A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images

Lorenzo Bruzzone, *Senior Member, IEEE*, and Lorenzo Carlin, *Student Member, IEEE*

***Abstract*—This paper proposes a novel pixel-based system for the supervised classification of very high geometrical (spatial) resolution images. This system is aimed at obtaining accurate and reliable maps both by preserving the geometrical details in the images and by properly considering the spatial-context information. It is made up of two main blocks: 1) a novel feature-extraction block that, extending and developing some concepts previously presented in the literature, adaptively models the spatial context of each pixel according to a complete hierarchical multilevel representation of the scene and 2) a classifier, based on support vector machines (SVMs), capable of analyzing hyperdimensional feature spaces. The choice of adopting an SVM-based classification architecture is motivated by the potentially large number of parameters derived from the contextual feature-extraction stage. Experimental results and comparisons with a standard technique developed for the analysis of very high spatial resolution images confirm the effectiveness of the proposed system.**

***Index Terms*—Hierarchical feature extraction, hierarchical segmentation, multilevel and multiscale analysis, spatial-context information, support vector machines (SVMs), very high spatial resolution images.**

I. Introduction

ONE of the most challenging problems addressed by the remote sensing community in current years is the development of effective data processing techniques for images acquired with the last generation of very high spatial resolution sensors. The development of these kinds of techniques appears even more important in light of recently launched commercial satellites (e.g., Ikonos and Quickbird), with on-board sensors characterized by very high geometrical resolution (from 2.5 to 0.60 m). The availability of images acquired by these sensors leads to a new set of possible applications, which require mapping the Earth surface both with great geometrical precision and a high level of thematic detail. In this context, great attention is devoted to the analysis of urban scenes, with applications such as road network extraction and road map updating, transportation infrastructure management, the monitoring of growth in urban areas, and discovering building abuse [1], [2]. Other applications are related to the monitoring of forests, like the definition of selective cutting planning and the analysis of forest status health [3], [4]. In addition, high-resolution remote sensing images can be used by public administrations to monitor, manage, and prevent natural disasters, to analyze evacuation planning in areas with high probability of floods or fires [5], etc. However, these are only a few examples of the wide range of potential applications of high geometrical resolution data.

The significant amount of geometrical details present in a high-resolution scene completely changes the perspective of data analysis compared with moderate-resolution images provided by previous-generation multispectral sensors [such as the Thematic Mapper (TM) and Enhanced Thematic Mapper Plus (ETM+)]. In particular, the improvement in spatial resolution simplifies the problem of mixed pixels[1] present in standard multispectral images, but at the same time, it increases the internal spectral variability (intraclass variability) of each land-cover class and decreases the spectral variability between different classes (interclass variability). Thus, on the one hand, the resulting high intraclass and low interclass variabilities lead to a reduction in the statistical separability of the different land-cover classes in the spectral domain, which in turn involves high classification errors [6], [7]. In addition, the limited spectral resolution of very high geometrical resolution sensors, which depends on technical constraints, further increases the complexity of the classification problem [6], [19]. On the other hand, due to the high spatial resolution of the images, the geometrical information of the scene can also be considered in the classification process according to proper feature-extraction methodologies.

In the recent literature, many papers have addressed the development of novel techniques for the classification of high-resolution remote sensing images. In [9], the authors present a technique for the identification of land developments across large-scale regions. The proposed technique uses straight lines, statistical measures (length, orientation, and periodicity of straight line), and a spatial coherence constraint to identify three classes, namely: 1) urban; 2) residential; and 3) rural. In [10], a standard maximum-likelihood classifier is used to discriminate four spectrally similar macroclasses. Subsequently, each macroclass can be hierarchically subdivided according to class-dependent spatial features and a fuzzy classifier. The main problem of these techniques is that they are highly problem dependent. This means that they cannot be considered as a general operational tool. In [11], the authors analyze the effectiveness of the gray-level cooccurrence matrix (GLCM) texture features in modeling the spatial context that characterizes high-resolution images. However, the fact that the analysis

Manuscript received July 11, 2005; revised March 7, 2006. This work was supported by the Italian Ministry of Education, University and Research (MIUR).

The authors are with the Department of Information and Communication Technology, University of Trento, 38050 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it).

Digital Object Identifier 10.1109/TGRS.2006.875360

[1] Mixed pixels are pixels that represent the spectral signature of more than one class due to the insufficient geometrical resolution of the sensor (more than one land-cover class is included in the ground-projected instantaneous field of view (GIFOV) of the sensor).

depends on a square window and different heuristic parameters and the intrinsic inability to model the shape of the objects do not yield satisfactory classification accuracies.

A more promising family of approaches to the analysis of high spatial resolution images, which is inspired by the behavior of the human view system, is based on object-oriented analysis and/or multilevel/multiscale strategies. The rationale of these approaches is that each image is made up of interrelated objects of different shapes and sizes. Therefore, each object can be modeled both with shape and topological measures which can be used and integrated with spectral features to improve the classification accuracy. Objects can be extracted from images according to one of the standard segmentation techniques proposed in the literature [6], [12]. In greater detail, the main idea of multilevel analysis is that for each level of detail, it is possible to identify different objects that are peculiar to the considered level and that should not appear in other levels. In other words, each object can be analyzed at its "optimal" representation level. Moreover, other aspects considered in this analysis are: 1) that objects at the same level are logically related to each other and 2) that each object at a generic level is hierarchically related to those at the higher and lower levels [7], [8], [13], [14]. For example, in the multiscale analysis of a high-resolution image, at finer levels, we can identify houses, gardens, streets, and single trees; at higher levels, we can identify urban aggregates, groups of trees, and agricultural fields; finally, at the coarser level, we can identify towns and cities, forests, and agricultural areas as one single object. The exploration of the hierarchical tree results in a precise analysis of the relations of objects. For example, we can count the number of houses that belong to an urban area [13].

In [15], the authors propose an approach based on the analysis of a high-resolution scene through a set of concentric windows. The concentric windows analyze the pixel under investigation and the effects of its neighbor system at different scales of resolution. To reduce the computational burden, the information contained in each analysis window is compacted using a Gaussian pyramidal resampling approach. The classification task is accomplished by a soft multilayer perceptron neural network that can be used adaptively as a pixel-based or an area-based classifier. One of the limitations of this approach is the fixed shape and choice of size of the analysis window. In [16], an object-based approach is proposed for classification of dense urban areas from pan-sharpened multispectral Ikonos imagery. This approach exploits a cascade combination of a fuzzy pixel-based classifier and a fuzzy object-based classifier. The fuzzy pixel-based classifier uses spectral and simple spatial features to discriminate between roads and buildings, which are spectrally similar. Subsequently, a segmented image is used to model the spectral and spatial heterogeneities and to improve the overall accuracy of the pixel-based thematic map. Shape features and other spatial features (extracted from the segmented image) as well as the previously generated fuzzy classification map are used as inputs to an object-based fuzzy classifier. In [17], morphological operators (such as opening and closing) are exploited within a multiscale approach to provide image structural information for the automatic recognition of man-made structures. In greater detail, the structural information is obtained by applying morphological operators with a multiscale approach and analyzing the residual images obtained as a difference between the multiscale morphological images at successive scales. A potential problem of this technique is the large feature space generated by the application of a series of opening and closing transforms. In [17], the authors overcome this problem by proposing the use of different feature-selection algorithms. An adaptive and supervised model for object recognition is presented in [7], where a scale-space filtering process that models a multiscale analysis for feature extraction is integrated in a unified framework within a multilayer perceptron neural network. This means that the error backpropagation algorithm used to train the neural network also identifies the most adequate filter parameters. The main problems of this technique are related to the choice of the number and type of filters to be used in the input filtering layer (first layer) of the neural network. In [18], an algorithm based on selective region growing is proposed to classify a high-resolution image. In the first step, the image is classified by taking into account only spectral information. In the second step, a classification procedure is applied to the previous map by taking into account not only spectral information but also a pixel distance condition to aggregate neighbor pixels. By reiteration, neighbor pixels that belong to the same class grow in a selective way, obtaining a final classification map.

Nevertheless, at present, the few techniques specifically developed for the automatic analysis of high spatial resolution images (compared with the very large literature on the classification of moderate-resolution sensors) do not exhibit sufficient accuracy to meet end-user requirements in all application domains. For this reason, it is important that the remote sensing community invests further efforts to define advanced effective methods for the classification of the aforementioned type of data.

In this paper, we propose a novel pixel-based approach to the classification of very high spatial resolution images, which is based on two modules (see Fig. 1): 1) a feature-extraction module that exploits an adaptive, multilevel, and complete hierarchical representation of the spatial context of each pixel in the scene under investigation and 2) a classification module based on support vector machines (SVMs). In greater detail, extending and developing concepts previously presented in the literature, a strategy for defining the spatial context of a pixel at different levels in an adaptive way is presented. The multilevel spatial-context information is then used to drive the feature-extraction phase. The resulting high-dimensional feature vectors are then analyzed according to a proper SVM-based multiclass architecture. The choice of the SVM depends on the effectiveness of this machine-learning methodology to manage classification problems in hyperdimensional feature spaces [21], [22]. It is worth noting that the contribution of this work concerning the importance of SVM in the classification of very high resolution images goes beyond the specific methodologies presented in this paper because the classification of high-resolution images generally requires the analysis of hyperdimensional feature vectors (e.g., when multiscale morphological filters are used, we can obtain a large feature set) and, thus, the exploitation of a classification technique robust

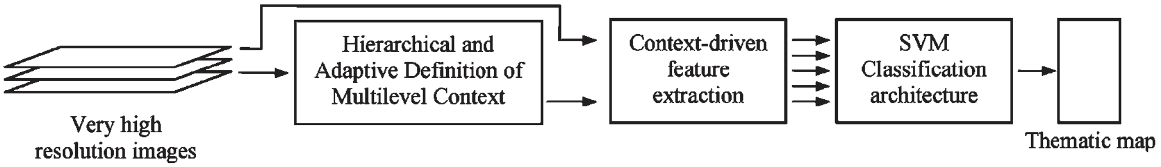

Very high resolution images
Hierarchical and Adaptive Definition of Multilevel Context
Context-driven feature extraction
SVM Classification architecture
Thematic map

Fig. 1. Block scheme of the proposed approach.

to the Hughes phenomenon. Unlike other methods presented in the literature, the proposed approach is general[2] and can be applied to any kind of very high geometrical resolution image.

Experimental results, obtained on two different data sets made up of very high spatial resolution images acquired by the Quickbird satellite in significantly different scenes (i.e., urban and rural areas), point out the effectiveness of the proposed system.

This paper is organized in five sections. Section II presents a detailed description of the proposed adaptive multilevel context-driven feature-extraction technique. Section III addresses the classification module and describes the adopted SVM-based classification architecture. Section IV presents the data sets used for the experiments and reports on experimental results. Finally, Section V provides a discussion on the proposed approach and draws the conclusion of this paper.

II. Proposed Adaptive Multilevel Context-Driven Feature-Extraction Technique

The rationale of the proposed feature-extraction technique consists of adaptively modeling the spatial context of each pixel according to a multilevel strategy. Each context level is defined according to predefined spectral and spatial constraints.

A. Adaptive Definition of the Multilevel Spatial Context

To adaptively characterize the spatial context of each pixel by taking into account a complete hierarchical multiscale context representation, extending and developing some concepts previously presented in the literature, we propose to decompose the scene under investigation from the pixel level to the highest levels of representation of its spatial context. A complete hierarchical modeling allows to capture and exploit the entire information present in the scene by working with adaptive context/neighborhood systems at different scales. This task is based on the application of a segmentation technique with a set of properly defined parameters that take into account both spectral and spatial constraints. This decomposition results in a multilevel representation of the spatial context of each pixel in the investigated scene. To satisfy a tree-based hierarchical requirement, this process is accomplished according to a specific set of rules. In this way, precise hierarchical relationships between each pixel in the image and the regions that adaptively define its context at different levels are established. In other words, we obtain a set of segmentation maps (one for each level) that characterize the context of each spatial position in the image hierarchically and in a nonambiguous way (Fig. 2).

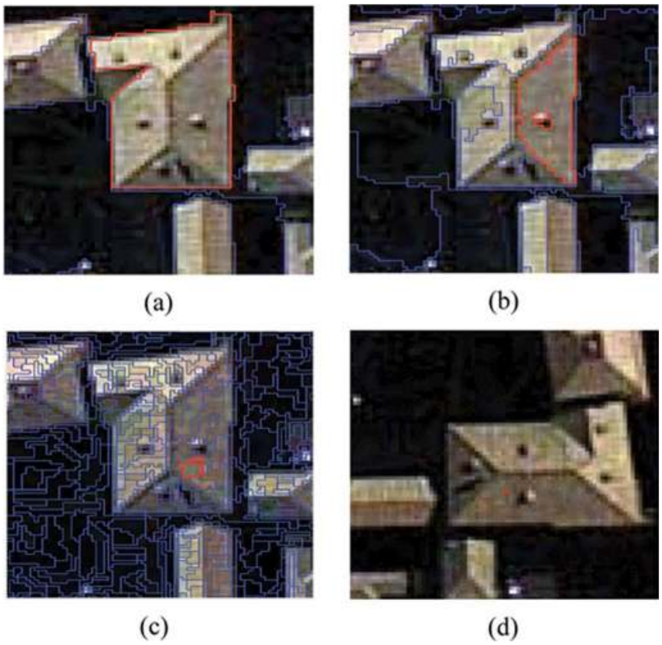


Fig. 2. Hierarchical multilevel segmentation applied to a multispectral Quickbird image. From (a) to (c), we use three different sets of parameters in the segmentation algorithm to adaptively model (at different levels) the context of the pixels in the image in (d). The selected rule guarantees that precise hierarchical relations between different levels are established.

It is worth noting that, unlike other approaches proposed in the literature and briefly described in Section I, hierarchical segmentation does not aim to identify the best level of representation of each object, but simply models the multilevel spatial context of each pixel. This should be considered as a preprocessing stage aimed at driving the feature-extraction phase.

A formal definition of the adopted segmentation procedure is given in the following. Let I denote the investigated image and H^l the homogeneity predicate at the generic level l $(l = 1, \ldots, L)$. Varying the homogeneity predicate means varying the level of definition of the adaptive spatial context of the pixel. This homogeneity predicate is defined according to different spatial and spectral attributes at different levels. According to the literature, the segmentation of I at a generic level l is a partition P^l in a set of N^l regions O_i^l $(i = 1, 2, \ldots, N^l)$,

[2] It is worth noting that, in this paper, the words "general" and "problem independent" mean that the proposed technique has not *a priori* constraints on the kinds of objects present in the scene, but it can be used with any kind of high-resolution image and in any application domain. On the contrary, many techniques proposed in the literature are not general and are problem dependent as they are specifically developed for addressing particular applications (e.g., analysis of urban areas) and are based on feature-extraction procedures and processing algorithms that cannot be applied to other scenes.

such that

$$\bigcup_{i=1}^{N^l} O_i^l = I \text{ with } O_n^l \cap O_m^l = \phi, \qquad n \neq m \quad (1)$$

$$H^l\left(O_i^l\right) = \text{true} \quad \forall i \quad (2)$$

$$H^l\left(O_n^l \cup O_m^l\right) = \text{false} \quad \forall O_n^l \text{ and } O_m^l \text{ adjacent.} \quad (3)$$

These three rules are valid for objects at a generic level l. To establish a precise hierarchy between the contexts of a pixel defined at different levels, we consider the following additional constraint:

$$\bigcup_{O_i^{l-1} \subseteq O_j^l} O_i^{l-1} = O_j^l. \quad (4)$$

This simple relation states that the adaptive neighborhood of a pixel at level $l-1$ cannot be included in more than one adaptive neighborhood at level l (it has only one father node). It is worth noting that level 1 represents the pixel level, i.e., the pixel for which the context is hierarchically defined.

We would like to stress that the idea of using hierarchical segmentation to represent the objects that compose the scene at different levels of abstraction is not a new contribution of this paper, but the novelty of this paper consists in the technique adopted for exploiting the results of the hierarchical segmentation. In this respect, any segmentation algorithm that satisfies the aforementioned constraints can be used in the proposed system (see, for example, [13] and [28]).

The multiresolution segmentation algorithm we adopted is a bottom-up region-merging technique starting from the pixel level (at the first step, each pixel represents an object). In an iterative way, at each subsequent step, image objects are merged into bigger ones. The aim of this procedure is to minimize the homogeneity predicate when two different objects are merged together (this constraint must be valid for all the couple of objects in the image). If the smallest growth exceeds a threshold defined by the user (the so-called "scale parameter"), the process stops. As briefly mentioned before, the homogeneity predicate takes into account spectral and spatial constraints. In detail, it can be defined as follows:

$$H^l\left(O_i^l\right) = w_{\text{spectral}}^l \cdot h_{\text{spectral}}^l\left(O_i^l\right) + w_{\text{shape}}^l \cdot h_{\text{shape}}^l\left(O_i^l\right) \quad (5)$$

where $w_{\text{spectral}}^l \in [0,\ldots,1]$, $w_{\text{shape}}^l \in [0,\ldots,1]$ are user-defined parameters and $w_{\text{shape}}^l = 1 - w_{\text{spectral}}^l$. The first part of (5) is a cost criterion for the spectral component of the objects, whereas the second part is a shape cost criterion. Hence, we can define an information loss function when two closest objects O_i^l and O_j^l, at a certain level l, are fused together as

$$\begin{aligned} C_{i,j}^l &= H\left(O_i^l \cup O_j^l\right) - H\left(O_i^l\right) - H\left(O_j^l\right) \\ &= w_{\text{spectral}}^l \cdot C_{i,j,\text{spectral}}^l + \left(1 - w_{\text{spectral}}^l\right) \cdot C_{i,j,\text{shape}}^l. \end{aligned} \quad (6)$$

We can stop the segmentation algorithm when $C_{i,j}^l \leq H_{\text{TH}}^l$, where H_{TH}^l is a level-dependent user-defined threshold defined for each level. The greater the value of H_{TH}^l, the greater the dimension of obtained objects. (In other words, we decrease the sensibility of the homogeneity predicate in the fusion of two adjacent regions.) The spectral information loss function $C_{i,j,\text{spectral}}^l$ in (6) can be defined as

$$C_{i,j,\text{spectral}}^l = \sum_{d=1}^{B} w^{l,d} \left[N_{i,j}^l \cdot \sigma_{i,j}^{l,d} - \left(N_i^l \cdot \sigma_i^{l,d} + N_j^l \cdot \sigma_j^{l,d}\right)\right] \quad (7)$$

where $w^{l,d}$, $d = 1,\ldots,B$ (B is the number of spectral bands), represents the weight associated to the dth spectral channel at level l in the combination process, and $\sum_{d=1}^{B} w^{l,d} = 1$; $N_{i,j}^l$ represents the number of pixels of the object obtained by merging O_i^l and O_j^l, and $\sigma_{i,j}^{l,d}$ represents its standard deviation on the spectral band d. N_i^l and N_j^l represent the number of pixels that compose objects O_i^l and O_j^l, respectively, and $\sigma_i^{l,d}$ and $\sigma_j^{l,d}$ represent their standard deviations calculated on the spectral band d.

The spatial information loss function, $C_{i,j,\text{shape}}^l$ in (6), takes into account the compactness and smoothness of the shape of the obtained object by merging O_i^l and O_j^l. It is defined as

$$C_{i,j,\text{shape}}^l = w_{\text{cmp}}^l \cdot C_{i,j,\text{cmp}}^l + \left(1 - w_{\text{cmp}}^l\right) \cdot C_{i,j,\text{smooth}}^l \quad (8)$$

where

$$C_{i,j,\text{cmp}}^l = N_{i,j}^l \cdot \frac{e_{i,j}^l}{\sqrt{N_{i,j}^l}} - N_i^l \cdot \frac{e_i^l}{\sqrt{N_i^l}} - N_j^l \cdot \frac{e_j^l}{\sqrt{N_j^l}} \quad (9)$$

$$C_{i,j,\text{smooth}}^l = N_{i,j}^l \cdot \frac{e_{i,j}^l}{r_{i,j}^l} - N_i^l \cdot \frac{e_i^l}{r_i^l} - N_j^l \cdot \frac{e_j^l}{r_j^l} \quad (10)$$

where $w_{\text{cmp}}^l \in [0,\ldots,1]$ is a user-defined parameter to weight the smoothness of the obtained objects with respect to the compactness; $e_{i,j}^l$ represents the perimeter of the object obtained by merging O_i^l and O_j^l, whereas $r_{i,j}^l$ represents the perimeter of the rectangle containing it; e_i^l and e_j^l represent the perimeter of the objects O_i^l and O_j^l, respectively, whereas r_i^l and r_j^l represent the perimeter of the rectangles that contain O_i^l and O_j^l, respectively. It is worth noting that the basic criteria of the aforementioned segmentation strategy are also implemented in commercial software packages [28].

The choice of the range of variation of the parameters defining the homogeneity criterion affects the number of levels[3] in which the scene is decomposed. The number of levels to be used for characterizing the spatial context of each pixel depends on many factors. The most important issues to take into account are: 1) geometrical resolution of the image [e.g., given a specific scene, Quickbird images (GIFOV equal to 0.6 m) require higher numbers of decomposition levels than SPOT 5 images (GIFOV equal to 2.5 m)] and 2) size of the objects present in the scene. An empirical rule, for obtaining indications on the number of levels to use, consists in computing the mean size of the regions at different decomposition levels and comparing this size with

[3] In this paper, we refer to a multilevel representation of the scene and not to a multiscale decomposition. This is due to the use of a segmentation procedure, which is driven not only on geometrical criteria but also on spectral parameters, for accomplishing the image decomposition task.

the expected average size of the objects in the scene, which can be defined by the end user. In greater detail, all the levels in the decomposition have to satisfy the following condition:

$$\frac{1}{N^l}\sum_{i=1}^{N^l} A_i^l \le \mathrm{EA}_{\mathrm{th}} \tag{11}$$

where N^l is the number of objects at level l, A_i^l represents the area of object i at level l (in pixels), and $\mathrm{EA}_{\mathrm{th}}$ is a user-defined parameter that corresponds to the expected average size of the objects in the scene (in pixels). It is also possible to define the quantity $\mathrm{EA}_{\mathrm{th,m^2}} = \mathrm{EA}_{\mathrm{th}} \cdot \mathrm{GIFOV}^2$, which represents the expected average size of the objects in square meters. Accordingly, (11) can be rewritten as

$$\left[\frac{1}{N^l}\sum_{i=1}^{N^l} A_i^l\right] \cdot \mathrm{GIFOV}^2 \le \mathrm{EA}_{\mathrm{th,m^2}}. \tag{12}$$

The definition of the value of $\mathrm{EA}_{\mathrm{th}}$ (or $\mathrm{EA}_{\mathrm{th,m^2}}$) is relatively easy in homogeneous scenes (e.g., urban areas). In heterogeneous scenes, the problem can be addressed according to two different strategies: 1) consider a tradeoff between the average sizes of different classes of objects and 2) select the aforementioned parameters according to the average size of the greatest class of objects (implicitly assuming the use of context information including more objects for the smallest components of the scene). Both strategies are consistent with the proposed approach. The choice of one of them should be based on end-user requirements.

It is worth noting that, in the discussion above, we considered the pixel level as the lowest level of the hierarchy, but it is also possible to define any level of the multiscale segmentation maps as the lowest level of the tree (in the latter case, we consider an object and its adaptive context).

B. Multilevel Context-Driven Feature Extraction

Given the hierarchical tree structure, it is then possible to exploit the relationships between pixels and regions at different levels to extract an effective set of features that describe each pixel and its adaptive context at each level. Depending on the level considered, different kinds of features can be extracted to characterize the spatial context with the most reliable attributes for the specific analyzed "scale." We can extract spectral, spatial, or relational features. Spectral features are derived analyzing directly the spectral information of a pixel and that of its adaptive neighborhood at different levels. Simple spectral features (such as mean and standard deviation) or more complex measures (such as entropy and high-order statistics) can be easily extracted to characterize both space-invariant and texture properties associated with the pixel. In addition, geometrical and relational measures can be computed to characterize the shape, size, and interrelation of the adaptive neighborhood of a pixel. In greater detail, geometrical features are related to the description of the shape and size of the spatial context at different levels of analysis (e.g., we can compute the area, the shape factor, and the perimeter of a generic region m at level l).[4] Concerning relational parameters, they can be expressed by a contextual analysis of neighboring regions at the same level or at different levels to model the relation between the spatial context of a pixel at the same or at different levels. Thus, we can define the feature vector $\underline{x}_i$, which describes the pixels and, through the hierarchical tree, the spatial context (objects) in which the pixel is included.

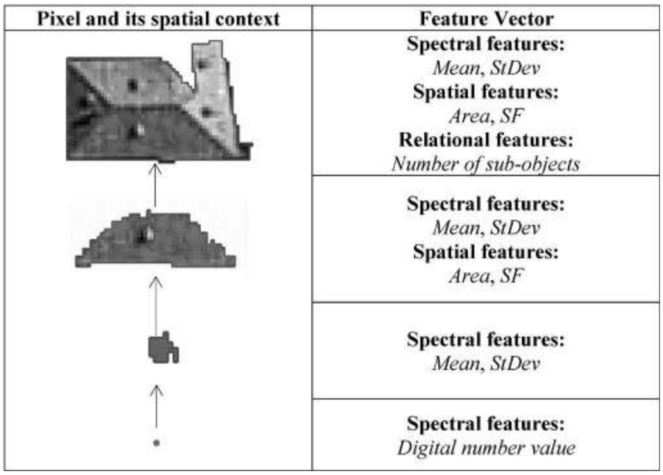


Fig. 3. Example of the features extracted for objects that, at different levels, characterize the context of the pixel under investigation. "Mean" and "StDev" represent the mean value and the standard deviation of pixels in a generic object, respectively. "Area" and "SF" represent the area and the shape factor (the ratio between length and width) of an object, respectively. "Number of sub-objects" represents the number of objects at level $l-1$ that make up an object at level l.

For a generic pixel i under analysis, we can write

$$\underline{x}_i = \left\{\underline{f}_1^i, \underline{f}_2^i, \ldots, \underline{f}_l^i, \ldots, \underline{f}_L^i\right\} \tag{13}$$

where $\underline{f}_l^i$ is the feature vector associated with the contextual information of pixel i at generic level l of the hierarchical tree, and L is the number of segmentation levels. It is worth noting that the components of $\underline{f}_1^i$ are the features that characterize the spatial position i in the image at pixel level. The subvector $\underline{f}_l^i$ is defined as

$$\underline{f}_l^i = \left\{f_{l,1}^i, \ldots, f_{l,j}^i, \ldots, f_{l,\mathrm{NF}^l}^i\right\} \tag{14}$$

where $f_{l,j}^i$ is the jth feature that models the context of pixel i at level l, and NF^l is the number of features extracted at level l.

As stated before, the component $f_{l,j}^i$ can be a spectral, a geometrical, or a relational feature. An important observation concerns the criterion to adopt for defining the set of features to be used at each level. As shown in the example reported in Fig. 3, at the pixel level, it is possible to use only the pixel spectral signature (there are no regions, and hence, it is not possible to compute any geometrical feature). At intermediate levels, the regions are typically small and represent only portions of the objects; thus, we recommend avoiding the use of geometrical features, as they do not contain relevant information about the geometry of the true objects present in the scene. At the higher levels, instead, the objects are better modeled by the

[4] Examples of these parameters are reported in Section IV-B.

segmentation algorithm, and geometrical and relational features can be properly used.

It is worth noting that the main ideas of the proposed approach are that: 1) all the segmentation levels, which should be selected according to the aforementioned general guidelines, can be used to obtain a complete hierarchical representation of the spatial context of each pixel and 2) all the features extracted to characterize the context information of each pixel at different levels can be used as input to the classification module. However, although this approach provides the classifier with a large amount of information, it has the disadvantage of leading to a very high dimensional (hyperdimensional) feature vector. This problem should be addressed in the classification module.

III. SVM Classification Approach

To achieve a good characterization of the spatial context of each pixel, we should use a sufficient number of segmentation levels. [The number of levels depends on the scene under analysis; see the criterion defined in (11).] As mentioned earlier, from the adaptive neighborhood of a pixel at each single level l, we can extract a large number of features that characterize the spectral, geometrical, and relational attributes of the regions. Hence, the number of components of the feature vector extracted from the hierarchical tree may be very high.

To obtain proper learning of the classifier and to achieve a good generalization capability while avoiding the course of dimensionality problem (the so-called Hughes phenomenon due to the small ratio between the number of training samples and the number of features [26]), we should collect a large number of independent training set samples to characterize all the possible spectral variations of each land-cover class. Although it is quite simple to collect ground truth samples by photo interpretation on very high resolution images, it is rather time consuming. In addition, the spatial autocorrelation of each sample reduces the spectral information of the neighboring samples and violates the sample-independent condition. This can lead to the so-called unrepresentative sample problem [24] that increases the complexity in the definition of training samples.

The following two possible alternatives to the problem of collecting a very large number of training samples can be considered: 1) applying a feature-selection procedure and 2) using a classifier intrinsically robust to the Hughes phenomenon. Concerning feature selection, in the considered problem, it is quite difficult to define a criterion function (aimed at evaluating the effectiveness of the considered subset of features) capable of dealing with the heterogeneity of the statistical models that characterize the different parameters extracted in the previous phase. Many feature-selection techniques assume Gaussian (or monomodal) distributions for the analyzed features, which do not fit some of the considered measures. For this reason, in the proposed approach, we prefer to avoid feature selection and to adopt a classification technique intrinsically less sensitive to the high dimensionality of the feature space. In particular, we consider a machine-learning classifier based on SVMs, which have been recently proved to be effective in hyperdimensional problems [21], [22].

Developed by Vapnik, SVMs are based on the structural risk minimization principle [27], and their popularity within the remote sensing community is constantly on the increase [20], due to their properties and intrinsic effectiveness. In the following, we briefly describe the main concepts of the mathematical formulation of SVMs for binary classification problems.

Let us consider a binary classification problem, with N training patterns in a d-dimensional feature space. Each pattern is associated with a target $y_i \in \{+1, -1\}$. The nonlinear SVM approach consists of mapping the data into a higher dimensional feature space, i.e., $\Phi(\underline{x})$, where a separation between the two classes is looked for by means of an optimal hyperplane defined by a weight vector $\underline{w}$ and a bias b. The decision rule is defined by the function $\mathrm{sign}[f(\underline{x})]$, where $f(\underline{x})$ represents the discriminant function of the hyperplane and is defined as

$$f(\underline{x}) = \underline{w} \cdot \Phi(\underline{x}) + b. \tag{15}$$

The optimal hyperplane is the one that minimizes a cost function that expresses a combination of two criteria, namely: 1) margin maximization and 2) error on training samples minimization. It is defined as

$$\Psi(\underline{w}, \xi) = \frac{1}{2}\|\underline{w}\| + C\sum_{i=1}^{N} \xi_i \tag{16}$$

and it is subject to the following constraints:

$$\begin{cases} y_i\left(\underline{w} \cdot \Phi(\underline{x}) + b\right) \geq 1 - \xi_i, & i = 1, 2, \ldots, N \\ \xi_i \geq 0, & i = 1, 2, \ldots, N \end{cases} \tag{17}$$

where ξ_i are called slack variables and are introduced to take into account nonseparable data. The constant C represents a regularization parameter that allows to tune the shape of the discriminant function. The above minimization problem can be reformulated through a Langrage functional for which the Lagrange multipliers can be found by means of a dual optimization leading to a quadratic programming solution. The final result is a discriminant function described (in the original feature space) by the following equation:

$$f(\underline{x}) = \sum_{i \in S} \alpha_i y_i K(\underline{x}_i, \underline{x}) + b \tag{18}$$

where $K(.,.)$ is a kernel function that should satisfy the Mercer's theorem. The set S is a subset of the indices $\{1, 2, \ldots, N\}$ corresponding to the nonzero Lagrange multipliers α_i. The training vectors associated with these multipliers are called support vectors. The solution of the dual-optimization problem avoids the problem of defining optimal transformation from the original to the hyperdimensional feature space. The most widely used kernel functions adopted in the remote sensing problems are

$$Polynomial\ kernel$$

$$K(\underline{x}_i, \underline{x}_j) = (\underline{x}_i \cdot \underline{x}_j + 1)^d \tag{19}$$

$$Radial\ basis\ function\ (RBF)\ kernel$$

$$K(\underline{x}_i, \underline{x}_j) = \exp\left(-\gamma\|\underline{x}_i - \underline{x}_j\|^2\right) \tag{20}$$

where d is the order of the polynomial kernel function, and γ is the spread of the RBF kernel.

To solve the multiclass problem, we propose to define an architecture made up of as many binary SVMs as the number of information classes. Each single SVM solves a one-against-all problem [20]. In greater detail, let $\Omega = \{\omega_1, \ldots, \omega_c\}$ be the set of information classes that characterize the considered problem. The ith SVM solves a binary problem between classes $\omega_A = \omega_i$ and $\omega_B = \Omega - \omega_i$ $(\omega_i \in \Omega)$. A generic pattern $\underline{x}$ is labeled according to a winner-takes-all rule, i.e.,

$$x \in \omega^* \Leftrightarrow \omega^* = \arg \max_{i=1,\ldots,C} \{f_i(\underline{x})\} \tag{21}$$

where $f_i(\underline{x})$ is the output of the ith SVM. However, other multiclass strategies could be considered (see [22]). We refer the reader to [20], [22], [23], and [27] for greater detail on SVMs.

IV. Experimental Results

To assess the effectiveness of the proposed approach, two different sets of experiments were conducted on two different data sets composed of Quickbird satellite images. The first data set represents a complex urban scene related to the city of Pavia (Italy), whereas the second data set represents a rural area close to the city of Trento (Italy).

A. Pavia Data Set: Urban Area

The image used in the experiments refers to the downtown area of the city of Pavia (northern Italy) and was acquired on June 23, 2002 from the Quickbird satellite. In particular, we used a panchromatic image (Fig. 4) and a pan-sharpened multispectral image obtained by applying a proper fusion technique to the panchromatic channel and the four bands of the multispectral image. The adopted technique is based on the Gram–Schmidt procedure implemented in the ENVI software package [25]. The final data set is made up of a panchromatic image and four pan-sharpened multispectral images of 1024 × 1024 pixels with a spatial resolution of 0.7 m. It is worth noting that the multiresolution fusion task artificially increases the spatial resolution of the multispectral channels on the one hand, whereas on the other, it may affect the spectral signatures of pixels. Nevertheless, this process was used both with the proposed method and with the standard approach adopted for comparison. We therefore do not expect it to affect the assessment of the effectiveness of the proposed approach compared with the standard method.

To assess the effectiveness of the proposed method in challenging classification problems, we define classes in a very detailed way, by considering land covers with similar spectral and/or geometrical attributes (e.g., buildings with different spectral signature). Table I shows the distribution of the samples (in pixels) in the training and test sets among the eight land-cover classes that characterize the considered scene. These samples have been collected by an accurate photo interpretation of the image for training and test samples and according to the following guidelines: 1) they have been extracted from different spatial positions in the image to properly represent classes in different portions of the scene and 2) training and test samples have been selected from different regions to have patterns as more uncorrelated as possible. In addition, unlike in standard accuracy assessment protocols, to better evaluate the performance of the proposed system in both homogeneous and edge (or boundary) areas, we have split the test set samples in two subsets. This allows to better understand the effectiveness of the different classification approaches in dealing with pixels with different properties in the image and results in a more precise accuracy assessment procedure.

To evaluate the effectiveness of the proposed approach, we conducted two different sets of experiments. One was aimed at assessing the effect of the number of context levels (segmentation levels) on classification accuracy. In the other set of experiments, we compared the performances of the proposed

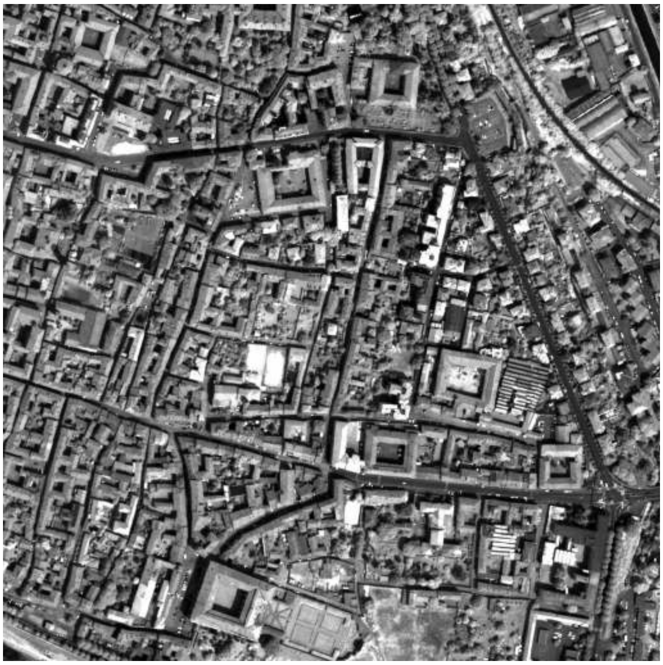


Fig. 4. Panchromatic image (1024 × 1024 pixels) acquired by the Quickbird satellite on the city of Pavia (northern Italy).

TABLE I
NUMBER OF SAMPLES (IN PIXELS) IN THE TRAINING AND TEST SETS (PAVIA DATA SET)

Class	Number of patterns		
	Training set	*Test set on edge areas*	*Test set on homogeneous areas*
Water	180	55	150
Tree areas	348	95	250
Grass areas	323	90	160
Roads	984	182	381
Shadow	750	297	325
Red buildings	2271	442	1040
Gray buildings	602	167	250
White building	275	98	100
TOTAL	5733	1426	2656

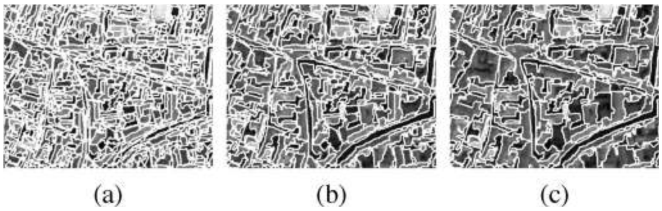


Fig. 5. Representation of the hierarchical modeling of the context at different levels. (a) Four. (b) Five. (c) Six levels of segmentation (Pavia data set).

TABLE II
(a) KAPPA ACCURACY AND (b) OVERALL ACCURACY PROVIDED BY THE PROPOSED APPROACH ON EDGE, HOMOGENEOUS, AND GLOBAL TEST AREAS VERSUS THE NUMBER OF CONSIDERED CONTEXT LEVELS. THE OPTIMAL VALUES OF THE REGULARIZATION PARAMETER C AND OF THE SPREAD γ OF THE KERNEL FUNCTIONS ARE REPORTED (PAVIA DATA SET)

Number of levels (including pixel level)	SVM parameters		Kappa accuracy on Test Sets		
	Regularization parameter (C)	*Spread of kernel function (γ)*	*Edge areas*	*Homogeneous areas*	*Homogeneous and edge areas*
1 - only pixel level	99	9.0	0.563	0.932	0.801
2	550	19.5	0.672	0.954	0.853
3	5	1.6	0.659	0.970	0.861
4	10	1.1	0.736	0.944	0.871
5	138	1.0	0.716	0.960	0.874
6	25	0.4	0.713	0.954	0.870

(a)

Number of levels (including pixel level)	SVM parameters		Overall accuracy (%) on Test Sets		
	Regularization parameter (C)	*Spread of kernel function (γ)*	*Edge areas*	*Homogeneous areas*	*Homogeneous and edge areas*
1 - only pixel level	99	9.0	64.59	94.65	84.15
2	550	19.5	73.35	96.35	88.31
3	5	1.6	72.79	97.67	88.98
4	10	1.1	78.68	95.63	89.71
5	138	1.0	77.14	96.84	89.96
6	24	0.4	76.93	96.39	89.59

(b)

system with those of a standard pixel-based classifier and of an alternative technique based on a classical feature-extraction method based on the generalized Gaussian pyramid decomposition of the image.

1) Experiment 1—Analysis of the Effectiveness of the Proposed Approach: In our experiments, we carried out several trials with hierarchies made up of a different number of levels (up to six levels). The first level is the pixel level, whereas the other five levels are obtained using the presented multiscale hierarchical segmentation technique with different parameters to tune the homogeneity predicate. On the one hand, the levels between two and four are characterized by very small regions. This means that from a general point of view, objects in these levels are highly oversegmented, as shown in Fig. 5(a). In other words, a small neighborhood system is adaptively defined for the pixel. On the other hand, the levels between five and six are characterized by regions of medium size [Fig. 5(b) and (c)]. In particular, at level 6, a single region defining the context of a pixel may contain different objects belonging to different information classes. Although this models the complex context of the object to which the pixel belongs, it may lead to classification errors.

We assessed the effectiveness of the proposed approach versus the number of levels considered (from two to six). The features extracted for the first level (i.e., the pixel level) were only the values of each spectral channel and the panchromatic image. For level 2, we only considered the mean value of the digital numbers of pixels defining each region in each spectral band and the panchromatic image. From levels 3 to 6, for each region and for each band, we considered the mean value and the standard deviation of the digital numbers. On the whole, 10, 20, 30, 40, and 50 features were considered for experiments with two, three, four, five, and six levels, respectively. In the experiments, we used an SVM classifier with RBF kernels, which have been proved effective in a number of different classification problems. According to a proper model selection technique [20], we have identified the best values of parameters (i.e., the regularization parameter C and the spread factor of Gaussian kernels γ) using the training samples and the global test samples (edge and homogeneous areas jointly) for validation. The highest accuracies obtained, as well as the related parameters, are shown in Table II.

These results confirm that the proposed classification system always exhibited a much greater overall accuracy compared with that obtained using only the pixel level. In detail, the greater increase in overall accuracy (i.e., about 13%) obtained with the presented approach relates to edge areas. Classification accuracy increased also on homogeneous areas, although the improvement is significantly smaller (i.e., about 2%) than that in edge areas.

The proposed criterion for the adaptive selection of the number of levels [see (11) and (12)] resulted in the choice of five levels. [The value of $\mathrm{EA}_{\mathrm{th}}$ used in (11) was related to the expected average size of buildings present in the scene.] This confirms the effectiveness of this simple criterion that selected the number of levels that provided the highest classification accuracy on the global test set. It is worth noting that the proposed approach provided stable accuracies for a number of levels close to the one identified by the automatic procedure (in the range between four and six levels), exhibiting Kappa values between 0.71 and 0.74 on edge areas and between 0.94 and 0.96 on homogeneous areas. This confirms its ability to model the spatial context of each analyzed pixel. As one can see from Table II, six context levels lead to a slight decrease of classification accuracies. This behavior is due to the significant undersegmentation of real objects at level 6, which may affect classification accuracy both on edge and homogeneous areas.

In the second part of this experiment, according to the obtained results, we considered only four, five, and six context levels, as they gave the highest classification accuracy. To better evaluate the performance of the proposed classification system, we also analyzed the classification maps obtained in all trials. We report only on a small representative portion of the obtained maps, to present examples that show both the advantages and the limitations of the proposed system.

Fig. 6 shows a small portion of the classification maps obtained with (a) four, (b) five, and (c) six segmentation levels and (d) with only the pixel level.

As can be seen, whereas the crossroad in the center of the images (within the red rectangle) is well modeled in Fig. 6(b) and (c), the results are inaccurate in Fig. 6(a) because of high fragmentation in the modeling of the spatial context at the higher level. In the map obtained using only the pixel level (without contextual information) reported in Fig. 6(d), we can

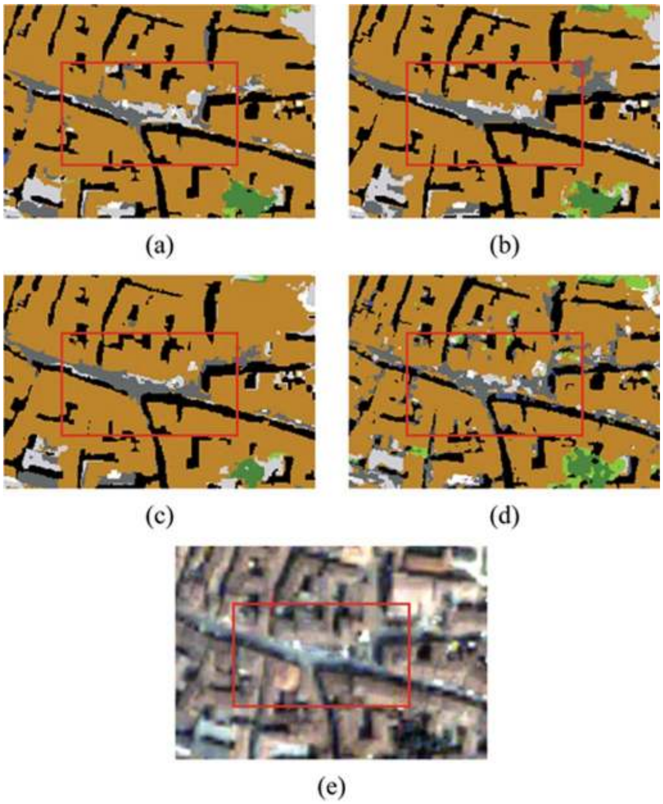


Fig. 6. Detail of classification maps obtained with (a) four, (b) five, and (c) six segmentation levels and (d) with only the pixel level. The real-color Quickbird image is reported in (e). The red rectangle shows an example of the effects of the different levels of the spatial context. The legend of the maps is reported in the caption of Fig. 8 (Pavia data set).

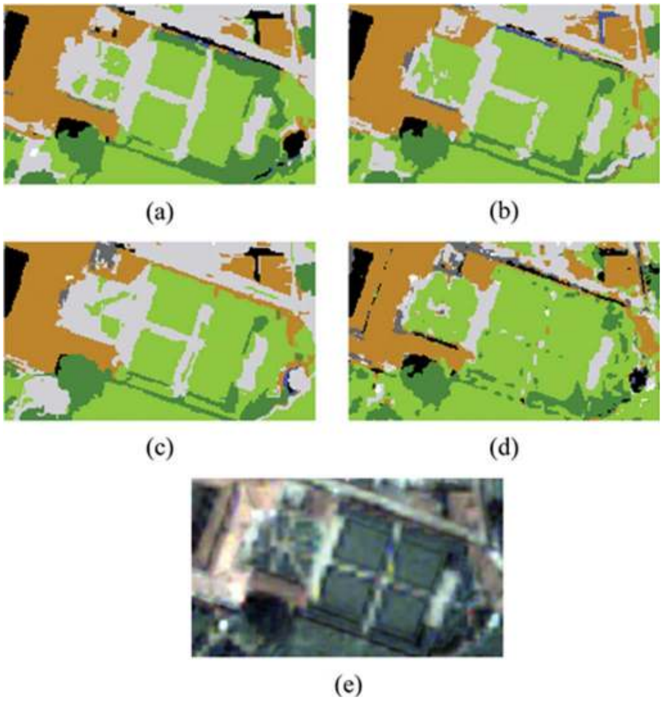


Fig. 7. Classification maps obtained with (a) four, (b) five, and (c) six levels of segmentation and (d) with only the pixel level. The real-color Quickbird image is reported in (e). The legend of the maps is reported in the caption of Fig. 8 (Pavia data set).

see that the shape of the buildings is not well modeled, and in many cases, homogeneous areas are not correctly classified. In addition, the crossroad is not properly recognized.

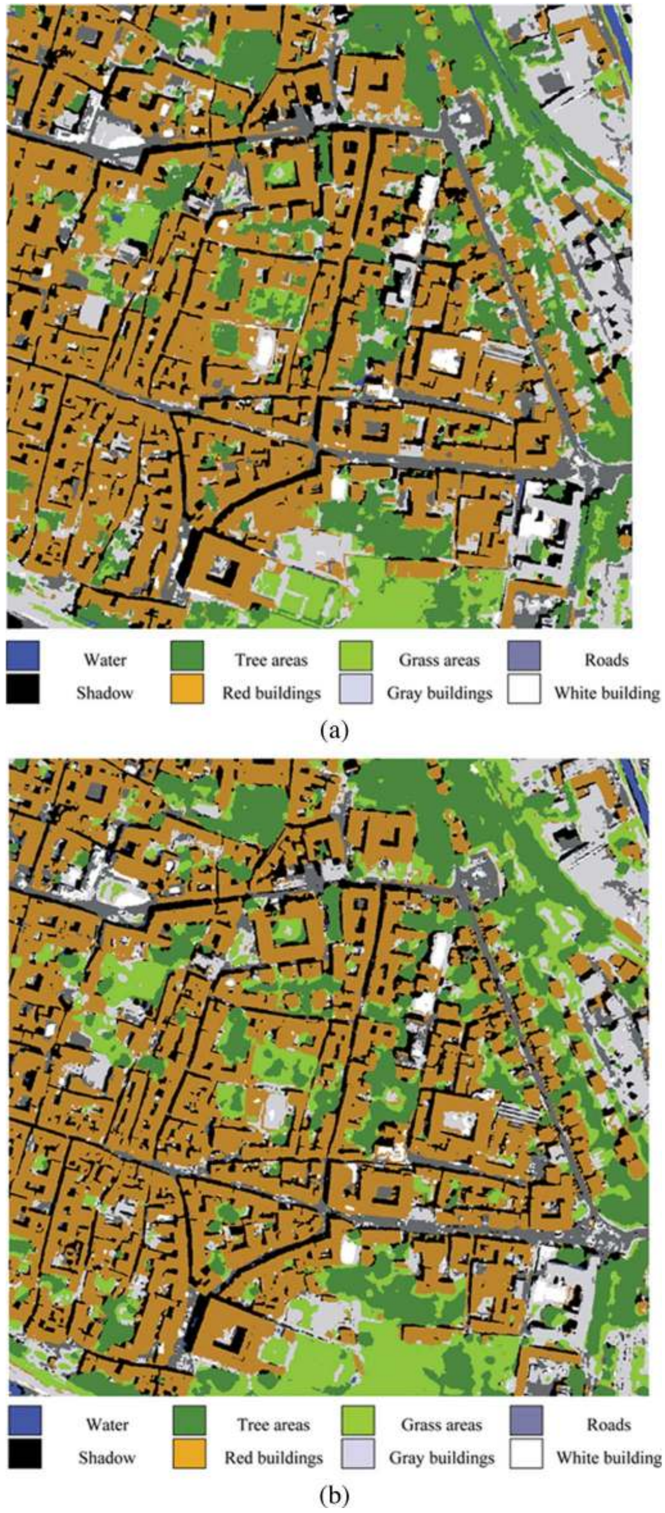
On the other hand, in some areas, using fewer levels (i.e., very small regions to characterize the adaptive neighborhood) leads to a better definition of small details. For example, in Fig. 7(a), small roads are well classified, whereas in Fig. 8(b) and (c), by exploiting more context levels, we obtain a poor representation of objects in the scene under investigation.

It is worth noting that we carried out also some trials by using geometrical (minimum rectangular fit, width-to-length ratio, etc.) and relational (number of neighbors of an object and number of sub-objects that compose an object at the upper level) features. These kinds of features were extracted only for the higher levels of the representation (levels 5 and 6). The obtained results did not improve both the classification accuracies and the quality of the classification maps. This behavior mainly depends on the criterion adopted for the definition of classes. Inasmuch as many classes share the same geometrical features (e.g., different kinds of building are discriminated only on the basis of the spectral signature), in this case, the use of the geometrical and relational information does not increase the separability among classes in the feature space.

2) Experiment 2—Comparisons With a Feature-Extraction Module Based on a Generalized Gaussian Pyramid Decomposition: The aim of the second set of experiments is to compare the proposed system with a different approach to multilevel feature extraction of very high resolution images based on the generalized Gaussian pyramid decomposition. In detail, the panchromatic and pan-sharpened images are iteratively analyzed by a Gaussian kernel low-pass filter, with 5 × 5 square analysis window, and are undersampled by a factor of 2. In this way, it is possible to obtain a simple multiscale decomposition of the scene. In our experiments, we exploit five levels of pyramidal decomposition (this is the number of levels that gives the highest accuracy between two and six) to characterize the spatial context of pixels and to label each pixel of the scene under investigation. The extracted feature vector was made up of 25 spectral features. The SVM classification module was also used in these trials.

The best accuracies obtained for the proposed technique and for the reference feature-extraction technique are reported in terms of Kappa coefficient and overall accuracy in Table III. These results show that the proposed feature-extraction technique provided an accuracy higher than the reference method. The accuracy obtained on test edge areas confirms the greater ability of the proposed approach (which increased Kappa values by 8.5% compared with the generalized Gaussian pyramid method) to model the geometrical details of objects in the scene, such as roofs and roads. A comparison between the accuracies obtained on homogeneous areas points out a gap of 2.4%. To better assess the effectiveness of the investigated methods, Fig. 8(a) and (b) shows the classification maps obtained using the proposed classification system and the reference system.

A qualitative analysis of the maps confirms the previous consideration based on the quantitative results. The adaptive and multilevel properties of the proposed feature-extraction



(a)

(b)

Fig. 8. Classification maps obtained by an SVM classifier (a) with the proposed feature extraction module and (b) with a feature extraction algorithm based on a pyramidal Gaussian decomposition (Pavia data set).

TABLE III
(a) Kappa Accuracy and (b) Overall Accuracy on Edge, Homogeneous, and Global Test Areas Versus the Proposed Feature Extraction Technique (Adaptive Hierarchical Context Modeling) and the Reference Technique (Generalized Gaussian Pyramid Reduction) (Pavia Data Set)

Test set considered	**Feature extraction technique**	
	Generalized Gaussian pyramid	***Proposed multilevel hierarchical decomposition***
Edge areas	0.631	0.716
Homogeneous areas	0.936	0.960
Homogeneous and edge areas	0.828	0.874

(a)

Test set considered	**Feature extraction technique**	
	Generalized Gaussian pyramid	***Proposed multilevel hierarchical decomposition***
Edge areas	70.13	77.14
Homogeneous areas	94.99	96.84
Homogeneous and edge areas	86.31	89.96

(b)

technique can better model the edge of objects in the scene, especially in areas with small details. The great complexity of the analyzed scene, which includes different types of buildings of different sizes and different types of roads, shows that the low-pass filter used in the generalized Gaussian pyramid decomposition is not suitable to model the boundaries of objects and complex structures accurately. On the contrary, when a multilevel segmentation algorithm is used to adaptively model the neighborhood of a pixel, a proper representation of the edges of the objects is obtained.

B. Trento Data Set: Rural Area

The image used in these experiments refers to the rural area of Trento (northern Italy) and was acquired on March 30, 2004 from the Quickbird satellite. The data set consists of a panchromatic image (Fig. 9) and four pan-sharpened images of 512 × 512 pixels with a spatial resolution of 0.7 m. The pan-sharpened images were obtained with the Gram–Schmidt procedure [25].

Table IV shows the distribution of the samples (in pixels) in the training and test sets among the eight land-cover classes that characterize the considered scene. We have selected the ground truth according to the guidelines followed in the previous data set on the Pavia area.

As in the case of the Pavia data set, to evaluate the effectiveness of the proposed approach, we conducted two different sets of experiments. The first one was aimed at assessing the effects of both the number of context levels and the different kinds of extracted features on the classification accuracies. The second one compared the performances of the proposed system with those of the feature-extraction method based on the generalized Gaussian pyramid decomposition of the image.

1) Experiment 1—Analysis of the Effectiveness of the Proposed Approach: In our experiments, we carried out several trials with hierarchies made up of two to seven context levels. The first level is the pixel level, whereas the other six levels are obtained according to the presented multiscale hierarchical segmentation technique with different parameters to tune the homogeneity predicate. As in experiments on the urban data set, levels between two and three are characterized by very small regions. This means that from a general point of view, objects in these levels are highly oversegmented. Levels 4 and

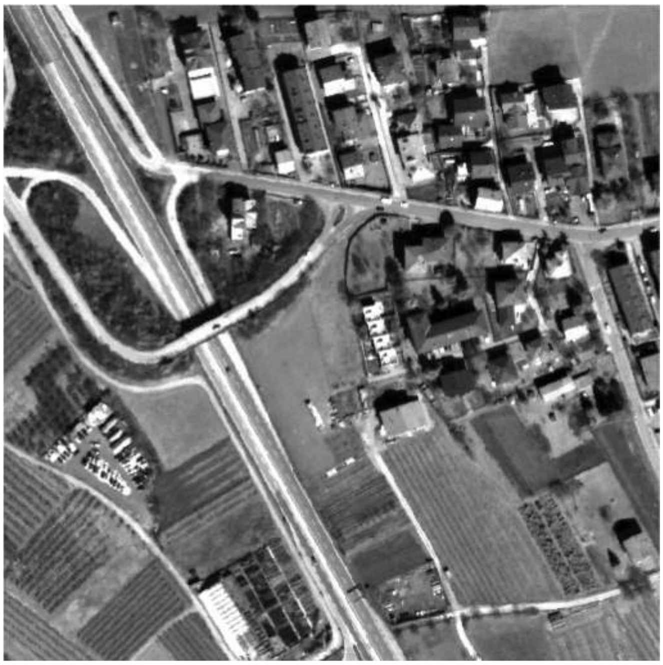


Fig. 9. Panchromatic image (512 × 512 pixels) acquired by the Quickbird satellite on the city of Trento (northern Italy).

TABLE IV
NUMBER OF SAMPLES (IN PIXELS) IN THE TRAINING AND TEST SETS (TRENTO DATA SET)

Class	Number of patterns		
	Training set	*Test set on edge areas*	*Test set on homogeneous areas*
Gray roof	1184	158	443
Red roof	309	141	219
Road	2351	376	2059
Shadow	857	289	473
Rural area 1	740	208	551
Rural area 2	3199	123	2964
Grass	703	139	579
Trees area	2047	165	1687
TOTAL	11390	1653	8975

5 are characterized by regions of medium size. Levels 6 and 7 contain regions that represent (or include) the objects present in the scene. The features extracted for the first level (i.e., the pixel level) were only the pixel values in all the spectral channels and the panchromatic image. For level 2, we only considered the mean value of the digital numbers of pixels defining each region in each spectral band and the panchromatic image. From levels 3 to 7, for each region and for each band, we considered the mean value and the standard deviation of the digital numbers. On the whole, 10, 20, 30, 40, 50, and 60 features were considered for experiments with two, three, four, five, six, and seven levels, respectively. In all the experiments, we used an SVM classifier with RBF kernels. According to a proper model selection technique [20], we identified the best values of the regularization parameter C and the spread factor of Gaussian kernels γ using the training set samples and the global test set samples for validation. The highest accuracies obtained, as well as the related parameter values, are shown in Table V.

These results confirm the effectiveness of the proposed classification system, which always exhibited a greater overall accuracy compared with that obtained using only the pixel level (much greater starting from three levels of context representation). In detail, the greatest increase in overall accuracy (i.e., about 10%) was obtained on edge test areas with six levels. With this number of levels, classification accuracy increased also on homogeneous test areas with an improvement of about 3%. It is worth nothing that these results confirm the effectiveness of the empirical criterion for the selection of the number of levels described in (11), which on this data set identified an optimal number of decomposition levels equal to six.[5] By analyzing Table V, one can see that the proposed approach provided stable accuracies versus the number of levels considered in the neighborhood of the optimal number of scales identified with the proposed empirical criterion (in the range between five and seven levels). In particular, it exhibited Kappa values between 0.62 and 0.64 on edge areas and between 0.97 and 0.98 on homogeneous areas. This confirms the ability of the presented methodology to model the spatial context of each analyzed pixel.

TABLE V
(a) KAPPA ACCURACY AND (b) OVERALL ACCURACY PROVIDED BY THE PROPOSED APPROACH ON EDGE, HOMOGENEOUS, AND GLOBAL TEST AREAS VERSUS THE NUMBER OF CONSIDERED CONTEXT LEVELS. THE OPTIMAL VALUES OF THE REGULARIZATION PARAMETER C AND OF THE SPREAD γ OF THE KERNEL FUNCTIONS ARE REPORTED (TRENTO DATA SET)

Number of levels (including pixel level)	SVM parameters		Kappa accuracy on Test Sets		
	Regularization parameter (C)	*Spread of kernel function (γ)*	*Edge areas*	*Homogeneous areas*	*Homogeneous and edge areas*
1 - only pixel level	25	32	0.525	0.942	0.860
2	10	28	0.546	0.944	0.866
3	10	1.0	0.583	0.966	0.891
4	10	1.5	0.574	0.979	0.899
5	25	2.5	0.623	0.979	0.909
6	15	1.0	0.637	0.980	0.913
7	15	2.0	0.629	0.971	0.904

(a)

Number of levels (including pixel level)	SVM parameters		Overall accuracy (%) on Test Sets		
	Regularization parameter (C)	*Spread of kernel function (γ)*	*Edge areas*	*Homogeneous areas*	*Homogeneous and edge areas*
1- only pixel level	25	32	59.30	95.43	88.62
2	10	28	61.31	95.54	89.11
3	10	1.0	64.53	97.30	91.15
4	10	1.5	63.81	98.33	91.84
5	25	2.5	67.88	98.32	92.60
6	15	1.0	69.18	98.40	92.91
7	15	2.0	68.41	97.75	92.21

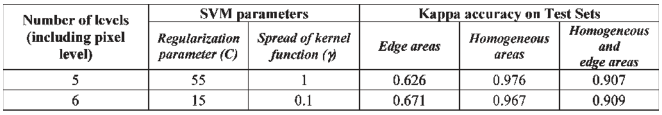
(b)

According to the previous results, in the second part of this experiment, we considered only five and six levels, as they gave the highest classification accuracies on the overall test set. To assess the importance of the use of geometrical features on this data set, we computed some geometrical parameters from the regions extracted at levels 5 and 6. We considered the following features.

1) *Width-to-length ratio*. It can be calculated as the ratio between the length and the width of the bounding box that contains the object under investigation.
2) *Shape index*. It can be obtained by the ratio of the border length of the object under analysis and four times the square root of its area.

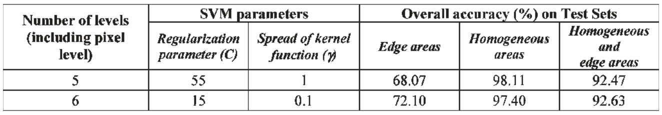
[5] In this case, we used a hybrid approach for defining the $\mathrm{EA}_{\mathrm{th}}$ parameter, by considering an average of the size of the objects that compose the rural scene (i.e., buildings and crops).

TABLE VI
(a) KAPPA ACCURACY AND (b) OVERALL ACCURACY PROVIDED BY THE PROPOSED APPROACH ON EDGE, HOMOGENEOUS, AND GLOBAL TEST AREAS VERSUS THE NUMBER OF CONSIDERED CONTEXT LEVELS WHEN GEOMETRICAL FEATURES WERE ALSO CONSIDERED. THE OPTIMAL VALUES OF THE REGULARIZATION PARAMETER C AND OF THE SPREAD γ OF THE KERNEL FUNCTIONS ARE REPORTED (TRENTO DATA SET)



Number of levels (including pixel level)	SVM parameters		Kappa accuracy on Test Sets		
	Regularization parameter (C)	*Spread of kernel function (γ)*	*Edge areas*	*Homogeneous areas*	*Homogeneous and edge areas*
5	55	1	0.626	0.976	0.907
6	15	0.1	0.671	0.967	0.909

(a)



Number of levels (including pixel level)	SVM parameters		Overall accuracy (%) on Test Sets		
	Regularization parameter (C)	*Spread of kernel function (γ)*	*Edge areas*	*Homogeneous areas*	*Homogeneous and edge areas*
5	55	1	68.07	98.11	92.47
6	15	0.1	72.10	97.40	92.63

(b)

3) *Rectangular fit*. It can be obtained as the ratio between the area not covered by a rectangle with the same area and proportion of the object under investigation, and the area of the object.

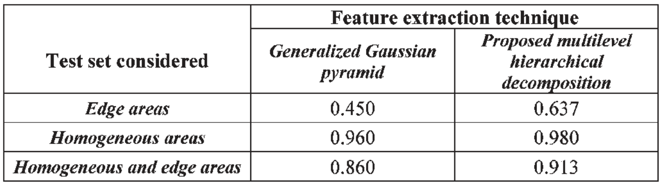
On the whole, 46 and 56 features were considered for experiments with five and six levels, respectively. Table VI reports the best accuracies obtained with these features after performing a new model selection for the SVM classifier.

These results point out that, on this data set, the use of geometrical features increases the accuracy on edge areas with respect to that obtained by using only spectral features, at the expense of a slight decrease of accuracy in homogeneous areas. In greater detail, in the six-level case, we obtained an increase of about 3% in terms of Kappa accuracy over edge areas and a decrease of 1% over homogeneous areas. This interesting result, which does not seem intuitive, can be explained as follows. On the one hand, the use of geometrical features allows a better characterization of pixels close to the border areas, which are better "attracted" from the geometry of objects to which they belong. On the other hand, oversegmentation errors slightly affect the accuracy on homogeneous areas, where spectral and textural features are sufficient for obtaining high accuracies.

2) Experiment 2—Comparisons With a Feature-Extraction Module Based on a Generalized Gaussian Pyramid Decomposition: Also, on this study area, the aim of the second set of experiments is to compare the proposed system with a multilevel feature extraction based on the generalized Gaussian pyramid decomposition. As in the previous case, we used five levels of pyramidal decomposition to characterize the spatial context of pixels (five levels resulted in the highest accuracies on the global test set) and adopted an SVM-based classification module. The extracted vector was made up of 25 features.

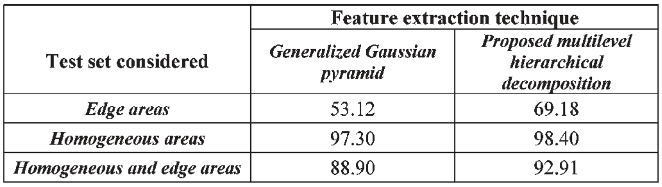
The best results (in terms of Kappa and overall accuracies) obtained with the proposed technique (with six levels and the same feature vector extracted in the first part of the previous experiment) and with the generalized Gaussian pyramid technique are reported in Table VII.

TABLE VII
(a) KAPPA ACCURACY AND (b) OVERALL ACCURACY ON EDGE, HOMOGENEOUS, AND GLOBAL TEST AREAS VERSUS THE PROPOSED FEATURE-EXTRACTION TECHNIQUE (ADAPTIVE HIERARCHICAL CONTEXT MODELING) AND THE REFERENCE TECHNIQUE (GENERALIZED GAUSSIAN PYRAMID REDUCTION) (TRENTO DATA SET)



Test set considered	Feature extraction technique	
	Generalized Gaussian pyramid	*Proposed multilevel hierarchical decomposition*
Edge areas	0.450	0.637
Homogeneous areas	0.960	0.980
Homogeneous and edge areas	0.860	0.913

(a)



Test set considered	Feature extraction technique	
	Generalized Gaussian pyramid	*Proposed multilevel hierarchical decomposition*
Edge areas	53.12	69.18
Homogeneous areas	97.30	98.40
Homogeneous and edge areas	88.90	92.91

(b)

These results confirm the better capability of the proposed approach to model the geometrical details of objects in the scene. In greater detail, it increased the Kappa value on the edge areas by about 19% compared with the generalized Gaussian pyramid method. In addition, a slight increase of accuracy on homogeneous areas was obtained (i.e., about 2%). To better assess the effectiveness of the investigated methods, Fig. 10(a) and (b) shows the classification maps obtained using the proposed classification system and the system based on the Gaussian pyramid feature extraction.

A qualitative analysis of maps in Fig. 10 confirms the previous consideration based on the quantitative results. The adaptive and multilevel properties of the proposed feature-extraction technique can better model the edge of objects in the scene. In greater detail, the map in Fig. 10(b) shows that the generalized Gaussian pyramid decomposition is not suitable to accurately model the boundaries of objects and complex structures due to a blurring effect. On the contrary, when the proposed multilevel segmentation algorithm is used to adaptively model the neighborhood of a pixel, a proper representation of the edges of the objects is obtained.

V. DISCUSSION AND CONCLUSION

In this paper, a novel system for the classification of very high resolution images has been presented. The system is made up of: 1) a feature-extraction module that adaptively models the spatial context of each pixel according to a complete hierarchical multilevel representation of the scene under investigation and 2) a proper classifier based on SVMs. In greater detail, a hierarchical segmentation is applied to the images to obtain segmentation results at different levels of resolution according to tree-based hierarchical constraints. In this way, precise hierarchical relationships are established between each pixel in the image and the regions that adaptively define its context at different levels. Each pixel is characterized by a feature vector that includes both the pixel-level information in the spectral channels of the sensor and the attributes of all the regions, which represent the multilevel relationships of the pixel and define its spatial context adaptively.

Depending on the level considered, different kinds of features are extracted to characterize the regions with the most reliable attributes for the specific scale analyzed and the specific scene considered. It is worth noting that in our technique, unlike other approaches proposed in the literature, all features associated both with the pixel level and all the region levels are jointly considered in the classification phase to label a pixel. This hierarchical representation allows to capture and exploit the entire information in the scene by working with adaptive regions at different scales. To deal with the large number of feature-vector components to be given to the classifier as input, we used a machine-learning classifier based on SVMs. This choice depends both on the effectiveness of SVMs classifiers and on their capabilities to analyze a high-dimensional feature space with a reduced effect of the Hughes phenomenon.

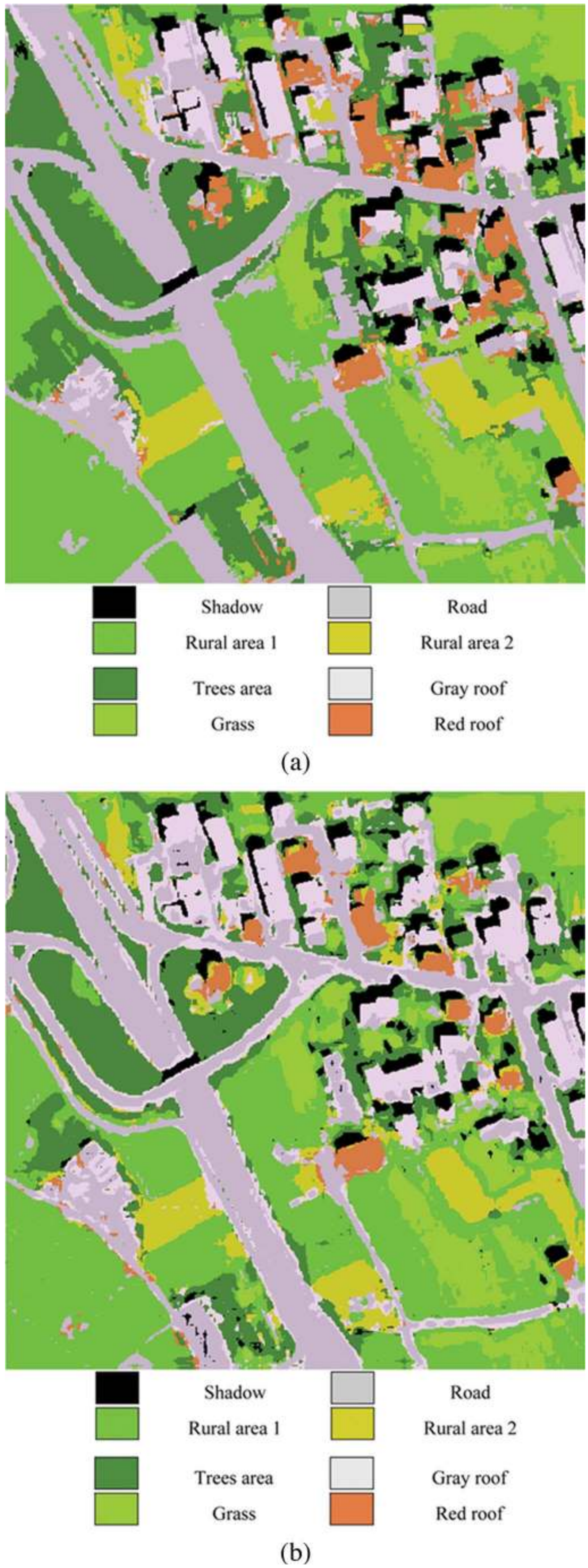


Fig. 10. Classification maps obtained by an SVM classifier (a) with the proposed feature extraction module and (b) with a feature extraction algorithm based on a pyramidal Gaussian decomposition (Trento data set).

Experimental results, obtained on two very high geometrical resolution Quickbird images acquired on a complex urban area and on a rural area, confirm the effectiveness of the proposed classification system. In detail, two main experiments have been carried out. In the first, we focused on the number of levels to be used to model the context of a pixel. The results show that varying the number of levels used to characterize the spatial context adaptively, in a range close to the “optimal” level identified by the empirical criterion proposed in (11) and (12), does not critically change the overall accuracy and the quality of classification maps. In the second set of experiments, we compared the proposed feature-extraction technique with a standard feature-extraction algorithm based on a generalized Gaussian decomposition pyramid. The SVM classifier was also used in these experiments. The experimental results confirm that the proposed feature-extraction module outperforms the reference method based on the Gaussian pyramidal reduction. This is due both to the adaptive and to the multilevel nature of the proposed feature-extraction module, which by exploiting a hierarchical segmentation algorithm can model the objects (shapes and relationships at different levels of resolution) in the scene under investigation better, compared with the feature-extraction module based on the generalized Gaussian pyramid decomposition.

ACKNOWLEDGMENT

The authors wish to thank A. Garzelli (University of Siena, Italy) for providing the Quickbird image of Pavia, R. Rigon (University of Trento, Italy) for providing the Quickbird image of Trento (which was acquired in the framework of the project ASI 175/02 funded by the Italian Space Agency), and A. Baraldi (Joint Research Centre, Ispra, Italy) for useful discussions about the segmentation algorithm. The authors are also grateful to the anonymous referees for their constructive comments.

REFERENCES

[1] F. Volpe and L. Rossi, “Quickbird high resolution satellite data for urban application,” in *Proc. 2nd GRSS/ISPRS Joint Workshop Data Fusion and Remote Sens. Over Urban Areas*, May 2003, pp. 1–3.

[2] S. R. Repaka, D. D. Truax, E. Kolstad, and C. G. O’Hara, “Comparing spectral and object based approaches for classification and transportation feature extraction from high resolution multispectral imagery,” in *Proc. ASPRS Annu. Conf.*, May 2004.

[3] S. P. Lennartz and R. G. Congalton, "Classifying and mapping forest cover types using Ikonos imagery in the northeastern United States," in *Proc. ASPRS Annu. Conf.*, May 2004.

[4] L. M. Moskal, "Historical landscape visualization of the Wilson's creek nationalbattlefield based on object oriented tree detection method from Ikonos imagery," in *Proc. ASPRS Annu. Conf.*, May 2004.

[5] S. J. Goetz, R. K. Wright, A. J. Smith, E. Zineckerb, and E. Schaub, "IKONOS imagery for resource management: Tree cover, impervious surfaces, and riparian buffer analyses in the mid-Atlantic region," *Remote Sens. Environ.*, vol. 88, no. 1/2, pp. 195–208, Nov. 2003.

[6] A. Carleer, O. Debeir, and E. Wolff, "Comparison of very high spatial resolution satellite image segmentation," in *Proc. SPIE Conf. Image and Signal Processing Remote Sensing IX*, vol. 5238, pp. 532–542.

[7] E. Binaghi, I. Gallo, and M. Pepe, "A neural adaptive model for feature extraction and recognition in high resolution remote sensing imagery," *Int. J. Remote Sens.*, vol. 24, no. 20, pp. 3947–3959, Oct. 2003.

[8] T. Blaschke, "Object-based contextual image classification built on image segmentation," in *Proc. IEEE Workshop Adv. Tech. Anal. Remote Sensed Data*, Oct. 2003, pp. 113–119.

[9] C. Unsalan and K. L. Boyer, "Classifying land development in high-resolution panchromatic satellite images using straight-line statistics," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 4, pp. 907–919, Apr. 2004.

[10] A. K. Shackelford and C. H. Davis, "A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 4, no. 9, pp. 1920–1932, Sep. 2003.

[11] M. De Martinao, F. Causa, and S. B. Serpico, "Classification of optical high resolution images in urban environment using spectral and textural information," in *Proc. IGARSS*, Jul. 2003, vol. 1, pp. 467–469.

[12] R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," *Comput. Vis. Graph. Image Process.*, vol. 29, no. 1, pp. 100–132, Jan. 1985.

[13] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, no. 3/4, pp. 239–258, Jan. 2004.

[14] C. Burnett and T. Blaschke, "A multi-scale segmentation/object relationship modeling methodology for landscape analysis," *Int. J. Ecol. Model. Syst. Ecol.*, vol. 168, no. 3, pp. 233–249, Oct. 2003.

[15] E. Binaghi, I. Gallo, and M. Pepe, "A cognitive pyramid for contextual classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2906–2922, Dec. 2003.

[16] A. K. Shackelford and C. H. Davis, "A combined fuzzy pixel-based and object-based approach for classification of high-resolution multispectral data over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 10, pp. 2354–2363, Oct. 2003.

[17] J. A. Benediktsson, M. Pesaresi, and K. Arnason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9, pp. 1940–1949, Sep. 2003.

[18] C. Mott, T. Andresen, S. Zimmermann, T. Schneider, and U. Ammer, "Selective region growing-an approach based on object-oriented classification routine," in *Proc. IGARSS*, Jun. 2002, vol. 3, pp. 1612–1614.

[19] R. A. Schowengerdt, *Remote Sensing. Models and Methods for Image Processing*, 2nd ed. Norwell, MA: Academic, 2002.

[20] N. Cristianini and J. Shaew-Taylor, *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[21] L. Bruzzone and F. Melgani, "Support vector machines for classification of hyperspectral remote-sensing images," in *Proc. IGARSS*, Jun. 2002, vol. 1, pp. 506–508.

[22] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote-sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[23] T. Joachims, *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999.

[24] A. Baraldi, L. Bruzzone, and P. Blonda, "Badly-posed classification of remotely sensed images—An experimental comparison of existing data mapping system," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 1, pp. 214–235, Jan. 2006.

[25] *ENVI User Manual*. Boulder, CO: RSI, 2003. [Online.] Available: http://www.RSInc.com/envi

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

[27] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[28] Definiens Imaging, *eCognition Professional User Guide 4*, 2003, Munich, Germany. [Online]. Available: http://www.definiens-imaging.com

Lorenzo Bruzzone (S'95–M'98–SM'03) received the laurea (M.S.) degree in electronic engineering (summa cum laude) and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

From 1998 to 2000, he was a Postdoctoral Researcher at the University of Genoa. From 2000 to 2001, he was an Assistant Professor at the University of Trento, Trento, Italy, and from 2001 to 2005, he was an Associate Professor at the same university. Since March 2005, he has been a Full Professor of telecommunications at the University of Trento, where he currently teaches remote sensing, pattern recognition, and electrical communications. He is currently the Head of the Remote Sensing Laboratory in the Department of Information and Communication Technology, University of Trento. His current research interests are in the area of remote-sensing image processing and recognition (analysis of multitemporal data, feature selection, classification, regression, data fusion, and machine learning). He conducts and supervises research on these topics within the frameworks of several national and international projects. Since 1999, he has been appointed Evaluator of project proposals for the European Commission. He is the author (or coauthor) of more than 150 scientific publications, including journals, book chapters, and conference proceedings. He is a Referee for many international journals and has served on the Scientific Committees of several international conferences.

Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (Seattle, July 1998). He was a recipient of the Recognition of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Best Reviewers in 1999 and was a Guest Editor of a Special Issue of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING on the subject of the analysis of multitemporal remote-sensing images (November 2003). He was the General Chair and Cochair of the First and Second IEEE International Workshop on the Analysis of Multi-temporal Remote-Sensing Images. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He is a member of the Scientific Committee of the India–Italy Center for Advanced Research. He is also a member of the International Association for Pattern Recognition and of the Italian Association for Remote Sensing (AIT).

Lorenzo Carlin (S'06) received the laurea (B.S.) and Laurea Specialistica (M.S.) degrees in telecommunication engineering (summa cum laude) from the University of Trento, Trento, Italy, in 2001 and 2003, respectively. He is currently working toward the Ph.D. degree in information and communication technologies at the same university.

He is currently with the Pattern Recognition and Remote Sensing group at the Department of Telecommunication and Information Technologies, University of Trento. His main research activity is in the area of pattern recognition applied to remote sensing images; in particular, his interests are related to classification of very high resolution remote sensing images. He conducts research on these topics within the frameworks of several national and international projects. He is a referee for the *Italian Journal of Remote Sensing* (AIT).

Mr. Carlin is a referee for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.