=R(\eta)$. The AWGN model in (10) follows from the linearity of R. Thus, the proof is complete.

APPENDIX III
PROOF OF THEOREM 2

The formulation in (10) is formally very similar to the decision problems encountered in digital communications when designing the optimum receiver for transmission over an ideal AWGN channel, i.e., a channel that introduces AWGN and does not distort the input signal [18]. For instance, in the case of binary communication, the received random signal, when conditioned to the transmission of either the "0" or "1" bit, can be expressed as the related class-conditional mean additively corrupted by AWGN. The close similarity between this mathematical model for digital communications and the probabilistic model in Proposition 1 makes it possible to generalize the theory of Bayesian optimum receiver design under AWGN conditions to the present problem.

More precisely, using the notations introduced in Section III-E, the subspace $\mathcal{P}$ of $\mathcal{H}$ spanned by the sequences $\mu^+(n)$ and $\mu^-(n)(n=0,1,2,\ldots)$ can be either one- or 2-D.

[5] We recall that the autocovariance function Σ is always positive semidefinite, i.e., the condition $\int\int_{[0,1]^2}\Sigma(t,\tau)u(t)u(\tau)dtd\tau\ge 0$ holds for all $u\in L^2[0,1]$. In particular, Σ is positive definite if the strict inequality holds for all u that do not identically vanish in [0, 1] [53].

Assuming $\dim \mathcal{P} = 2$ (the case $\dim \mathcal{P} = 1$ is analogous), an orthonormal basis of $\mathcal{P}$ is made of two sequences $\phi(n)$ and $\psi(n)$ and can be easily obtained, for instance, by the classical Gram–Schmidt procedure [18]. Hence, $\bar{x}_i(n)$ can be uniquely decomposed as

$$\bar{x}_i(n) = x_i^{\mathcal{P}}(n) + x_i^{\perp}(n) \tag{29}$$

where

$$x_i^{\mathcal{P}}(n) = \langle \bar{x}_i, \phi \rangle_{\mathcal{H}} \phi(n) + \langle \bar{x}_i, \psi \rangle_{\mathcal{H}} \psi(n) \tag{30}$$

is the projection of $\bar{x}_i(n)$ onto $\mathcal{P}$, and $x_i^{\perp}(n)$ is the component of $\bar{x}_i(n)$ orthogonal to $\mathcal{P}$. The theorem of irrelevance ensures that the probability of error of a Bayesian classifier based on the projection components $\langle \bar{x}_i, \phi \rangle_{\mathcal{H}}$ and $\langle \bar{x}_i, \psi \rangle_{\mathcal{H}}$ is unaffected by the addition of arbitrary further features extracted from $x_i^{\perp}(n)$[18]. Due to the linearity of the projection operator, the 2-D vector $\boldsymbol{z}_i = [\langle \bar{x}_i, \phi \rangle_{\mathcal{H}}, \langle \bar{x}_i, \psi \rangle_{\mathcal{H}}]^{\top}$ is Gaussian when conditioned to either class. As in the case of optimum receiver design [18], it is easy to prove that $\boldsymbol{z}_i$ has the same covariance matrix for both classes. Therefore, the same arguments used in Section III-D can be used to write the energy-difference function corresponding to the application of the minimum-energy rule in $\mathcal{F}$ as

$$\Delta U_i(\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}) = \tilde{\boldsymbol{w}}^{\top} \boldsymbol{z}_i + b + \beta \varepsilon_i \tag{31}$$

where $\tilde{\boldsymbol{w}}, b$ and ε_i have the same meaning as in Appendix I. If $\tilde{\boldsymbol{w}} = [w_\phi, w_\psi]^{\top}$, then the sequence $\bar{w}(n) = w_\phi \phi(n) + w_\psi \psi(n)$ belongs to $\mathcal{P}$ and $\tilde{\boldsymbol{w}}^{\top} \boldsymbol{z}_i = \langle \bar{w}, \bar{x}_i \rangle_{\mathcal{H}}$ because $x_i^{\perp}(n)$ is orthogonal to $\mathcal{P}$. Therefore

$$\begin{aligned} \Delta U_i(\boldsymbol{x}_i, \boldsymbol{y}_{\partial i}) &= \langle \bar{w}, \bar{x}_i \rangle_{\mathcal{H}} + b + \beta \varepsilon_i \\ &= \langle w, \Phi(\boldsymbol{x}_i) \rangle_{\mathcal{F}} + b + \beta \varepsilon_i \end{aligned} \tag{32}$$

where $w = B^{-1}(\bar{w})(w \in \mathcal{F})$ and where the equality $\langle \bar{w}, \bar{x}_i \rangle_{\mathcal{H}} = \langle w, \Phi(\boldsymbol{x}_i) \rangle_{\mathcal{F}}$ holds because B is an isometry. This result is identical to the one obtained in (19) in the finite-dimensional case, which allows the same conclusions to be reached here as well.

The discussed methodological arguments are feasible because the Markovian minimum-energy rule can always be cast in a Bayesian fashion (see Section III-C), and the aforementioned projection onto a one- or 2-D subspace does not affect Bayesian decisions (i.e., the applications of Bayesian rules before and after the projection are equivalent to each other). On one hand, any other transformation that maps the infinite-dimensional space associated with a kernel to a finite-dimensional subspace and, at the same time, preserves Bayesian decisions could be used instead of the proposed approach, which is rooted in the theory of optimum receiver design under AWGN. On the other hand, methods that map an infinite-dimensional space onto a finite-dimensional subspace (e.g., kernel principal component analysis [71]), but for which the equivalence between Bayesian rules before and after the transformation is not theoretically ensured, cannot be straightforwardly plugged into the proposed arguments for SVM-MRF integration.

ACKNOWLEDGMENT

The authors would like to thank Prof. D. A. Landgrebe (Purdue Univ.) and Dr. P. Gamba (University of Pavia, Italy), for providing the "Indian Pine" and "Pavia" images, respectively, Dr. M. De Martino for her assistance in data preparation and experiments, and Mr. P. Irrera for his contribution to implementation and experimental validation. The support is gratefully acknowledged.

REFERENCES

[1] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.
[2] V. N. Vapnik, *Statistical Learning Theory*. Hoboken, NJ: Wiley, 1998.
[3] G. Camps-Valls and L. Bruzzone, Eds., *Kernel Methods for Remote Sensing Data Analysis*. Hoboken, NJ: Wiley, 2009.
[4] P. Mantero, G. Moser, and S. B. Serpico, "Partially supervised classification of remote sensing images using SVM-based probability density estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 559–570, Mar. 2005.
[5] S. Li, *Markov Random Field Modeling in Image Analysis*. Berlin, Germany: Springer-Verlag, 2009.
[6] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
[7] 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.
[8] S. B. Serpico, M. D'Incà, F. Melgani, and G. Moser, "A comparison of feature reduction techniques for classification of hyperspectral remote-sensing data," in *Proc. SPIE Conf. Image Signal Process. Remote Sens. VIII*, Crete, Greece, Sep. 22–27, 2002, pp. 347–358.
[9] S. B. Serpico and G. Moser, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 2, pp. 484–495, Feb. 2007.
[10] F. Melgani and S. B. Serpico, "A statistical approach to the fusion of the spectral and spatio-temporal contextual information for the classification of remote sensing images," *Pattern Recognit. Lett.*, vol. 23, no. 9, pp. 1053–1061, Jul. 2002.
[11] A. H. S. Solberg, T. Taxt, and A. K. Jain, "A Markov random field model for classification of multisource satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 100–113, Jan. 1996.
[12] F. Melgani and S. B. Serpico, "A Markov random field approach to spatio-temporal contextual image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 11, pp. 2478–2487, Nov. 2003.
[13] G. Moser and S. B. Serpico, "Unsupervised change detection from multichannel SAR data by Markovian data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2114–2128, Jul. 2007.
[14] G. Moser, S. B. Serpico, and G. Vernazza, "Unsupervised change detection from multichannel SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 2, pp. 278–282, Apr. 2007.
[15] G. Storvik, R. Fjortoft, and A. H. S. Solberg, "A Bayesian approach to classification of multiresolution remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 539–547, Mar. 2005.
[16] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.
[17] F. Lafarge, X. Descombes, and J. Zerubia, "Textural kernel for SVM classification in remote sensing: Application to forest fire detection and urban area extraction," in *Proc. ICIP*, Genoa, Italy, 2005, pp. III-1096–III-1099.
[18] J. M. Wozencraft and I. M. Jacobs, *Principles of Communications Engineering*. Hoboken, NJ: Wiley, 1965.
[19] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Stat. Soc.*, vol. 48, no. 3, pp. 259–302, 1986.
[20] S. B. Serpico and G. Moser, "Weight parameter optimization by the Ho–Kashyap algorithm in MRF models for supervised image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 12, pp. 3695–3705, Dec. 2006.
[21] G. Moser and S. B. Serpico, "Automatic parameter optimization for support vector regression for land and sea surface temperature estimation

from remote-sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 909–921, Mar. 2009.
[22] F. Bovolo and L. Bruzzone, "A context-sensitive technique based on support vector machines for image classification," in *Proc. PReMI*, vol. 3776, *LNCS*, 2005, pp. 260–265.
[23] J. Platt, "Advances in large margin classifiers," in *Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods*, A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000.
[24] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, Aug. 2004.
[25] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson, "Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 5, pp. 1267–1279, Oct. 2010.
[26] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.
[27] D. Liu, M. Kelly, and P. Gong, "A spatial-temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery," *Remote Sens. Envinron.*, vol. 101, no. 2, pp. 167–180, 2006.
[28] Z. Bing, L. Shanshan, J. Xiuping, G. Lianru, and P. Man, "Adaptive Markov random field approach for classification of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 973–977, Sep. 2011.
[29] A. A. Farag, R. M. Mohamed, and A. El-Baz, "A unified framework for MAP estimation in remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1617–1634, Jul. 2005.
[30] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.
[31] G. Camps-Valls, T. Bandos-Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.
[32] M. Marconcini, G. Camps-Valls, and L. Bruzzone, "A composite semisupervised SVM for classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 234–238, Apr. 2009.
[33] D. Tuia, F. Ratle, A. Pozdnoukhov, and G. Camps-Valls, "Multisource composite kernels for urban-image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 88–92, Jan. 2010.
[34] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. L. Rojo-Alvarez, and M. Martinez-Ramon, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1822–1835, Jun. 2008.
[35] B.-C. Kuo, C.-S. Huang, C.-C. Hung, Y.-L. Liu, and I.-L. Chen, "Spatial information based support vector machine for hyperspectral image classification," in *Proc. IGARSS*, 2010, pp. 832–835.
[36] H. Sahbi, J.-Y. Audibert, and R. Keriven, "Context-dependent kernels for object classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 699–708, Apr. 2011.
[37] G. Camps-Valls, N. Shervashidze, and K. M. Borgwardt, "Spatio-spectral remote sensing image classification with graph kernels," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 741–745, Oct. 2010.
[38] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden Markov support vector machines," in *Proc. ICML*, Washington, DC, 2003.
[39] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, Sep. 2005.
[40] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Proc. NIPS*, Vancouver, BC, Canada, 2003.
[41] B. Taskar, V. Chatalbashev, and D. Koller, "Learning associative Markov networks," in *Proc. ICML*, 2004, pp. 807–814.
[42] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin, "Learning structured prediction models: A large margin approach," in *Proc. ICML*, 2005, pp. 896–903.
[43] T. Finley and T. Joachims, "Parameter learning for loopy Markov random fields with structural support vector machines," in *Proc. ICML*, Corvallis, OR, 2007.
[44] P. Schnitzspan, M. Fritz, and B. Schiele, "Hierarchical support vector random fields: Joint training to combine local and global features," in *Proc. ECCV*, 2008, pp. 527–540.
[45] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Proc. Adv. Neural Inf. Process. Syst.*, 1998, vol. 11, pp. 487–493.
[46] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 203–210, Mar. 2005.
[47] Q. Jackson and D. Landgrebe, "Adaptive Bayesian contextual classification based on Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2454–2463, Nov. 2002.
[48] Z. Kato, J. Zerubia, and M. Berthod, "Satellite image classification using a modified Metropolis dynamics," in *Proc. ICASSP*, 1992, pp. 573–576.
[49] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Amer. Stat. Assoc.*, vol. 82, no. 397, pp. 76–89, Mar. 1987.
[50] L. Debnath and P. Mikusinski, *Introduction to Hilbert Spaces*. New York: Academic, 1990.
[51] J. Conway, *A Course in Functional Analysis*. New York: Springer-Verlag, 1985.
[52] M. Murat-Dundar and D. Landgrebe, "Toward an optimal supervised classifier for the analysis of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 1, pp. 271–277, Jan. 2004.
[53] R. Ash, *Information Theory*. Hoboken, NJ: Wiley, 1965.
[54] M.-W. Chang and C.-J. Lin, "Leave-one-out bounds for support vector regression model selection," *Neural Comput.*, vol. 17, no. 5, pp. 1188–1222, May 2005.
[55] V. Vapnik and O. Chapelle, "Bounds on error expectation for support vector machines," *Neural Comput.*, vol. 12, no. 9, pp. 2013–2036, Sep. 2000.
[56] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 131–159, 2002.
[57] J. Guo, N. Takahashi, and T. Nishi, "An efficient method for simplifying decision functions of support vector machines," *IEICE Trans. Fundam.*, vol. E89-A, no. 10, pp. 2795–2802, Oct. 2006.
[58] W. H. Press, S. A. Teukolsky, W. T. Wetterling, and B. P. Flannery, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge Univ. Press, 2002.
[59] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*. Reading, MA: Addison-Wesley, 1974.
[60] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, 2004.
[61] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.
[62] S. B. Serpico, G. Moser, and A. F. Cattoni, "Feature reduction for classification purpose," in *Hyperspectral Data Exploitation: Theory and Applications*, C.-I. Chang, Ed. Hoboken, NJ: Wiley, 2007, pp. 245–274.
[63] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA: Academic, 1990.
[64] M. V. Ibanez and A. Simo', "Parameter estimation in Markov random field image modeling with imperfect observations. A comparative study," *Pattern Recognit. Lett.*, vol. 24, no. 14, pp. 2377–2389, Oct. 2003.
[65] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *J. R. Stat. Soc. B*, vol. 51, no. 2, pp. 271–279, 1989.
[66] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[67] G. Moser and S. B. Serpico, "Classification of high-resolution images based on MRF fusion and multiscale segmentation," in *Proc. IGARSS*, Boston, MA, 2008, vol. II, pp. 277–280.
[68] G. Moser and S. B. Serpico, "Contextual high-resolution image classification by Markovian data fusion, adaptive texture extraction, and multiscale segmentation," in *Proc. IGARSS*, Vancouver, BC, Canada, 2011, pp. 1155–1158.
[69] M. D. Martino, F. Causa, and S. B. Serpico, "Classification of optical high resolution images in urban environment using spectral and textural information," in *Proc. IGARSS*, Toulouse, France, 2003, pp. 467–469.
[70] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 2002.
[71] B. Scholkopf, A. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

Gabriele Moser (S'03–M'05) received the Laurea (M.Sc. equivalent) degree (*summa cum laude*) in telecommunications engineering and the Ph.D. degree in space sciences and engineering from the University of Genoa, Genoa, Italy, in 2001 and 2005, respectively.

Since 2010, he has been an Assistant Professor of telecommunications at the University of Genoa. Since 2001, he has cooperated with the Signal Processing and Telecommunications research laboratory of the University of Genoa. From January to March 2004, he was a Visiting Student at the *Institut National de Recherche en Informatique et en Automatique*, Sophia Antipolis, France. His research activity is focused on the development of image-processing and pattern-recognition methodologies for remote-sensing data interpretation. His current research interests include contextual classification, multitemporal image classification and change detection, SAR data analysis, hyperspectral image classification, and geo/biophysical parameter estimation.

Dr. Moser has been a Reviewer for several international journals. He has been an Associate Editor of the international journals IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and *Pattern Recognition Letters* since 2008 and 2011, respectively.

Sebastiano B. Serpico (M'87–SM'00–F'08) received the Laurea degree in electronic engineering and the Doctorate degree from the University of Genoa, Genoa, Italy, in 1982 and 1989, respectively.

He is a Full Professor of telecommunications at the Faculty of Engineering of the University of Genoa and the Head of the Signal Processing and Telecommunications laboratory of the Department of Telecommunications, Electronic, Electrical, and Naval Engineering of the University of Genoa. His current research interests include pattern recognition for remote sensing images and for biomedical images. He is the Chairman of the Institute of Advanced Studies in Information and Communication Technologies. He has been the Project Manager of numerous research contracts and an Evaluator of project proposals for various programs of the European Union. He is an author (or coauthor) of more than 200 scientific articles published in journals and conference proceedings.

Dr. Serpico is an Associate Editor of the international journal IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS). He was a Guest Editor of two special issues of TGRS on the subject of the analysis of hyperspectral image data (July 2001 issue) and on the subject "Advances in techniques for the analysis of remote sensing data" (March 2005 issue). From 1998 to 2002, he was the chairman of a SPIE/EUROPTO series of conferences on signal and image processing for remote sensing.