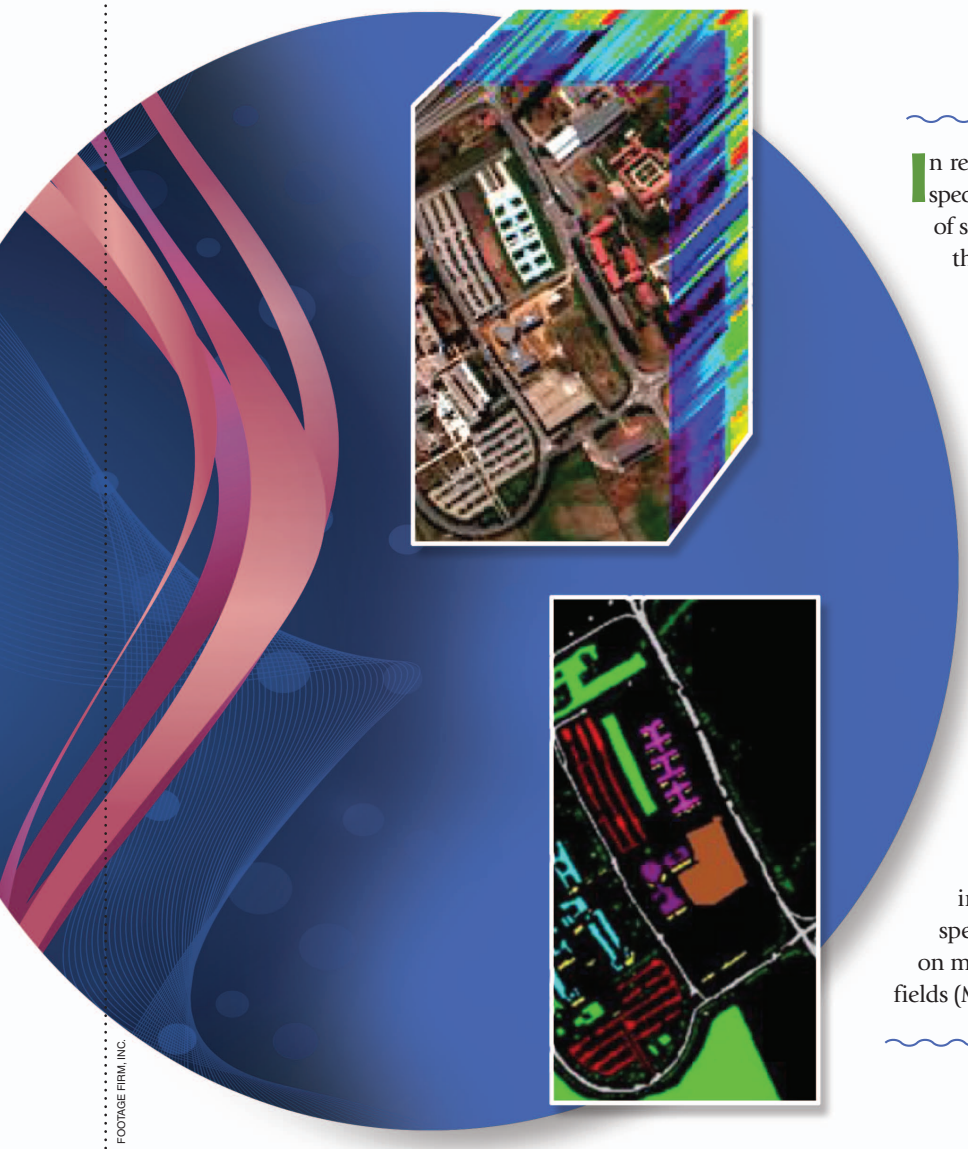
New Frontiers in Spectral-Spatial Hyperspectral Image Classification

The latest advances based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning



FOOTAGE FIRM, INC.

In recent years, airborne and spaceborne hyperspectral imaging systems have advanced in terms of spectral and spatial resolution, which makes the data sets they produce a valuable source for land cover classification. The availability of hyperspectral data with fine spatial resolution has revolutionized hyperspectral image (HSI) classification techniques by taking advantage of both spectral and spatial information in a single classification framework.

The ECHO classifier, proposed in 1976, might be the first spectral-spatial classification approach of its kind to be used in the remote-sensing community. Since then and especially in recent years, increased attention has been dedicated to developing sophisticated spectral-spatial classification methods. There is a rich literature on this topic, which includes several fast-growing branches. This article critically reviews the latest advances in spectral-spatial classification of hyperspectral data. More than 25 approaches based on mathematical morphology, Markov random fields (MRFs), segmentation, sparse representation

PEDRAM GHAMISI, EMMANUEL MAGGIORI, SHUTAO LI, ROBERTO SOUZA, YULIYA TARABALKA, GABRIELE MOSER, ANDREA DE GIORGI, LEYUAN FANG, YUSHI CHEN, MINGMIN CHI, SEBASTIANO B. SERPICO, AND JÓN ATLI BENEDIKTSSON

Digital Object Identifier 10.1109/MGRS.2018.2854840
Date of publication: 24 September 2018

(SR), and deep learning are addressed, with an emphasis on their methodological foundations. We also present examples of experimental results using three benchmark hyperspectral data sets, including both well-known, long-used data and a recent data set resulting from an international contest. In addition, training and test sets for these data sets along with several codes and libraries are shared with the online community.

HYPERSPECTRAL IMAGING CLASSIFICATION

Hyperspectral imaging sensors capture data that are usually visible through the near-infrared wavelength ranges and consist of hundreds of (narrow) spectral channels with continuous spectral information that can accurately discriminate diverse materials of interest on the immediate surface of the Earth. Therefore, HSIs are considered a valuable source of information for object identification and classification [1].

An HSI is a stack of n pixel vectors, where n indicates the number of pixels in the image. The length of each pixel vector is equal to the number of bands or spectral channels. Supervised classification plays a vitally important role for analyzing HSIs and is used to differentiate between the diverse land covers of interest available in a scene [1]. A classification technique assigns unknown pixels to one of the available classes according to a set of representative samples for each class, known as *training samples*. Detailed information about advanced supervised classifiers for HSIs can be found in [2].

The first attempts dedicated to HSI classification were based on techniques developed for multispectral images that have only a small number of spectral channels, i.e., generally fewer than 13. However, most of the commonly used methods designed for the analysis of gray scale, color, or multispectral images are inappropriate and even useless for HSIs. In fact, despite of all the similarities among HSIs and other optical images (panchromatic, red-green-blue, and multispectral), analyzing an HSI is more challenging for a number of reasons, including the high dimensionality of HSI data, the existence of extreme redundancy within HSIs, the existence of different types of noise, and the uncertainty of the sources observed.

Hyperspectral imaging often deals with inherently nonlinear relations between the captured spectral information and the corresponding material. This nonlinear relation is due to a wide variety of reasons, such as 1) the undesired scattering from other objects during the acquisition process, 2) the different atmospheric and geometric distortions, and 3) the intraclass variability of similar objects. Conversely, training samples are usually collected by the manual labeling of a small number of pixels in an image or are based on some field measurements, both of which are expensive and/or time consuming. As a result, the number of available training samples is usually limited compared to the available number of bands in HSIs, which makes the supervised classification of HSIs extremely challenging.

Additionally, neighborhood pixels in HSIs are highly correlated because remote sensors acquire a considerable amount of energy from adjacent pixels. Moreover, homogeneous structures in an image scene are generally larger than the size of a pixel [1]. This is particularly evident for images of very high spatial resolution (VHR), and it has triggered research efforts dedicated to spectral-spatial classification because the integration of these two sources of information can substantially improve the discrimination power of classifiers in complex scenes. To this end, spatial and contextual data can provide useful information about the shape of different structures. Additionally, such information reduces the labeling uncertainty that exists when only spectral information is taken into account and also helps to address the "salt and pepper" appearance of the resulting classification map.

To extract spatial information from HSIs, most methodological approaches can be broadly related to two common strategies: the crisp neighborhood system [3]–[5] and the adaptive neighborhood system [1], [6], [7]. Methodologies based on the crisp neighborhood system extract spatial and contextual information using a neighborhood of predefined shapes. Conversely, methodologies based on the adaptive neighborhood system are conceptually more flexible and make use of neighborhoods of various shapes. In this context, two-dimensional (2-D) convolutional neural networks (CNNs) [5] and the MRF family [3], [4], [8] are mostly categorized as spectral-spatial classification approaches using the crisp neighborhood system. In contrast, methodologies based on segmentation [9], [11], morphological profiles (MPs) [11], [12], attribute profiles (APs) [6], [13], and extinction profiles (EPs) [14], [15] can extract spatial and contextual information using adaptive neighborhood systems.

Figure 1 demonstrates the importance of HSI classification in our community. The number of papers can be

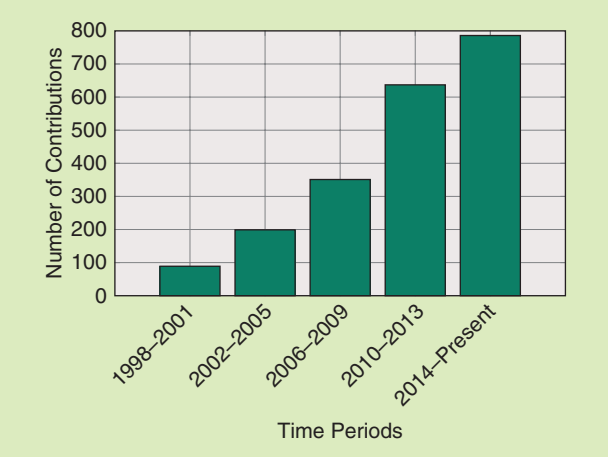


FIGURE 1. The number of journal and conference papers available in IEEE *Xplore* on the subject of hyperspectral imaging within different periods since 1998. This graph was prepared based on the number of contributions through 1 March 2017.

obtained by checking the keywords of "hyperspectral" and "classification" used in the abstract of published journals and conference papers that appear in IEEE *Xplore*. To highlight the growth in the number of published papers, relevant time periods have been equally divided (i.e., 1998–2001, 2002–2005, 2006–2009, 2010–2013, and 2014–2017). As can be seen, the number of papers dedicated to this subject has increased dramatically.

Due to the rapid growth and importance of HSI classification in the remote-sensing community, this article attempts to critically and systematically review the latest advances in spectral-spatial HSI classification. The focus is on the methodological foundations of the considered families of techniques and on their mutually complementary methodological rationales, which can provide the reader with a comprehensive picture of the current evolution of HSI spectral-spatial classifiers. To this end, computational properties are also examined, and examples of experimental results are discussed for all of the considered algorithms. Three benchmark data sets, which include both widely known and long-used data as well as a recent data set released during the 2013 Data Fusion Contest of the IEEE Geoscience and Remote Sensing Society (GRSS), are used for this purpose. In this context, we review more than 25 methods categorized into five branches: mathematical-morphology-based techniques, MRFs, segmentation approaches, SR methods, and deep-learning-based classifiers. For each category, the main methodological ideas are studied, and a few key techniques are explained and demonstrated using the data sets described previously. Finally, some possible future directions are highlighted. Several codes and libraries as well as the training and test sets used in this article are shared and made available publicly. It should be noted that this article places particular emphasis on methodologies developed since 2013 (after the publication of a previous survey paper on spectral-spatial classification [7]).

HYPERSPECTRAL IMAGING OFTEN DEALS WITH INHERENTLY NONLINEAR RELATIONS BETWEEN THE CAPTURED SPECTRAL INFORMATION AND THE CORRESPONDING MATERIAL.

HSI classification is crucial for a wide variety of real-world applications:

- ecological science (e.g., estimating biomass and carbon, studying biodiversity in dense forest zones, and monitoring land cover changes)
- geological science (e.g., recovering physicochemical mineral properties, such as composition and abundance)
- mineralogy (e.g., identifying a wide range of minerals)
- hydrological science (e.g., determining changes in wetland characteristics, surveying water quality, and monitoring estuarine environments and coastal zones)
- precision agriculture (e.g., categorizing agricultural classes and extracting nitrogen content for the purpose of precision agriculture)
- military applications (e.g., target detection and classification).

This article, however, emphasizes the methodological aspects of recent publications on spectral-spatial classification.

NOTATIONS

In this article, matrices are denoted by bold and capital letters. The comma (,) and semicolon (;) are used for horizontal and vertical concatenation, respectively, of the elements in a matrix. $\hat{\mathbf{X}}$ stands for the estimate of the variable $\mathbf{X}$, and $\mathbf{X}^m$ denotes the estimate of the variable $\mathbf{X}$ at the mth iteration of some iterative method. $|\,.\,|$ is the absolute value, $\|\,.\,\|_F$ is the Frobenius norm, and $\|\,.\,\|_n$ is the ℓ_n norm. The Kronecker product is denoted by $\otimes$. The identity matrix of size $p \times p$ is denoted by $\mathbf{I}_p$.

A hyperspectral data cube, which consists of d spectral channels and $n(= n_1 \times n_2)$ pixels in each spectral channel, is denoted with an $n \times d$ matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_i$ refers to the spectral vector of the ith pixel. A classification approach tries to assign unknown pixels to one of the classes in $\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$, where C represents the number of classes, using a set of training samples for these classes. Vector $\mathbf{Y} = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ collects the classification labels of all of the pixels.

DATA SETS

Three benchmark data sets have been used to illustrate the considered spectral-spatial methods through examples of experimental results. Two are very well known and have been used by the hyperspectral community for many years. The third is quite recent and was made available to the remote-sensing community during the 2013 IEEE GRSS Data Fusion Contest [16].

The first data set was acquired by the airborne visible/infrared imaging spectrometer (AVIRIS) sensor over the agricultural Indian Pines test site in northwestern Indiana, United States. The spatial dimensions of this data set are 145×145 pixels, and the spatial resolution is 20 m. This data set originally included 220 spectral channels; however, 20 water absorption bands (104–108, 150–163, and 220) have been removed, and the rest (200 bands) have been taken into account for the experiments. The reference data contain 16 classes of interest, which represent mostly different types of crops and are detailed in Table 1. Figure 2 shows a three-band false color image and its corresponding reference samples.

The second data set was captured in the city of Pavia, Italy, by the reflective optics spectrographic imaging system (ROSIS)-03 airborne instrument (Table 2). The flight over the city of Pavia was operated by the German Aerospace Agency (Deutschen Zentrum für Luft- und Raumfahrt) within the context of the HySens project, which is managed and sponsored by the European Union. The ROSIS-03 sensor has 115 data channels with a spectral

TABLE 1. THE AVIRIS INDIAN PINES DATA SET: THE NUMBER OF TRAINING AND TEST SAMPLES.

CLASS		NUMBER OF SAMPLES	
NUMBER	NAME	TRAINING	TEST
1	Corn-Notill	50	1,384
2	Corn-Mintill	50	784
3	Corn	50	184
4	Grass-Pasture	50	447
5	Grass-Trees	50	697
6	Hay-Windrowed	50	439
7	Soybean-Notill	50	918
8	Soybean-Mintill	50	2,418
9	Soybean-Clean	50	564
10	Wheat	50	162
11	Woods	50	1,244
12	Building-Grass-Tree-Drives	50	330
13	Stone-Steel-Towers	50	45
14	Alfalfa	50	39
15	Grass-Pasture-Mowed	50	11
16	Oats	50	5
Total		695	9,671

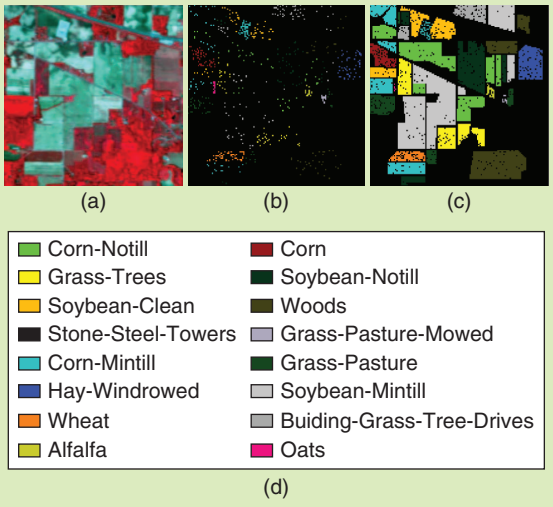


FIGURE 2. The AVIRIS Indian Pines hyperspectral data set. (a) A three-band false color composite, (b) training samples, (c) test samples, and (d) color codes for the classes.

coverage ranging from 0.43 to 0.86 μm. Twelve channels have been removed due to noise, the remaining 103 spectral channels have been processed, and the spatial resolution is 1.3 m. The data set covers the Engineering School of the University of Pavia and consists of different classes, including trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow, and soil. This data set comprises 640×340 pixels. Figure 3 presents a false color image of the ROSIS-03 Pavia University data and their corresponding reference samples.

The third data set, named *grss_dfc_2013* [17], was captured in June 2012 by the Compact Airborne Spectrographic Imager (CASI) over the campus of the University of Houston, Texas, United States, and the neighboring urban area. The size of the data is $349 \times 1{,}905$ with a spatial resolution of 2.5 m. This data set is composed of 144 spectral bands ranging from 0.38–1.05 μm. It consists of 15 classes, including grass healthy, grass stressed, grass synthetic, tree, soil, water, residential, commercial, road, highway, railway, parking lot 1, parking lot 2, tennis court, and running track. Parking lot 1 includes parking garages at the ground level and also in elevated areas, while parking lot 2 corresponds to parked vehicles. Table 3 demonstrates different classes with the corresponding number of training and test samples. Figure 4 shows a three-band false color image and its corresponding training and test samples.

In this article, we have incorporated a split of the ground truth of each considered data set into the training and the test sets, which is rather common in the hyperspectral community to make the results fully comparable with several studies in the literature. The sets of training and test samples

TABLE 2. THE ROSIS-03 PAVIA UNIVERSITY DATA SET: THE NUMBER OF TRAINING AND TEST SAMPLES.

CLASS		NUMBER OF SAMPLES	
NUMBER	NAME	TRAINING	TEST
1	Asphalt	548	6,304
2	Meadow	540	18,146
3	Gravel	392	1,815
4	Tree	524	2,912
5	Metal Sheet	256	1,113
6	Bare Soil	532	4,572
7	Bitumen	375	981
8	Brick	514	3,364
9	Shadow	231	795
Total		3,921	40,002

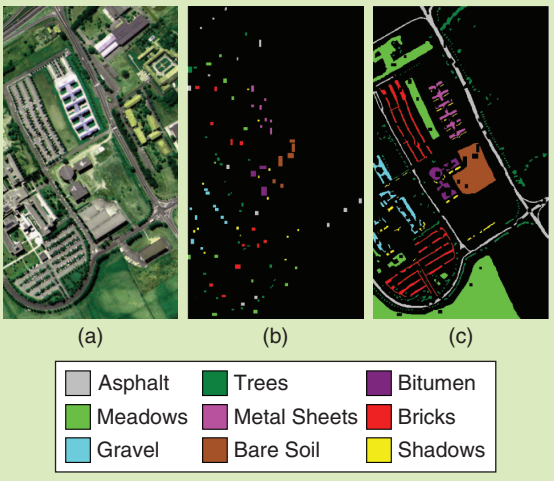


FIGURE 3. The ROSIS-03 Pavia University hyperspectral data set. (a) A three-band false color composite, (b) the training samples, and (c) the test samples.

TABLE 3. THE CASI UNIVERSITY OF HOUSTON DATA SET: THE NUMBER OF TRAINING AND TEST SAMPLES.

CLASS		NUMBER OF SAMPLES	
NUMBER	NAME	TRAINING	TEST
1	Grass–Healthy	198	1,053
2	Grass–Stressed	190	1,064
3	Grass–Synthetic	192	505
4	Tree	188	1,056
5	Soil	186	1,056
6	Water	182	143
7	Residential	196	1,072
8	Commercial	191	1,053
9	Road	193	1,059
10	Highway	191	1,036
11	Railway	181	1,054
12	Parking Lot 1	192	1,041
13	Parking Lot 2	184	285
14	Tennis Court	181	247
15	Running Track	187	473
Total		2,832	12,197

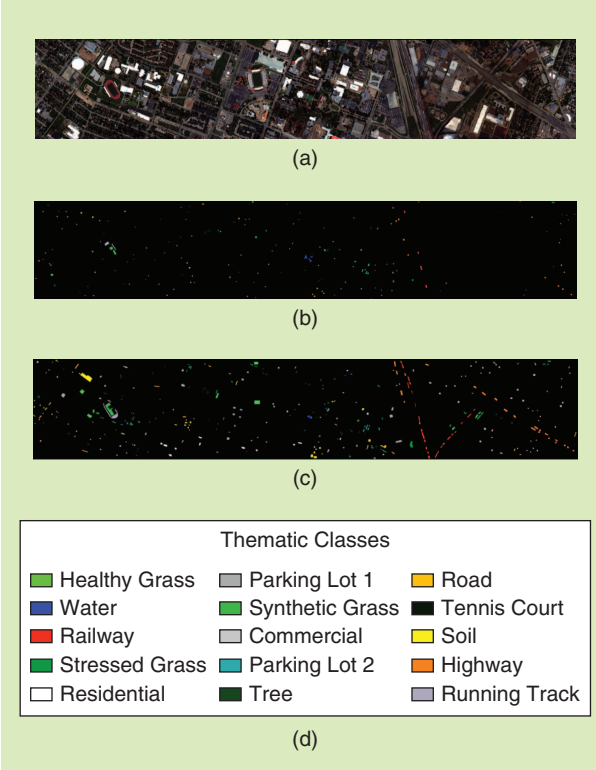


FIGURE 4. The CASI University of Houston data set. (a) A color-composite representation of the hyperspectral data using bands 70, 50, and 20, as red, green, and blue, respectively; (b) training samples; (c) test samples; and (d) color codes for the classes.

used in this article can be found at https://pghamisi.wixsite.com/mysite.

MATHEMATICAL-MORPHOLOGY-BASED SPECTRAL-SPATIAL CLASSIFIERS

BACKGROUND

The concept of MPs was introduced in 2001 [11]; since then, it has been used as a powerful approach to model an image's spatial information (e.g., contextual relations) by extracting structural features (e.g., size, geometry, and so on). MPs, which are constructed from the successive use of opening/closing operations with a structuring element (SE) of increasing size, led to the creation of a "morphological spectrum" for each pixel. In [18], the concept of MPs was successfully generalized to deal with HSIs [these are known as *extended MPs* (*EMPs*)]. A detailed survey of MPs and their extensions can be found in [1] and [7]. Although the MP can improve the discrimination ability of a spectral-spatial classification framework, the concept has a few limitations: 1) the shape of SEs is fixed, which makes it impossible for MPs to precisely model the shape of different objects, and 2) SEs are able to extract information with respect only to the size of existing objects and cannot characterize information concerning the gray-level characteristics of the regions.

To address these shortcomings, the morphological AP was introduced in [13] as a generalization of the MP; this provides a multilevel characterization of an image using the sequential morphological attribute filters (AFs). Compared to MPs, APs offer a more flexible tool because they can extract spatial and contextual features based on multiple attributes, which can be purely geometric, related to the spectral values of the pixels, or based on different characteristics such as spatial relations to other connected components. In [19], the concept of the AP was generalized and applied to HSIs, known as *extended APs* (*EAPs*) or *extended multi-APs* (*EMAPs*), if multiple types of attributes are taken into account. A detailed survey about APs and their extensions can be found in [1] and [6].

This section takes a closer look at EPs, a very recent variant of MPs [14]. We first briefly discuss the so-called tree-representation (max-tree), which is a crucial step for efficient implementation of EPs. Then, we present a brief discussion on AFs and extinction filters (EFs) highlighting the main differences between these two approaches. In addition, we discuss EPs and evaluate the performance of different mathematical-morphology-based spectral-spatial classifiers through experiments on three widely used hyperspectral data sets.

MAX-TREE

Max-tree is a data structure that represents a gray-scale image as a tree based on the hierarchical property of threshold decomposition. It was proposed by Salembier et al. [20] as an efficient structure to implement antiextensive (and extensive by duality) connected filters. Extensivity and

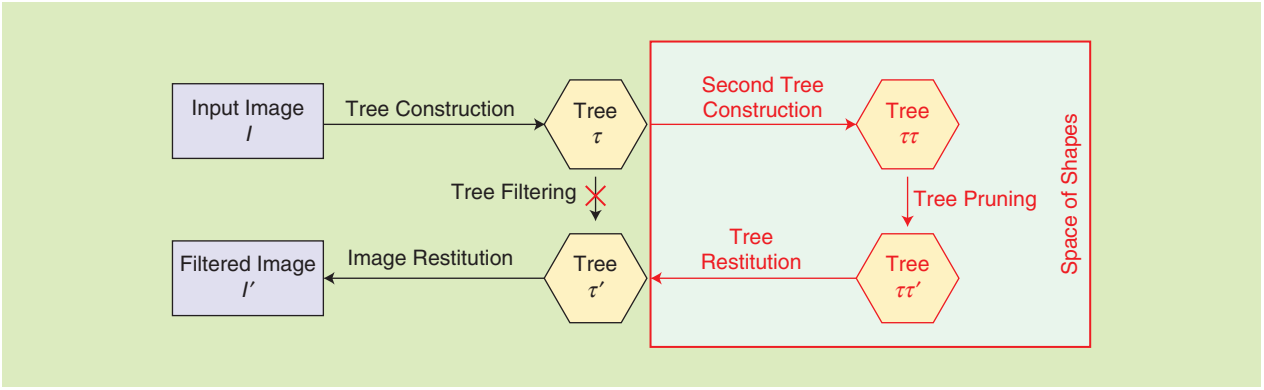


FIGURE 5. The max-tree and space of shapes fluxogram.

antiextensivity refer to transformation ψ, which is extensive if, for each pixel, the transformation output is greater than or equal to the original image and can be mathematically shown for a gray-scale image, $\mathbf{X}$, as $\mathbf{X} \leq \psi(\mathbf{X})$. By duality, the corresponding property is antiextensive if it satisfies $\mathbf{X} \geq \psi(\mathbf{X})$ for all pixels in the image.

There are algorithms that allow the max-tree construction in quasilinear time [21], [22]: the max-tree processing pipeline is depicted by the black path in Figure 5. The processing time required for max-tree filtering and image reconstruction is usually negligible compared to the construction time; thus, max-tree is even more efficient when used to perform a succession of filtering steps, such as the ones used to construct EPs [14]. In [23], the principles of max-tree representation, along with the corresponding algorithms and applications, were reviewed.

ATTRIBUTE FILTERS

A gray-scale image can be seen as a stack of binary images obtained at a different upper threshold ($\mathbf{X} \geq t$) ranging from the minimum to the maximum gray level of the image. Using this interpretation, the image's gray level is determined by the sum of the binary images in the stack. As an informal definition, AFs are connected filters that remove the connected components of each image in the stack that does not meet the threshold criteria. AFs may use either a single attribute or a set of attributes to decide which connected components to remove. There is a wide variety of AFs, such as area-open [24], hmax [20], vmax [25], ultimate opening [26], statistical AFs [27], and vector AFs [28]. These filters can be efficiently implemented on the max-tree structure [20].

The AF procedure on the max-tree is as follows:

1) Build the max-tree if implementing antiextensive filters; build the min-tree if implementing extensive filters of the image.
2) Mark all nodes that do not meet the threshold criteria based on the attribute being analyzed.
3) Filter the nodes marked in the previous steps.
4) Reconstruct the image from the filtered tree.

APs are constructed by the sequential application of attribute thinning and thickening with a set of progressively stricter threshold values proposed by Dalla Mura et al. [13]. A filter applied to the min-tree is a thickening operator, and a filter applied to the max-tree is a thinning operator. Since then, APs have been thoroughly investigated for the classification of HSIs. A detailed survey of the use of APs for the classification of HSIs can be found in [6].

EXTINCTION VALUES

Extinction values, which can be formally defined, are a measure of the persistence of extrema (minima or maxima) proposed by Vachier [25]. The measure of persistence is related to an attribute, which initially (as defined by Vachier) had to be increasing. Extinction values of the height attribute are known as *dynamics* [29]. Let M be a regional maximum of a gray-scale image $\mathbf{X}$, and let $\Psi = (\psi_\lambda)_\lambda$ be a family of decreasing and connected antiextensive transformations. The extinction value corresponding to M with respect to Ψ and denoted by $\varepsilon_\Psi(M)$ is the maximal λ value, such that M is still a regional maxima of $\psi_\lambda(\mathbf{X})$. This definition can be expressed as

$$\varepsilon_\Psi(M) = \sup\left\{\lambda \geq 0 \,\middle|\, \forall \mu \leq \lambda, M \subset \mathrm{Max}(\psi_\mu(\mathbf{X}))\right\}. \quad (1)$$

Extinction values of minima can be defined similarly. The height extinction values of maxima of a one-dimensional (1-D) signal are illustrated in Figure 6. It is important to emphasize that extinction values are not directly related

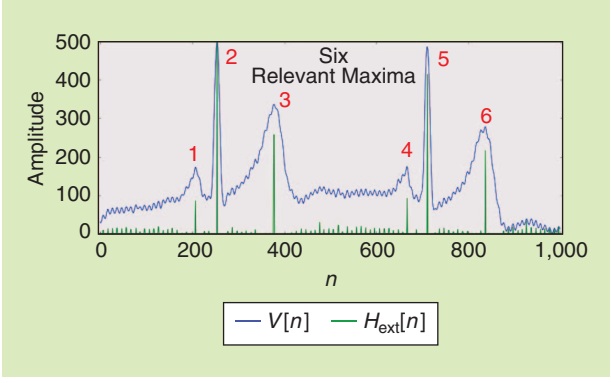


FIGURE 6. The height extinction values of the maxima of a 1-D signal. The six maxima with the highest extinction values are highlighted.

to the amplitude of the peak but depend on the adjacent extrema. In this illustration, the six most relevant maxima are not necessarily the six highest peaks in the signal. There are algorithms with linear complexity to compute extinction values [30], [31] from the max-tree [20], [32].

EXTINCTION FILTERS FOR INCREASING ATTRIBUTES

EFs for increasing attributes are connected idempotent filters, i.e., they do not blur the image and only alter the image the first time they are applied. They are extrema oriented and have three parameters to be set: the kind of extrema to be filtered (minima or maxima), the attribute being analyzed, and the number of extrema to be preserved.

EXTINCTION VALUES, WHICH CAN BE FORMALLY DEFINED, ARE A MEASURE OF THE PERSISTENCE OF EXTREMA (MINIMA OR MAXIMA) PROPOSED BY VACHIER.

Natural (real) images are contaminated by noise; therefore, they contain many irrelevant extrema, i.e., extrema with low extinction values. For example, a satellite high-resolution panchromatic image of an urban area like the city of Rome, Italy, acquired by the QuickBird satellite is depicted in Figure 7(a). The image is $972 \times 1{,}188$ pixels and has 67,960 regional maxima. More than 50% of the maxima has an area extinction value of 1 [Figure 7(b)]; therefore, if we apply an area-open [24] filter set to filter structures smaller or equal to 1, more than 50% of the image maxima would be filtered.

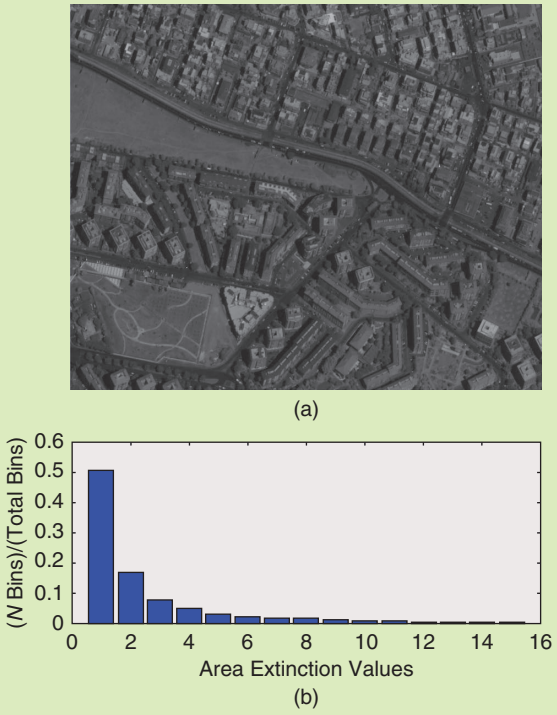


FIGURE 7. (a) A Rome satellite image and (b) its area-normalized extinction histogram.

EFs can be efficiently implemented using the max-tree structure [33]. The general description of the EF operation on the max-tree is as follows:

1) Build the image max-tree if filtering maxima (antiextensive) or the image min-tree if filtering minima (extensive).
2) Compute the leaves' extinction values of the increasing attribute being analyzed.
3) Mark all nodes on the paths starting from the n max-tree leaves with the highest extinction values assigned to the root.
4) Filter the nodes that were not marked in the previous step.
5) Reconstruct the image from the filtered tree.

The formal definition of EF for increasing attributes when filtering maxima is as follows: consider that $\mathrm{Max}(\mathbf{X}) = \{M_1, M_2, \ldots, M_N\}$ denotes the set of regional maxima of the image $\mathbf{X}$. M_i is an image of the same size as $\mathbf{X}$, with zero in all other positions except for the pixels that compose the regional maximum M_i: here, where the gray value is the value of the maximum. Each regional maxima M_i has an extinction value ϵ_i corresponding to the increasing attribute being analyzed. The EF of $\mathbf{X}$ that preserves the n' maxima with highest extinction values, $\mathrm{EF}^{n'}(\mathbf{X})$, is given as

$$\mathrm{EF}^{n'}(\mathbf{X}) = R^{\delta}_{\mathbf{X}}(\mathbf{G}), \tag{2}$$

where $R^{\delta}_{\mathbf{X}}(\mathbf{G})$ is the reconstruction by dilation [34] of the mask image $\mathbf{X}$ from the marker image $\mathbf{G}$. The marker image $\mathbf{G}$ is given by

$$\mathbf{G} = \max_{i=1}^{n'} \{M'_i\}, \tag{3}$$

where max is the pixel-wise maximum operation. M'_1 is the maximum with the highest extinction value, M'_2 has the second highest extinction value, and so on.

SPACE OF SHAPES

Xu et al. [35] proposed building max-trees of tree-based image representations, i.e., building a max-tree from a max-tree or a max-tree from a tree of shapes [36]. This second max-tree construction takes into account a shape-attribute threshold on the first tree nodes, as opposed to thresholding image gray levels. Moreover, the connectivity rule is already defined by the initial tree, while, in the first tree construction, it is necessary to define a connectivity rule, which is usually either four-connectivity (vertical and horizontal neighbors of the pixel) or eight-connectivity (all neighbors of the pixel).

The second max-tree construction takes us to the space of shapes [35], allowing for the creation of a novel class of connected operators from the leveling family and more complex morphological analysis, such as the computation

of extinction values for nonincreasing attributes. This methodology has been used for blood vessel segmentation, a generalization of constrained connectivity [37], and hierarchical segmentation [38]. The space of shapes fluxogram is depicted in the red path of Figure 5. An example of the second max-tree construction using the aspect ratio attribute of the initial max-tree nodes for the second max-tree construction on a synthetic image is depicted in Figure 8. The nodes marked in blue are going to be preserved. The result of the filtering procedure in the space of shapes is depicted in Figure 9.

EXTINCTION FILTERS FOR NONINCREASING ATTRIBUTES

After building the max-tree of the initial tree representation (max-tree or min-tree in our case) and using a nonincreasing attribute (and, therefore, working on the space of shapes), the height of the attribute used to compute the

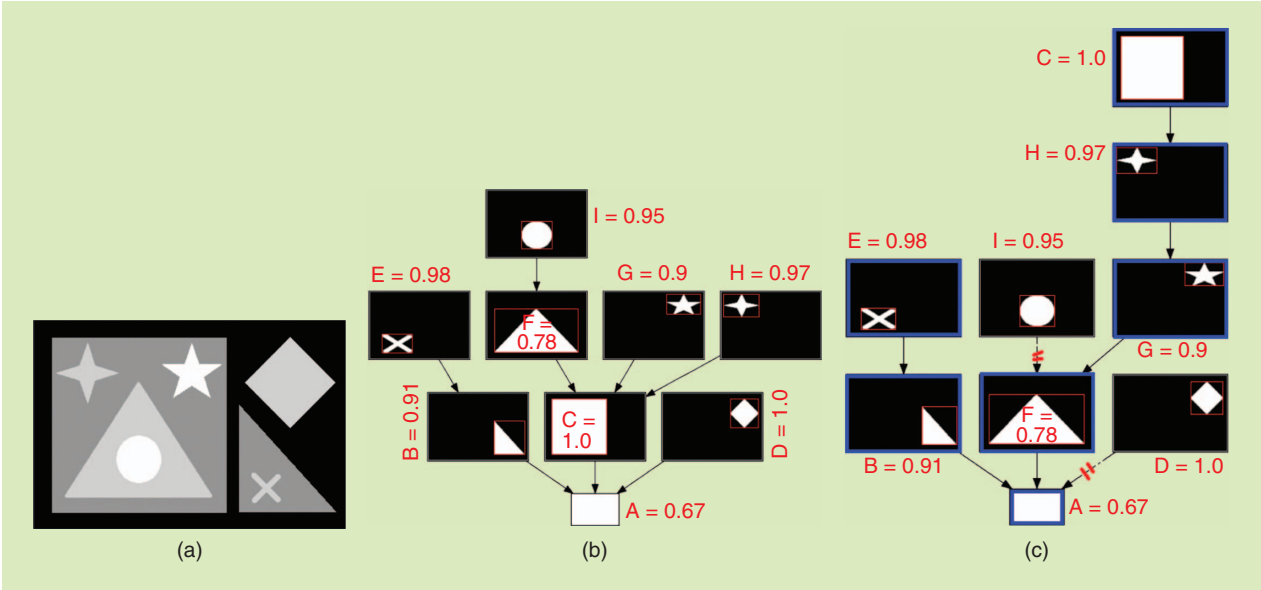


FIGURE 8. (a) A synthetic image, (b) its max-tree, and (c) a second max-tree using an aspect ratio as the attribute for the second tree construction.

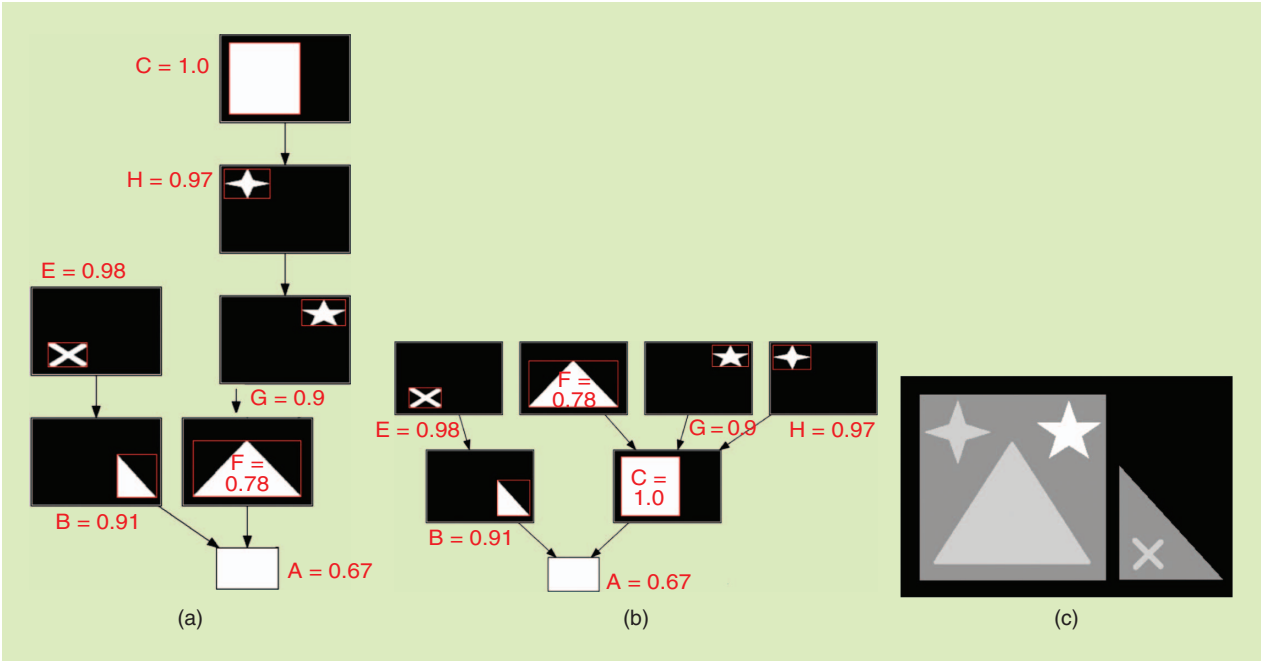


FIGURE 9. (a) The second max-tree after the filtering step, (b) the recovered initial max-tree after the filtering step, and (c) the resulting image after the filtering procedure.

second max-tree increases in this space. Therefore, it is possible to compute extinction values and EFs for nonincreasing attributes.

The procedure for computing EFs for nonincreasing attributes using the max-tree is as follows:

1) Build the image max-tree if filtering maxima (antiextensive) or the image min-tree if filtering minima (extensive).
2) Compute the second tree (max-tree) of the initial tree representation using the nonincreasing attribute chosen.
3) On the second tree, compute the height-extinction values for the nonincreasing attribute.
4) On the second tree, mark all nodes on the paths starting from the n' max-tree leaves with highest extinction values to the root.
5) On the second tree, filter the nodes that were not marked in the previous step.
6) Recover the initial tree (max-tree or min-tree) from the second tree.
7) Reconstruct the image.

EFs for nonincreasing attributes do not have the same extrema-preservation property as EFs for increasing attributes. They can also be seen as second max-tree increasing attribute EFs and belong to a class of filters known as *shape-based filters* [35].

EFS FOR NONINCREASING ATTRIBUTES DO NOT HAVE THE SAME EXTREMA-PRESERVATION PROPERTY AS EFS FOR INCREASING ATTRIBUTES.

EXTINCTION PROFILES FOR GRAY-SCALE IMAGES

To obtain EPs, several EFs (which are a sequence of thinning and thickening transformations) are used with progressively higher threshold values. In this manner, one can extract spatial and contextual information from the input data comprehensively [14]. Therefore, the EP for the input gray-scale image $\mathbf{X}$ is constructed by

$$\mathrm{EP}(\mathbf{X}) = \begin{Bmatrix} \Pi_{\phi^{\lambda_s}}, & s = (s-i+1), \quad \forall i \in [1,s]; \\ \Pi_{\gamma^{\lambda_s}}, & s = (i-s), \quad \forall i \in [s+1, 2s]. \end{Bmatrix}, \tag{4}$$

where $\Pi_{\phi^{\lambda}}$ is the thickening EP and $\Pi_{\gamma^{\lambda}}$ is the thinning EP computed with a generic-ordered criterion λ (also called *threshold* or *criteria*); s is the number of thresholds (i.e., criteria). In the set of ordered thresholds $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ for $\lambda_i, \lambda_j \in \lambda$ and $j \geq i$, the relation $\lambda_i \leq \lambda_j$ holds for thickening and $\lambda_i \geq \lambda_j$ holds for thinning. Note that, for the EP, the higher value of extrema can provide more detail. This contrasts with the conventional thresholding approach applied to APs, where the higher value of the threshold causes more smoothness. In other words, for the EP, the feature produced by the higher number of extrema is placed closer to the input image in the profile. There is a hierarchical relationship between the images generated by the EP, i.e., $\Pi_{\phi^{\lambda_1}} \geq \Pi_{\phi^{\lambda_2}} \geq \cdots \geq \Pi_{\phi^{\lambda_s}} \geq \Pi_{\gamma^{\lambda_s}} \geq \prod_{\gamma^{\lambda_{s-1}}} \geq \cdots \geq \prod_{\gamma^{\lambda_1}}$.

EXTINCTION PROFILES FOR HYPERSPECTRAL IMAGES

MPs, APs, and EPs, as discussed previously, produce several additional features from a single gray-scale image (i.e., the input image). It is possible to apply such profiles to all of the bands of the hyperspectral data individually and concatenate them. However, this results in producing many redundant features that need to be handled by the subsequent classifier. As a result, if the number of training samples is limited and the classification approach is not capable of handling high-dimensional data, the accuracies of the classification step will be downgraded due to the Hughes phenomenon [39]. This is the main reason that the number of bands is first reduced by using a dimensionality reduction approach. Then, a few informative features are fed to the MP, AP, or EP to produce spatial and contextual features. In more detail (and to generalize MPs, APs, and EPs from a gray-scale image to an HSI), we first need to reduce the dimensionality of the data from $E \subseteq \mathbf{Z}^{d_1}$ to $E' \subseteq \mathbf{Z}^{d_2}$ $(d_2 \leq d_1)$ using a generic transformation $\Psi: E \rightarrow E'$ (e.g., independent component analysis). Then, the EP can be performed on the most informative features $\mathbf{Q}_i$ $(i = 1, \ldots, d_2)$ among the extracted ones, which can mathematically be given as

$$\mathrm{EEP}(\mathbf{Q}) = \{\mathrm{EP}(\mathbf{Q}_1), \mathrm{EP}(\mathbf{Q}_2), \ldots, \mathrm{EP}(\mathbf{Q}_{d_2})\}. \tag{5}$$

Another extension of the EPs on an HSI is the extended multi-EP (EMEP) [15], which concatenates different extended EPs (EEPs) (e.g., area, height, volume, diagonal of bounding box, and standard deviation on different extracted features) into a single-stacked vector as

$$\mathrm{EMEP} = \{\mathrm{EEP}_{a_1}, \mathrm{EEP}_{a_2}, \ldots, \mathrm{EEP}_{a_w}\}, \tag{6}$$

where a_k, $k = \{1, \ldots, w\}$ denotes different types of attributes. It is easy to understand that, because different extinction attributes provide complementary spatial and contextual information, the EMEP is better able to extract spatial information than is a single EP [14], [15].

In [41], random forest ensembles and EMEPs are integrated to shape a spectral-spatial classification framework. In [41] and [42], EMEPs are used along with composite-kernel (CK) learning to perform spectral-spatial classification on HSIs. EMEPs have also been also investigated for fusing spectral and spatial features of hyperspectral and light detection and ranging lidar data using total variation [43], CK learning [44], deep CNNs [45], and sparse and low-rank feature fusion [46].

SOME NOTES ON COMPUTATIONAL TIME

EMEPs and EPs demand approximately the same computational time because the most time-consuming part is the construction of the max-tree and min-tree, which are computed only once for each gray-scale image [14], [15]. In terms of increasing attributes (i.e., a, bb, v, and h), the computation of both EPs and APs with the same size and for the same attribute leads to similar processing times. The only

difference is that EFs need to compute the extinction values for the attribute; however, this can be done simultaneously with the number of nodes and, consequently, does not add much to the processing time. In terms of nonincreasing attributes (standard), however, EPs need to construct a second max-tree (min-tree), which is not the case for APs. It should be noted that the second max-tree (min-tree) can be constructed much more quickly because its complexity is proportional to the number of nodes (m) of the first tree instead of the number of pixels (n) in the original image $(m << n)$ [47]. A detailed analysis of the computational complexity can be found in [14].

EXPERIMENTAL RESULTS

EXPERIMENTAL SETUP

For the experiments, a random forest with 200 trees is used to classify input features [see Figures 10(b), 11(b), and 12(b)]. Because an EMP considers only the attribute area, e.g., (a), to make a fair comparison with the EMP, we designed two scenarios: 1) EAPa and EEP$_a$, which only consider the attribute area, and 2) EMAPs and EMEPs, which consider five attributes (i.e., the area, height, volume, diagonal of the bounding box, and standard deviation) described in [14] and in the previous subsections. The EMP is composed of seven openings/closings by reconstruction with a circular SE of size 2, 4, 6, 8, 10, 12, and 14 pixels. The EMAP is generated using the following attributes and thresholds:

- λ_a = [100, 500, 1,000, 2,000, 3,000, 4,000, and 5,000]
- λ_h = [100, 500, 1,000, 2,000, 3,000, 4,000, and 5,000]
- λ_v = [100, 500, 1,000, 2,000, 3,000, 4,000, and 5,000]
- λ_{bb} = [10, 25, 50, 75, 100, 125, and 150]
- λ_{std} = [10, 20, 30, 40, 50, 60, and 70].

For the EMEP [see Figures 10(f) and 11(f)], the threshold values for all attributes are automatically set using $\lambda = 3^j, j = 0, 1, ..., s-1$, where s is set to 7 to produce the same number of features as the EMAP and EMP for each profile. The producer's accuracy has been used as class-specific accuracy, and its average value is reported as average accuracy (AA). Kappa and OA represent the kappa coefficient and overall accuracy, respectively.

In terms of the CASI University of Houston data, we have also run extra experiments using a support vector machine (SVM) [48] with weighted-summation CKs [see Figure 12(f)]. The weight parameter for both spectral and spatial kernels was set to 0.5.

RESULTS AND DISCUSSIONS

With respect to Tables 4–6, the following conclusions can be reached.

- Although the EMP, EAP$_a$, and EEP$_a$ include the same attribute (i.e., area) and number of features, the EEP$_a$ results in the highest classification accuracies. The main reason is that, as shown in [14], EPs are more effective than APs in terms of simplification for recognition. The main reason that the EEP$_a$ can improve the EMP in terms of classification accuracies is that the shape of the SE to produce EMPs is fixed, which imposes a constraint on modeling spatial structures within a scene.
- As can be seen, one needs to adjust a number of threshold values for the EMAP, which is a time-consuming

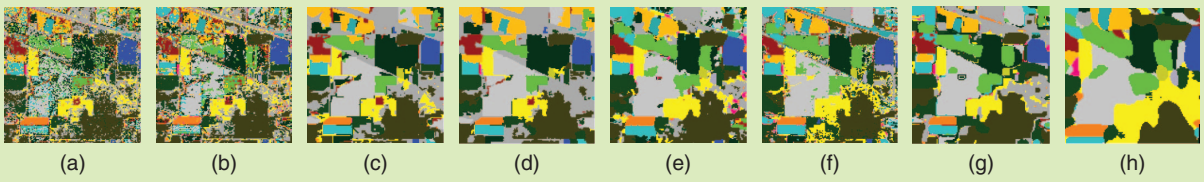


FIGURE 10. Classification maps obtained on AVIRIS Indian Pines data: (a) an SVM, (b) a random forest, (c) a binary partition tree (BPT) $\alpha = 0$, (d) a BPT $\alpha = 0.5$, (e) a multiscale adaptive SR (MASR), (f) an EMEP, (g) an MSVC with graph cuts, and (h) a Gabor-CNN.

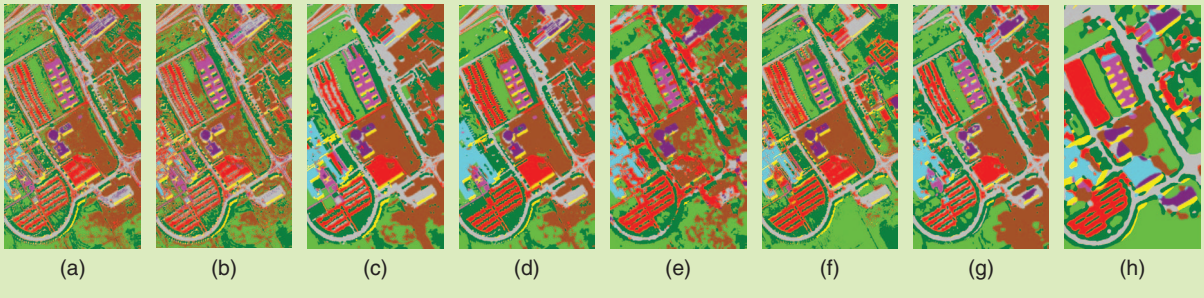


FIGURE 11. Classification maps obtained on ROSIS-03 Pavia University data: (a) an SVM, (b) a random forest, (c) a BPT $\alpha = 0$, (d) a BPT $\alpha = 0.5$, (e) an MASR, (f) an EMEP, (g) an MSVC with loopy belief propagation (LBP), and (h) a Gabor-CNN.

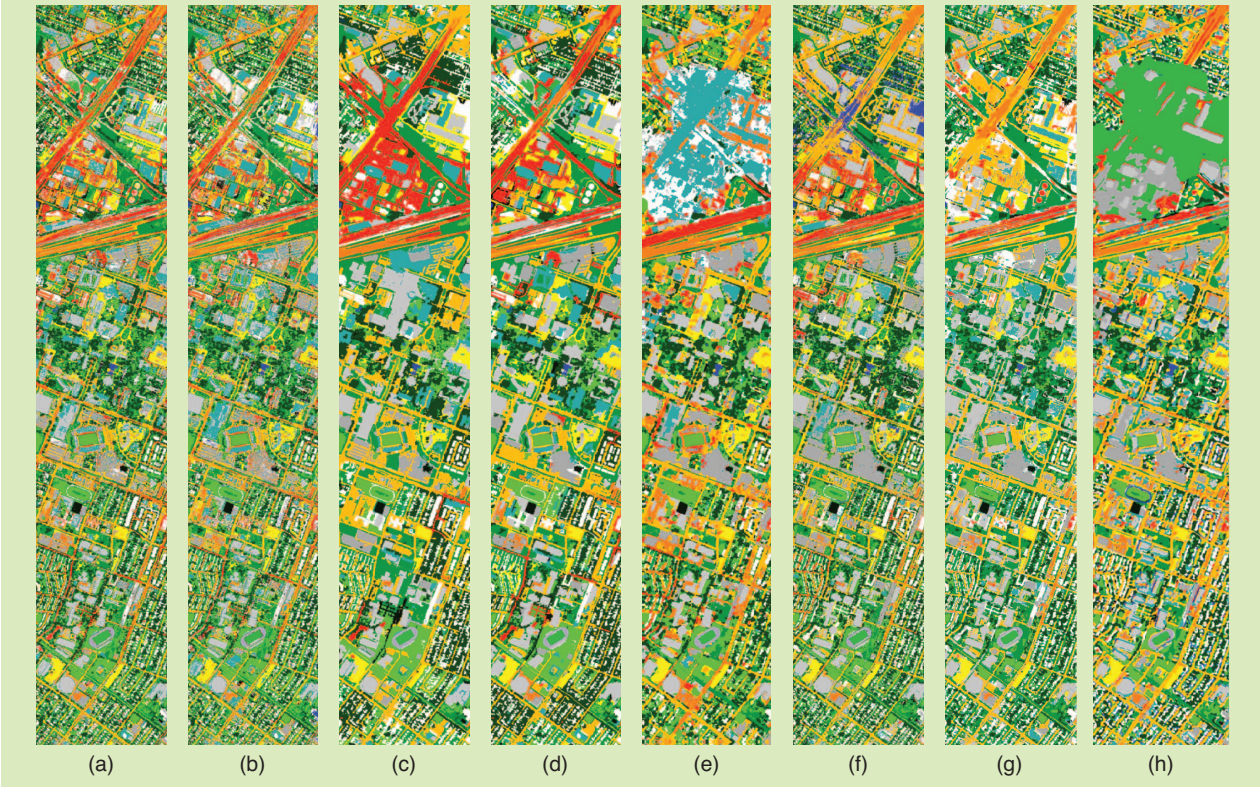


FIGURE 12. Classification maps obtained on the CASI University of Houston data: (a) an SVM, (b) a random forest, (c) a BPT $\alpha = 0$, (d) a BPT $\alpha = 0.5$, (e) an MASR, (f) a CK_{EMEP}, (g) an MSVC with LBP, and (h) a Gabor-CNN.

TABLE 4. THE AVIRIS INDIAN PINES DATA—THE CLASSIFICATION ACCURACIES [%] OBTAINED BY MATHEMATICAL-MORPHOLOGY-BASED APPROACHES AND THE CORRESPONDING CENTRAL PROCESSING UNIT (CPU) PROCESSING TIME (IN SECONDS).

CLASSES	RANDOM FOREST	EMP	EAP_a	EEP_a	EMAP	EMEP
1	55.13	85.04	86.56	85.40	84.10	87.43
2	55.61	92.98	90.56	96.05	95.66	95.79
3	82.61	96.74	96.74	98.37	98.37	98.91
4	85.68	93.51	95.53	95.08	95.53	95.53
5	79.91	96.84	96.13	94.84	96.84	95.98
6	94.08	99.54	99.09	98.41	99.32	99.09
7	78.21	90.85	89.43	92.70	90.63	92.81
8	59.35	89.83	87.30	92.68	89.16	93.13
9	60.82	86.17	85.99	89.18	86.88	87.06
10	95.06	98.15	97.53	98.15	98.77	99.38
11	87.86	97.03	98.23	95.10	94.13	97.03
12	54.85	99.09	98.79	98.48	98.18	99.09
13	100	100	100	100	100	100
14	53.85	94.87	94.87	94.87	94.87	94.87
15	81.82	100	100	100	100	100
16	100	100	80	100	100	100
OA	69.36	91.99	91.38	92.99	91.65	93.7
AA	76.55	95.04	93.54	95.58	95.15	96
Kappa	0.6541	0.9085	0.9015	0.9199	0.9046	0.9279
Time(s)	2	3	3	3	7	7

procedure. However, for MPs and EMEPs, one needs to adjust only the number of features.

- As discussed in the "Space of Shapes" section, the advantage of using an AP over an EP is its nonincreasing attributes (standard deviation). In this case, the EP needs to produce the second tree based on the space of the shapes.
- All of the EMPs, EAP_a, and EEP_a can provide results very swiftly.

Table 7 schematically compares the EMPs, EAPs, and EEPs in terms of classification accuracy, simplicity, and how close each is to being automatic. The best performance is shown using three bullets while the worst performance is represented by one bullet.

MARKOV RANDOM FIELDS

RANDOM FIELDS AND PROBABILISTIC GRAPHICAL MODELS

While mathematical morphology captures spatial information as part of the feature-extraction stage in a pattern recognition pipeline, a relevant family of methods for incorporating spatial information into the classification stage is based on random fields and probabilistic graphical

models. A random field is a stochastic process defined on some multidimensional domain such as, most remarkably, a 2-D pixel lattice. A probabilistic graphical model for an image makes use of a topological description based on graphs and a probabilistic description based on random fields to characterize the dependency properties of the image, usually involving suitable Markovian conditions [49]. These methodological tools make it possible to capture spatial dependencies in an HSI on a probabilistic basis.

Conventional image classifiers drawn from the pattern recognition literature (e.g., neural networks, random forests, or SVMs) are usually formalized under the assumption of independent and identically distributed (i.i.d.) pixels [50], [51]. While this noncontextual approach was found to be effective for remote-sensing data at moderate spatial resolutions, it is generally inadequate in the case of VHR, including VHR hyperspectral imaging [4]. Probabilistic graphical models allow non-i.i.d. pixels to be characterized in a Bayesian framework. From a signal-processing perspective, this is equivalent to moving from a white stationary model to a correlated and possibly nonstationary (or piecewise stationary) model for the spatial image behavior [4], [7]. From a machine-learning standpoint, classifiers based on probabilistic graphical models belong to the area of structured output learning, which includes algorithms whose output is supposed to exhibit dependency structures [52].

The main family of probabilistic graphical models extensively applied to HSI classification is given by MRFs, which provide powerful and flexible spatial-contextual models for prior distribution in Bayesian image analysis [53]–[55]. They have recently been used for HSI classification in conjunction with SVMs [3], [56]–[58], active learning [59], multinomial logistic regression (MLR) [56], [60], subspace projections [57], hierarchical statistical region merging [61], blind source separation and mean-field approximations [62], multidimensional wavelets [63], sparse modeling and Dirichlet distributions [64], and ensemble classifiers [65], [66]. In [61] and [67], MRF-based methods were also developed for HSI segmentation.

A further class of probabilistic graphical models is given by conditional random fields (CRFs), which model as Markovian the posterior distribution directly [68]. HSI

TABLE 5. THE ROSIS-03 PAVIA UNIVERSITY DATA—THE CLASSIFICATION ACCURACIES [%] OBTAINED BY MATHEMATICAL-MORPHOLOGY-BASED APPROACHES AND THE CORRESPONDING CPU PROCESSING TIME (IN SECONDS).

CLASSES	RANDOM FOREST	EMP	EAP_a	EEP_a	EMAP	EMEP
1	80.17	94.18	96.58	95.93	91.30	96.05
2	55.95	93.37	84.71	92.49	91.83	93.45
3	52.83	87.71	70.84	79.23	72.22	81.37
4	98.73	99.15	97.88	99.87	99.80	99.87
5	99.18	99.93	99.93	99.93	99.93	99.93
6	78.82	68.78	96.12	98.87	99.62	99.26
7	84.59	99.55	99.77	99.85	99.70	99.85
8	91.20	99.38	97.26	99.48	99.27	99.43
9	97.89	99.89	98.20	99.89	99.79	100
OA	71.51	91.82	90.33	94.82	93.52	95.46
AA	82.15	93.54	93.47	96.17	94.82	96.57
Kappa	0.6498	0.8912	0.8771	0.9332	0.9165	0.9407
Time(s)	8	9	8	8	21	23

TABLE 6. THE CASI UNIVERSITY OF HOUSTON DATA—THE CLASSIFICATION ACCURACIES [%] OBTAINED BY MATHEMATICAL-MORPHOLOGY-BASED APPROACHES AND THE CORRESPONDING CPU PROCESSING TIME (IN SECONDS).

CLASSES	RANDOM FOREST	EMP	EAP_a	EEP_a	EMAP	EMEP	CK_{EMEP}
1	83.38	75.02	76.45	74.83	77.59	77.78	80.53
2	98.40	88.06	76.97	77.54	80.64	76.88	98.03
3	98.02	99.80	99.80	100	100	100	100
4	97.54	84.38	83.62	82.67	85.42	82.77	96.02
5	96.40	95.83	95.64	95.93	96.12	96.02	99.05
6	97.20	94.41	95.10	95.80	95.10	95.80	95.10
7	82.09	70.43	71.92	72.20	72.29	72.95	78.08
8	40.65	84.14	84.43	79.39	70.28	82.05	81.20
9	69.78	63.74	59.40	61.19	57.79	63.17	81.68
10	57.63	55.98	66.51	66.99	67.86	67.57	61.39
11	76.09	82.45	78.84	84.06	74.86	82.83	86.62
12	49.38	78.19	77.91	85.11	81.36	84.53	89.53
13	61.40	73.33	71.93	75.79	77.19	73.68	78.95
14	99.60	99.60	99.19	99.60	99.19	99.60	100
15	97.67	96.41	99.37	99.37	97.89	98.94	98.52
OA	77.47	80.01	79.5	80.32	78.92	80.83	86.64
AA	80.34	82.78	82.47	83.36	82.23	83.64	88.31
Kappa	0.7563	0.7834	0.777	0.7866	0.7721	0.792	0.8831
Time	26	23	21	21	57	60	162

TABLE 7. THE PERFORMANCE EVALUATION OF EMP, EAP, AND EEP IN TERMS OF CLASSIFICATION ACCURACIES, SIMPLICITY, AND HOW CLOSE IT IS TO AUTOMATIC.

TECHNIQUES	ACCURACY	AUTOMATION	SPEED
EMP	••	••	•••
EAP	••	•	•••
EEP	•••	•••	•••

The best performances are shown using three bullets, while the worst performance is represented by one bullet.

classification methods have recently been developed using CRFs along with SVMs and Mahalanobis distances [8], [69], MLR [71], decision tree ensembles [71], extreme learning machines [72], deep belief networks (DBNs) [73], segmentation and object-based image analysis [74], game theory [75], and adaptive differential evolution for decision fusion with lidar data [76]. Here, we focus on MRFs, first reviewing the basics and then discussing advanced methods that integrate the MRF and SVM approaches into HSI classification.

SETTING A NEIGHBORHOOD SYSTEM ON THE PIXEL LATTICE IS EQUIVALENT TO CONSTRUCTING AN UNDIRECTED GRAPH IN WHICH EACH NODE IS A PIXEL AND EACH EDGE IS DETERMINED BY A PAIR OF NEIGHBORING PIXELS.

KEY IDEAS OF MARKOV RANDOM FIELDS MODELING

MRF models formalize spatial interactions on a local basis using neighborhoods. A neighborhood system is defined on the 2-D regular lattice of the n image pixels if, for every ith pixel, a subset ∂i of neighboring pixels is specified $(i = 1, 2, \ldots, n)$. The neighborhood relation is supposed to be symmetric (i.e., if one pixel is a neighbor to another, then the opposite holds as well) and irreflexive (i.e., no pixel is a neighbor to itself) [54]. Classical examples include the first- and second-order neighborhood systems, in which ∂i is the set of the four pixels adjacent to the ith pixel and the eight pixels surrounding it, respectively. Higher-order or adaptive neighborhoods can be defined as well [54].

Setting a neighborhood system on the pixel lattice is equivalent to constructing an undirected graph in which each node is a pixel and each edge is determined by a pair of neighboring pixels. Given this topological structure, the random field of the labels of all pixels is an MRF if its joint probability distribution is strictly positive and if the following Markovian property holds $(i = 1, 2, \ldots, n)$ [53], [54]:

$$P(\gamma_i \mid \gamma_j, j \neq i) = P(\gamma_i \mid \gamma_j, j \in \partial i). \tag{7}$$

While the strict positivity of the joint distribution is a technical assumption meant to ensure mathematical tractability [55], (7) means that the distribution of the label of each pixel, given the labels of all other pixels, is equivalent only to conditioning the labels of the neighbors—a condition that extends the analogous properties of 1-D Markov chains [77] to 2-D images.

In an HSI classification problem, establishing an MRF model for the labels has a remarkable impact on Bayesian decision rules. Collecting all HSI data in the $n \times d$ matrix $\mathbf{X}$ and all labels in the n-dimensional discrete vector $\mathbf{Y}$ (see the "Notations" section), it is possible to prove through the Hammersley-Clifford theorem that, under mild assumptions, the joint posterior distribution $P(\mathbf{Y} \mid \mathbf{X})$ of all the labels given all the image data is a Gibbs distribution proportional to $\exp[-U(\mathbf{Y} \mid \mathbf{X})]$, where U, named *energy*, is defined locally according to the neighborhood system [55].

Focusing for the sake of clarity on a common subclass of MRF models (i.e., the MRFs with only nonzero pairwise clique potential), the functional form of the energy [54], [78] is

$$U(\mathbf{Y} \mid \mathbf{X}) = \sum_{i=1}^{n} D_i(\mathbf{x}_i, \gamma_i) + \beta \sum_{i=1}^{n} \sum_{j \in \partial i} V_{ij}(\gamma_i, \gamma_j), \tag{8}$$

where $D_i(\mathbf{x}_i, \gamma_i)$ is a pixel-wise (or unary) term associated with the spectral feature vector $\mathbf{x}_i$ and the label of the ith pixel; $V_{ij}(\gamma_i, \gamma_j)$, named *pairwise potential*, determines the spatial relation among the ith and jth pixels and their labels; and β is a parameter $(i = 1, 2, \ldots, n; j \in \partial i)$. Based on (8), the Bayesian maximum a posteriori rule is equivalent to minimizing the energy $U(\mathbf{Y} \mid \mathbf{X})$ with respect to $\mathbf{Y}$, given the input HSI $\mathbf{X}$. Within this minimization, the pixel-wise spectral information described by D_i and the spatial interactions encoded by V_{ij} are fused for spectral-spatial classification purposes, while β weighs the tradeoff between the two contributions.

The unary term generally comes from the pixel-wise negative class-conditional log-likelihood of the spectral data, estimated through parametric [3], [64], [79], [80] or nonparametric algorithms [59], [63], [65]. The pairwise potential determines the adopted MRF model and is defined to favor the desired spatial behavior. Well-known models can be used to favor smooth, edge-preserving, isotropic or anisotropic, and stationary or nonstationary behaviors [53], [54], [65], [80]. More advanced MRFs also allow multiscale, multiresolution, multisensor, and multitemporal fusion; hierarchical structures; segmentation results; or textures to be incorporated [4], [61], [81]–[83]. This remarkable flexibility is among the reasons for the current prominence of MRF approaches to spectral-spatial classification.

Another major reason is the availability of computationally efficient energy- minimization methods, which rely on graph cut (GC) and belief propagation concepts and have attracted increasing interest during the last decade. In the case of binary classification, GCs make use of a reformulation based on the min-flow/max-cut theorem to reach a global energy minimum with low-order polynomial complexity, provided the pairwise potential satisfies a suitable condition [84]. In the multiclass case, GC algorithms iteratively define a sequence of suitable binary problems and, under appropriate assumptions on the pairwise potential, converge with local minima that have strong optimality properties [78], [85], [86].

Belief propagation-type methods formalize the intuitive idea of passing messages along the graph to decrease the energy [87]. In particular, the max-product loopy belief propagation (LBP) technique operates on graphs with loops, such as those that are usually associated with MRF neighborhoods. It may generally not converge, but when it does, it obtains a local minimum with good optimality properties [87], [88]. The complexity of efficient formulations of LBP

is linear with respect to the numbers of pixels and classes [89]. The tree-reweighted (TRW) message-passing method combines belief propagation with the construction of suitable spanning trees [90] and can be endowed with specific convergence properties by using an appropriate sequential formulation (TRW-S) [91]. The complexity of this formulation is linear with respect to the number of its edges in the graph, of its classes, and of its iterations [91].

Among earlier methods, which have been consolidated since the 1980s, we recall simulated annealing (SA) and iterated conditional mode (ICM). SA makes use of Gibbs or Metropolis random sampling and converges to a global minimum under certain conditions, although with long computation times [55]. ICM is a deterministic algorithm that has a much lower computational burden but converges to a generic local minimum and may be sensitive to initialization [92].

Recent applications of energy minimization methods to hyperspectral imaging can be found, e.g., in [58], [60], and [64]–[66]. For more details on MRF and energy minimization, we refer readers to [53] and [54].

SUPPORT VECTOR MACHINES, MARKOV RANDOM FIELDS, AND ENERGY MINIMIZATION

Among the noncontextual classifiers, SVMs are known for their remarkable generalization capability even in applications for high-dimensional feature spaces—a property that justifies their consolidated use for spectral HSI classification. Hence, combining SVMs with contextual MRF models nicely fits the requirements of spectral-spatial HSI classification and has received substantial attention [57], [58], [61], [63]. Here, we do not review the basics of SVMs; rather, we refer the reader to well-known textbooks such as [51] and [93] and only note that merging SVMs and MRFs is not straightforward because the latter are framed within probabilistic Bayesian modeling, whereas the former are non-Bayesian learning machines.

A common workaround is to postprocess the SVM discriminant function through the algorithms in [94] and [95], which use parametric modeling, maximum likelihood, and numerical analysis concepts to approximate pixel-wise posteriors. The resulting probabilistic output is plugged into the unary energy. This approach is computationally efficient and has recently led to accurate results with hyperspectral images [57], [61], [63]. However, it methodologically mixes i.i.d. and non-i.i.d. assumptions in the parameter estimation and MRF modeling stages, respectively.

An alternate approach, which aims at merging the analytical formulations of SVM and MRF, has been proposed in [58] and [96]. Focusing on binary classification and denoting the two classes as +1 and –1, let the random field of the class labels be an MRF with pairwise potential V_{ij} and weight parameter β, and let K be a kernel. By definition, this means that computing $K(\mathbf{x},\mathbf{x}')$ $(\mathbf{x},\mathbf{x}' \in \mathbb{R}^d)$ is equivalent to evaluating an inner product in some transformed space $\mathcal{F}$ [51]. The key idea of the approach in [58] and [96] is to apply the MRF minimum-energy rule directly in the space $\mathcal{F}$ implied by the kernel. On one hand, this is not straightforward, because $\mathcal{F}$ may be infinite-dimensional (it is a separable Hilbert space [97]) and is normally not even specified explicitly in a kernel machine [51]. On the other hand, this approach leads to integrating the SVM and MRF into a unique framework, where energy-minimization algorithms can be formulated for spectral-spatial classification.

More precisely, two main results have been proven in this framework. First, under mild assumptions, the difference ΔU_i between the energy contribution associated with the ith pixel and with label $\gamma_i = -1$ and that associated with $\gamma_i = 1$ can be expressed as an SVM-like kernel expansion $(i = 1, 2, \ldots, n)$

$$\Delta U_i = \sum_{s \in \mathcal{S}} \alpha_s \gamma_s K^{\mathrm{MRF}}(\mathbf{x}_i, \varepsilon_i; \mathbf{x}_s, \varepsilon_s) + b, \tag{9}$$

provided that a case-specific Markovian kernel K^{MRF} and a spatial additional feature ε_i are used $(\mathbf{x},\mathbf{x}' \in \mathbb{R}^d; \varepsilon, \varepsilon' \in \mathbb{R}; i = 1, 2, \ldots, n)$ [58],

$$\begin{aligned} K^{\mathrm{MRF}}(\mathbf{x},\varepsilon;\mathbf{x}',\varepsilon') &= K(\mathbf{x},\mathbf{x}') + \beta\varepsilon\varepsilon' \\ \varepsilon_i &= \sum_{j \in \partial i} [V_{ij}(-1,\gamma_j) - V_{ij}(1,\gamma_j)], \end{aligned} \tag{10}$$

and that the set $\mathcal{S}$ of support vectors and the coefficients α_s $(s \in \mathcal{S})$ and b are computed by training an SVM with kernel K^{MRF} [58]. The labels of the support vectors in (9) are obviously known from the training set.

Secondly, if the pairwise potential satisfies the additional condition that, for each ith pixel, the sum $\Sigma_{j \in \partial i}[V_{ij}(-1,\gamma_j) + V_{ij}(1,\gamma_j)]\Sigma_{j \in \partial i}[V_{ij}(-1,\gamma_j) + V_{ij}(1,\gamma_j)]$ is a constant independent from the labels of the neighbors, then the energy $U(\mathbf{Y} \mid \mathbf{X})$ can be written as $-\Sigma_{i=1}^{n} \gamma_i \Delta U_i$, or, equivalently, in terms of the following unary and pairwise terms $(i = 1, 2, \ldots n; j \in \partial i)$ [96],

$$\begin{aligned} D_i^{\mathrm{SVM}}(\mathbf{x}_i,\gamma_i) &= -\gamma_i \sum_{s \in \mathcal{S}} \alpha_s \gamma_s K(\mathbf{x}_i,\mathbf{x}_s) - b\gamma_i \\ V_{ij}^{\mathrm{SVM}}(\gamma_i,\gamma_j) &= V_{ij}(\gamma_i,\gamma_j) - V_{ij}(-\gamma_i,\gamma_j), \end{aligned} \tag{11}$$

and of a suitable weight parameter. In general, the aforementioned condition on the pairwise potential is a restriction. Nevertheless, it is satisfied by several popular MRF models, such as the widely used spatial Potts model [53] or the multitemporal model in [82], [83], and [98].

In (9) and (10), the additional feature ε_i is determined by the adopted spatial MRF (as described by the related pairwise potential) and by the Markovian kernel merging spectral and spatial terms in a linear combination. In (11), this formulation even provides a full representation of the Markovian energy associated with a classification problem in the transformed space implied by the kernel. Comments

on the assumptions behind these theorems and the related proofs can be found in [58] and [96].

Given the integration of the SVM and MRF in (9) and (11), energy-minimization algorithms can be formulated to design spectral-spatial classifiers. Several such algorithms (e.g., SA and ICM) can be entirely expressed in terms of the energy difference ΔU_i, so that they can be combined with the kernel expansion (9) [58]. More generally, (11) provides a full representation of the global energy U, which makes it possible to apply arbitrary energy-minimization methods, including GC, LBP, and TRW [96]. The resulting classifiers iteratively alternate: 1) the update of the additional spatial feature as a function of the current classification map, 2) the training of an SVM with the Markovian kernel, and 3) the update of the classification map through the considered energy minimization algorithm. We recall that the complexity of current algorithms for SVM training is generally at least quadratic with respect to the number of training samples [99] and that the complexity of the calculation of D_i^{SVM} on the entire image is linear with respect to the numbers of pixels and of support vectors.

IN PARTICULAR, THESE RESULTS POINT OUT THE ABILITY OF THE MSVC FRAMEWORK TO SIMULTANEOUSLY BENEFIT FROM THE SPATIAL MODELING CAPABILITY OF MRFS, THE FLEXIBLE NONPARAMETRIC FORMULATION OF KERNEL LEARNING, AND ITS EFFECTIVENESS IN HIGH-DIMENSIONAL FEATURE SPACES.

The resulting classification methods will be collectively termed *Markovian support vector classifiers* (*MSVCs*) in subsequent sections. More algorithmic details as well as comments on the automatic optimization of the parameters of the methods (i.e., β and the SVM hyperparameters) can be found in [58] and [96].

EXPERIMENTAL RESULTS

EXPERIMENTAL SETUP

The MSVC framework is tested with the considered data sets in conjunction with three energy-minimization algorithms, i.e., GC, LBP, and TRW. Multiclass labeling is accomplished using the one-versus-one approach [93], i.e., each minimization method is applied to a separate energy of the form (11) with regard to each pair of distinct classes. Accordingly, the GC approach is applied to binary subproblems in its max-flow/min-cut formulation. Regarding LBP, for each iteration of the MSVC approach, both variants discussed in [101] (which differ in the schedules for exchanging messages among the pixels) are used, and the solution with the lower energy is selected. In the case of TRW, TRW-S is used to favor a convergent behavior.

The results of MSVC are discussed in comparison with those of state-of-the-art contextual HSI classification methods based on MRF or kernel concepts: 1) the MRF- and kernel-based method in which approximate pixel-wise posteriors are derived from the output of a purely spectral SVM through the algorithm in [95] and are plugged into the unary term (*MRF-SVM-Post* in subsequent sections); 2) the MRF- but not the kernel-based classifier in which unaries are computed through a Gaussian class-conditional model (*MRF-Gauss* subsequently), and 3) the kernel- but not MRF-based contextual SVM (CSVM) technique in [101]. A CSVM incorporates spatial information into an SVM for HSI classification using suitable embeddings in a reproducing kernel Hilbert space. A somewhat similar analytical formulation can be achieved using graph-kernel concepts [102]. Alternately, as discussed in the "Mathematical-Morphology-Based Spectral-Spatial Classifiers" section, if spatial information is characterized in the feature extraction rather than

TABLE 8. THE AVIRIS INDIAN PINES DATA—THE CLASSIFICATION ACCURACIES [%] OBTAINED BY THE MSVC FRAMEWORK USING THREE ENERGY-MINIMIZATION ALGORITHMS AND BY PREVIOUS SPECTRAL-SPATIAL CLASSIFIERS BASED ON MRF AND KERNEL APPROACHES.

	MSVC			MRF-SVM-POST	MRF-GAUSS	CSVM
CLASSES	GC	TRW-S	LBP			
1	88.01	91.26	93.50	86.34	77.31	83.31
2	94.64	96.81	96.68	83.80	80.10	92.47
3	97.28	97.83	97.83	100	95.11	95.65
4	95.30	93.51	93.74	96.20	91.05	95.97
5	96.56	97.42	97.99	97.85	93.26	89.53
6	91.12	94.31	92.03	99.32	98.18	95.22
7	87.15	81.70	87.80	86.93	93.25	91.07
8	92.47	90.74	89.04	88.50	63.98	86.64
9	86.35	84.04	85.11	96.45	88.48	89.36
10	100	99.38	99.38	99.38	99.38	99.38
11	91.80	92.93	92.36	88.59	94.21	97.27
12	92.73	92.42	92.12	100	76.06	93.64
13	100	93.33	97.78	100	95.56	100
14	82.05	76.92	74.36	92.31	79.49	94.87
15	100	81.82	81.82	100	81.82	100
16	100	100	100	100	80.00	100
OA	91.66	91.40	91.80	90.54	82.04	90.35
AA	93.47	91.53	91.97	94.73	86.7	94.02
Kappa	0.9044	0.9014	0.9062	0.8919	0.7968	0.8900

the classification stage, composite kernels can be used to fuse spectral and spatial features [41], [42], [44], [48].

Before applying MRF-Gauss, dimensionality reduction is performed through nonparametric-weighted feature extraction [103] to prevent the impact of the Hughes phenomenon on Gaussian density estimation. In all kernel methods, the Gaussian radial basis function kernel is used, and the hyperparameters of the SVM are automatically optimized by using the method in [58], which numerically minimizes the span bound on the SVM error [104]. In all Markovian methods, the Potts model is used, i.e., $V_{ij}(\gamma_i,\gamma_j) = -1$ if $\gamma_i = \gamma_j$ and $V_{ij}(\gamma_i,\gamma_j) = 0$ otherwise. We recall that, with this choice, both conditions for the applicability of (11) and for the convergence of GCs to a global minimum hold true. The parameter β is automatically optimized using the method in [98], which is based on the Ho-Kashyap's algorithm.

RESULTS AND DISCUSSION

The accuracies obtained by the previously described methods on the test samples of the three data sets are collected in Tables 8–10. In terms of the state-of-the-art MRF-SVM-Post and MRF-Gauss methods, only the results of the energy minimization algorithm that provides the highest OA are shown in these tables (for brevity).

The MSVC framework obtains OA values of approximately 91–92%, 82–87%, and 85–87% in the cases of the AVIRIS Indian Pines, ROSIS-03 Pavia University, and CASI University of Houston data, respectively. Accurate results are also generated by the two previous contextual kernel methods. Yet MSVC obtains higher OA values than MRF-SVM-Post in the case of all three data sets— and higher than CSVM in the application to two data sets using all energy-minimization algorithms and to the third data set using one of these algorithms. MRF-Gauss, which is based on a parametric Gaussian model for the class-conditional statistics, achieves lower accuracies than all of the previously mentioned nonparametric kernel methods.

On one hand, all of the considered Markovian approaches yield improvements over purely spectral classifiers (e.g., random forest and SVM in Tables 4–6 and Tables 11–13), an expected conclusion that has largely been demonstrated in the literature (e.g., [4], [7], and [105]). On the other hand, the experimental results confirm that MRF models, and especially their combination with kernel machines, are powerful tools for HSI classification. In particular, these results point out the ability of the MSVC framework to simultaneously benefit from the spatial modeling capability of MRFs, the flexible nonparametric formulation of kernel learning, and its effectiveness

TABLE 9. THE ROSIS-03 PAVIA UNIVERSITY DATA—THE CLASSIFICATION ACCURACIES [%] OBTAINED BY THE MSVC FRAMEWORK USING THREE ENERGY-MINIMIZATION ALGORITHMS AND BY PREVIOUS SPECTRAL-SPATIAL CLASSIFIERS BASED ON MRF AND KERNEL APPROACHES.

CLASSES	MSVC GC	TRW-S	LBP	MRF-SVM-POST	MRF-GAUSS	CSVM
1	95.29	96.73	96.89	93.99	84.84	92.56
2	67.45	77.47	69.58	67.73	72.56	73.60
3	80.72	81.93	82.59	70.80	65.12	71.68
4	95.19	95.36	96.91	96.53	96.63	98.97
5	100	98.65	100	99.91	99.91	100
6	98.25	96.85	97.86	97.44	92.34	96.35
7	95.51	81.45	85.22	92.46	91.95	92.46
8	95.07	97.68	97.35	98.10	94.59	97.41
9	90.19	93.46	86.79	99.50	98.99	95.09
OA	82.35	86.93	83.60	82.19	81.78	84.58
AA	90.85	91.06	90.35	90.72	88.55	90.90
Kappa	0.7769	0.8312	0.7917	0.7745	0.7676	0.8031

TABLE 10. THE CASI UNIVERSITY OF HOUSTON DATA—THE CLASSIFICATION ACCURACIES [%] OBTAINED BY THE MSVC FRAMEWORK USING THREE ENERGY-MINIMIZATION ALGORITHMS AND BY PREVIOUS SPECTRAL-SPATIAL CLASSIFIERS BASED ON MRF AND KERNEL APPROACHES.

CLASSES	MSVC GC	TRW-S	LBP	MRF-SVM-POST	MRF-GAUSS	CSVM
1	82.91	83.10	82.81	82.24	80.34	83.76
2	100	100	100	98.31	97.74	97.65
3	99.21	99.80	99.80	99.80	99.01	99.80
4	97.35	97.35	98.30	98.86	93.28	98.77
5	99.81	99.91	99.91	98.39	95.74	99.43
6	99.30	97.90	98.60	98.60	90.91	100
7	91.70	91.79	92.26	88.62	69.78	78.17
8	53.85	57.08	55.84	48.34	75.59	47.86
9	84.70	86.59	89.24	83.19	82.25	81.78
10	76.64	73.65	77.41	74.61	46.04	75.87
11	74.48	72.01	74.76	86.81	80.46	84.16
12	83.09	79.25	81.27	76.27	82.04	75.50
13	83.51	80.70	82.11	71.93	76.84	84.56
14	100	100	100	99.60	99.19	100
15	97.25	91.75	94.93	97.46	92.39	98.31
OA	86.07	85.48	86.60	85.05	82.04	84.29
AA	88.25	87.39	88.48	86.87	84.11	87.04
Kappa	0.8489	0.8424	0.8546	0.8379	0.8062	0.8300

in high-dimensional feature spaces. This belief is also confirmed by previous experiments, suggesting that no feature reduction was generally necessary prior to MSVC (see [58]), and by a visual analysis of the classification maps, which points out the spatial regularity favored by MRF modeling [Figures 10(g), 11(g), and 12(g)].

The three considered energy-minimization algorithms exhibit similar behaviors overall. They obtain very similar accuracies in the cases of the AVIRIS Indian Pines and CASI University of Houston data sets, while in the case of ROSIS-03 Pavia data set, TRW-S reaches an OA 3–4% higher than that of GC and LBP. The high accuracies achieved confirm the effectiveness of current advanced GC and message-passing techniques for MRF energy minimization in an HSI classification task—a conclusion that has been drawn in numerous image processing and computer vision applications [101]. The performances obtained using all three methods also confirm the flexibility of the MSVC framework in incorporating arbitrary energy minimization algorithms. In [58] and [98], this flexibility is combined with the potential to fully automate the resulting classifiers through the previously mentioned parameter optimization methods. We also recall that a previous MSVC formulation using ICM was originally developed in [58] and experimentally validated with various data modalities, including hyperspectral imagery.

SEGMENTATION METHODS PARTITION AN IMAGE INTO NONOVERLAPPING HOMOGENEOUS REGIONS WITH RESPECT TO SOME CRITERION OF INTEREST OR HOMOGENEITY CRITERION (E.G., BASED ON THE INTENSITY OR ON THE TEXTURE).

TABLE 11. THE AVIRIS INDIAN PINES DATA—THE CLASSIFICATION ACCURACY VALUES OBTAINED BY BINARY PARTITION TREES.

CLASSES	SVM	BPT $\alpha = 0$	BPT $\alpha = 0.5$
1	53.25	57.66	54.99
2	52.17	59.70	58.16
3	83.70	97.83	100
4	87.25	95.53	95.53
5	82.50	86.23	91.96
6	92.03	99.54	99.54
7	72.11	98.58	98.69
8	47.56	80.65	86.35
9	71.63	89.54	96.81
10	96.91	99.38	98.77
11	79.34	90.19	90.43
12	72.73	99.70	99.70
13	95.56	97.78	100
14	56.41	97.44	94.87
15	81.82	90.91	100
16	100	0	100
OA	65.64	82.46	84.36
AA	76.56	83.79	91.61
Kappa	0.6141	0.8013	0.8224

SEGMENTATION

An important family of methods involves the segmentation of images and the classification of each of the individual segments. Segmentation methods partition an image into nonoverlapping homogeneous regions with respect to some criterion of interest or homogeneity criterion (e.g., based on the intensity or on the texture) [106]. Hence, each region in the segmentation map can be seen as a connected spatial neighborhood for all of the pixels within this region. One of the pioneering spectral-spatial techniques belongs to this category: the well-known ECHO classifier [107], which has been extensively used by the remote-sensing community. It is based on region growing to find homogeneous groups of adjacent pixels that are then classified as single objects by a Gaussian maximum likelihood method. Since then, different techniques have been proposed for HSI segmentation, such as watershed, partitional clustering, and hierarchical segmentation (HSeg) [108]–[110]. From a segmentation map, any pixel-wise classifier and majority voting can be applied to combine spectral and spatial information; for every region in the segmentation map, all pixels are assigned to the most frequent class within this region, based on pixel-wise classification results [110].

It is, however, a challenging task to perform HSI segmentation automatically. The performance is highly dependent on both the measure of region homogeneity and the algorithm's parameters. Several alternatives have been proposed to deal with this challenge. Tarabalka et al. [10], [111] proposed performing a marker-controlled segmentation for this purpose. The classification probabilities are used to automatically select the most reliably classified pixels (i.e., pixels with the highest probability belonging to the assigned class). The classification map is then obtained by building a minimum-spanning forest from the image graph

TABLE 12. THE ROSIS-03 PAVIA UNIVERSITY DATA—THE CLASSIFICATION ACCURACY VALUES OBTAINED BY BINARY PARTITION TREES.

CLASSES	SVM	BPT $\alpha = 0$	BPT $\alpha = 0.5$
1	84.21	96.94	94.14
2	69.95	71.27	72.27
3	67.71	82.26	99.89
4	98.08	97.73	98.15
5	99.47	100	99.47
6	93.39	97.97	97.99
7	90.42	99.90	96.23
8	92.87	95.63	99.32
9	97.48	94.01	88.18
OA	80.62	84.80	85.74
AA	88.17	92.87	93.96
Kappa	0.7542	0.8066	0.8185

rooted on the selected markers. This method has a principle similar to the widely used MRF-based GC approach [112] presented in the previous section (in the sense that the classification probabilities serve as the basis for the following spatial regularization process).

The second widely used class of approaches for automatic segmentation consists of first building a hierarchy of segmentations according to differing levels of details and then selecting from this hierarchy the regions at different scales that correspond to the objects of interest. Valero et al. proposed using a binary partition tree (BPT) model for this purpose [113]. In this method, a BPT is first constructed by iteratively clustering similar regions based on a criterion specifically designed for HSI. Each BPT node is then modeled by its mean spectrum and classified using an SVM. A misclassification rate is then computed for each node and can be understood as the error incurred by assigning the entire node to the wrong class. Finally, a spectral-spatial classification map is built in a bottom-up traversal of the tree by extracting regions with a low misclassification rate. Another BPT-based model has recently been proposed in [114] and extended in [115]; here, the object-based classification problem is formulated as an energy-minimization task. While the GC-based approach has been mentioned in the previous section, we detail in the following section the BPT-based segmentation method and demonstrate its performances for the hyperspectral data sets.

BINARY PARTITION TREE MODEL

BPTs were studied by Salembier and Garrido [116] as a way of representing a set of meaningful image regions in a compact and structured manner. A BPT is a hierarchical partition of an image: the root node represents the entire image, the following level partitions the image into two nonoverlapping regions, and so on. The BPT is constructed in a bottom-up fashion, by iteratively clustering pairs of similar regions together. The starting point is an initial subdivision of the image represented by a region adjacency graph (RAG), where every node conveys a region and the edges link spatial neighbors.

The typical initial RAG is the pixel grid, although nothing prevents the approach from being used with other inputs as well. Every edge in the RAG is labeled with a dissimilarity value that compares the two associated regions. BPTs are typically constructed using a global and mutual best-fitting region-merging approach [117], and, at each iteration, the two most similar regions in the current subdivision are merged. When a merge occurs, a new region is added to the BPT, connected to its two corresponding children. The process concludes when there are no more edges left in the RAG.

Once a tree is constructed, an exhaustive segmentation of the image can be obtained by performing a horizontal "cut" on the structure (see Figure 13). In this procedure, commonly referred to as *pruning*, branches can be selected at different scales, which is an inherent advantage of such hierarchical structures.

The key elements that define the behavior of a BPT are the region model, i.e., how regions are represented, and the dissimilarity function, i.e., the function of comparing the region models, which are used to define the priority of the merges during tree construction.

REGION MODEL

There are essentially two alternatives for representing the spectrum of each region: parametric and nonparametric models. Nonparametric models (e.g., per-band histograms of the pixel values) have proven to be a better approach than their parametric counterpart (e.g., average spectrum) because they represent the real, observed distributions and can thus describe the internal variability of a region [113].

TABLE 13. THE CASI UNIVERSITY OF HOUSTON DATA—THE CLASSIFICATION ACCURACY VALUES OBTAINED BY BINARY PARTITION TREES.

CLASSES	SVM	BPT $\alpha = 0$	BPT $\alpha = 0.5$
1	83.01	82.05	83.10
2	96.80	83.65	82.99
3	99.60	100	100
4	97.82	87.78	94.03
5	96.12	92.71	99.43
6	94.41	95.10	95.10
7	86.94	87.50	91.23
8	51.57	46.53	51.29
9	81.40	93.39	88.20
10	66.51	42.57	64.29
11	81.59	99.05	94.02
12	60.04	57.83	73.97
13	62.81	68.42	62.46
14	100	100	100
15	98.10	100	100
OA	81.91	79.69	83.78
AA	83.78	82.44	85.34
Kappa	0.8040	0.7799	0.8242

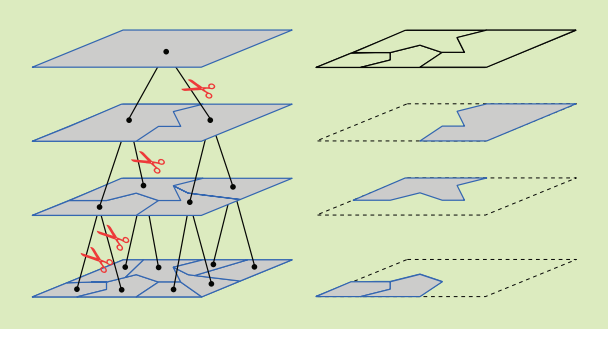


FIGURE 13. A BPT is a hierarchical subdivision of an image. An exhaustive partitioning can be extracted by "cutting" branches at different scales.

In addition to spectral data, the model usually stores the area of the region because it is commonly used in the dissimilarity function. Other shape descriptors such as solidity, rectangularity index, elongatedness, and compactness can also be efficiently stored and computed from the children nodes [118].

THE BPT IS CONSTRUCTED IN A BOTTOM-UP FASHION, BY ITERATIVELY CLUSTERING PAIRS OF SIMILAR REGIONS TOGETHER.

DISSIMILARITY FUNCTION

To establish a priority for merging during BPT construction, it is necessary to provide a means for comparing the models of two regions. A dissimilarity function $O(R_1, R_2)$ typically used for this purpose consists of two factors:

$$O(R_1, R_2) = \min\left(|R_1|, |R_2|\right)^{\beta} D(R_1, R_2), \tag{12}$$

where $|R_i|$ denotes the area of region R_i. The first part of (12), $\min\left(|R_1|, |R_2|\right)^{\beta}$, is the area-weighting factor. This is an agglomerative force intended to cluster regions that are very small compared to the rest of the elements in the RAG. The second factor, $D(R_1, R_2)$, compares both regions based on their spectra. Kullback-Leiber divergence and Bhattacharyya distance are popular choices to compute D [113], [119]. However, using cross-bin measures, which go beyond individual bins, has proven to be more robust [113]. The average of the Earth mover's distances [120] among the histograms of all bands can be used as a robust and efficient cross-bin dissimilarity function.

To better face the internal class variability issue, Maggiori et al. [114] proposed including within the dissimilarity function an additional force that clusters regions belonging to the same class despite being spectrally dissimilar:

$$\begin{aligned} O(R_1, R_2) = \min\left(|R_1|, |R_2|\right)^{\beta} [(1-\alpha) D(R_1, R_2) \\ - \alpha \log P\left(\omega_{R_1} = \omega_{R_2} \,|\, R_1, R_2\right)]. \end{aligned} \tag{13}$$

As in (12), there is an area-weighting factor and an unsupervised term $D(R_1, R_2)$, which is computed by comparing the spectral histograms of regions. Equation (13) adds a supervised term, $P(\omega_{R_1} = \omega_{R_2} \,|\, R_1, R_2)$, which represents the probability of assigning the same label to both regions. In this way, while the unsupervised term penalizes spectral dissimilarity, the supervised term will encourage merging regions that are likely to belong to the same class. The tradeoff between both terms is controlled by parameter α.

The term $P(\omega_{R_1} = \omega_{R_2} \,|\, R_1, R_2)$ is computed by marginalizing over the classes as

$$P\left(\omega_{R_1} = \omega_{R_2} \,|\, R_1, R_2\right) = \sum_{j=1}^{K} P(\omega_j \,|\, R_1) P(\omega_j \,|\, R_2), \tag{14}$$

where $P(\omega_j \,|\, R_k)$ with $k \in \{1, 2\}$ represents the probability of assigning a certain label L_j to segment R_k. To compute $P(\omega_j \,|\, R_k)$, the authors proposed first estimating per-pixel class probabilities, $P(\omega_j \,|\, \mathbf{x}_i)$, $j = 1, \ldots, C$, with an SVM and then averaged these individual probabilities within each region:

$$P(\omega_j \,|\, R_k) = \frac{1}{|R_k|} \sum_{\mathbf{x}_i \in R_k} P(\omega_j \,|\, \mathbf{x}_i). \tag{15}$$

OBJECT-BASED CLASSIFICATION WITH BINARY PARTITION TREES

Let $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^d, i = 1, 2, \ldots, n\}$ be a d-band image seen as a set of n pixel vectors. Object-based classification consists of an exhaustive partitioning of the pixels into a nonoverlapping set of regions $R = (R_j)$ with associated labels $\Omega = (\omega_j)$, where every label ω_j belongs to the set Ω of available information classes. For each class, we suppose that we are given training examples from which we can derive posterior probabilities $P(\omega_j \,|\, \mathbf{x}_i)$ for assigning a certain label ω_j after the spectral observation $\mathbf{x}_i$ is taken into account. Such posterior probability may be derived from an SVM [95]. The negative log-likelihood, $-\log P(\omega_j \,|\, \mathbf{x}_i)$, is typically used to express a cost that penalizes the assignment of label ω_j to pixel $\mathbf{x}_i$.

As proposed in [114], a classification task consists of finding the labeled partitioning (R, Ω) from a BPT that minimizes the energy:

$$E(R, \Omega) = \lambda \| \mathrm{R} \| - \sum_{R_j \in \mathrm{R}} \sum_{\mathbf{x}_i \in R_j} \log P(\omega_j \,|\, \mathbf{x}_i). \tag{16}$$

The first term is a regularizer on the number of regions in the partition $\| \mathrm{R} \|$, which controls the coarseness of the output through parameter λ. The regularization term can be either set manually or directly learned from training samples [118].

From a BPT, the best possible labeled segmentation with respect to (16) can be extracted efficiently by searching for a minimal horizontal s-t cut on the tree with a source at every leaf and a sink at the root [121]. Let us denote $C(R)$ as the energy of the cut on R with minimal (16) among all possible cuts. Considering that the branches in the tree are independent, the globally optimal cut can be found by a dynamic programming algorithm. Let us denote $\mathcal{E}(R) = \min_{\omega \in \Omega} E(\{R\}, \{\omega\})$ as the lowest possible energy of a region R (by assigning the label that incurs the lowest cost). The tree is traversed in a bottom-up manner. Whenever a region R is visited, the following property is evaluated:

$$\mathcal{E}(R) \leqslant C(R_{\text{left}}) + C(R_{\text{right}}), \tag{17}$$

where R_{left} and R_{right} are the children of R. If the property does not stand, we set $C(R) = C(R_{\text{left}}) + C(R_{\text{right}})$ and keep the best cuts of both children. Otherwise, we set $C(R) = \mathcal{E}(R)$ and replace the cuts by R with label L. The traversal finishes when C(root) is computed, i.e., the optimal partition of the whole image.

EXPERIMENTAL RESULTS

EXPERIMENTAL SETUP

In these experiments, SVMs are used to classify the samples. We train multiclass one-versus-one SVMs with Gaussian kernels on the CASI University of Houston, the AVIRIS Indian Pines, and the ROSIS-03 Pavia University data sets, deriving posterior probabilities from each [95]. The SVM training hyperparameters are set using fivefold cross-validation (University of Houston: $c = 10, \gamma = 0.1$; Indian Pines: $c = 1024, \gamma = 2^{-7}$: Pavia University: $c = 128, \gamma = 0.125$).

A BPT is built for each of the data sets. We first set $\alpha = 0$ in (13), thus ignoring the class probabilities during BPT construction [see Figures 10(c), 11(c), and 12(c)]. Alternatively, we set $\alpha = 0.5$, assigning equal importance to the SVM probabilities and the spectral similarity terms [see Figures 10(d), 11(d), and 12(d)]. In this case, the BPT is constructed in a supervised manner. To extract the segmentation, we choose in every case the scale λ in (16) that optimizes the OA. The BPTs are constructed with mild area weighting ($\beta = 0.1$) using a nonparametric model to represent regions, with 30 bins per histogram. The dissimilarity measure used to compare the histograms is based on the Earth mover's distance, as described in the previous section.

RESULTS AND DISCUSSIONS

The numerical results are summarized in Tables 11–13. The unsupervised BPT construction ($\beta = 0$) significantly improves the results over the initial SVM classification in the AVIRIS Indian Pines data set. This data set contains large homogeneous areas with similar spectral characteristics that are grouped together by the BPT, thus enhancing the classification. However, for the much more cluttered scene in the CASI University of Houston data set, the BPT fails at clustering semantically significant objects, downgrading the SVM performance in certain individual classes as well as overall. When the supervised BPT building strategy is used ($\alpha = 0.5$), the BPT clusters significant objects together by combining spectral similarity with class probabilities, outperforming both the SVM and the unsupervised BPT. The results on the ROSIS-03 Pavia University data set confirm the benefits of the supervised BPT construction. In this data set, we observe a consistently good performance of the BPT approach for most classes but a lower performance in the case of meadow and shadow classes (2 and 9, respectively). This is expected because BPTs particularly exploit the notion of objects, while these two classes define vague areas without precise boundaries.

To illustrate the relevance of BPTs for object-based classification, in Figure 14 we show visual close-ups of results on the Pavia Center data set [114] under a similar experimental setup. This data set shows a cluttered urban scene in which single objects are composed of dissimilar parts, challenging the construction of BPTs with a purely spectral dissimilarity criterion. A random color is assigned to each segmented region of the tile building class, as extracted by the BPT-based classification method described in this section. In the unsupervised BPT construction case, while most of the tile surfaces are satisfactorily detected, the objects that compose those regions hardly coincide with real objects. However, the supervised BPT construction better clusters objects into single nodes of the BPT, enabling the extraction of significant objects as entire segments from the tree.

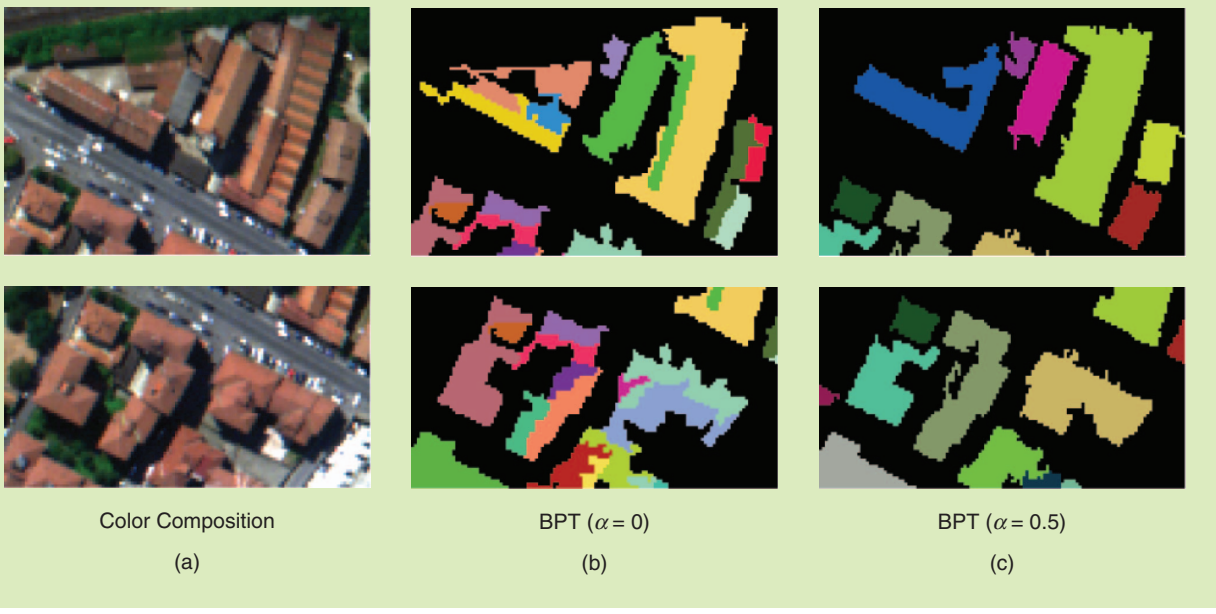


FIGURE 14. A supervised BPT construction ($\alpha = 0.5$) clusters significant objects in single tree nodes. (a) a color composition, (b) a BPT ($\alpha = 0$), and (c) a BPT ($\alpha = 0.5$).

AN OVERVIEW OF SPARSE-REPRESENTATION-BASED CLASSIFIERS

SR has been demonstrated to be a powerful tool for many computer vision problems (e.g., face recognition, image superresolution, and data segmentation) [122], [123]. Recently, SR has also been successfully extended to HSI classification [124]–[126]. In [124], Chen et al. first proposed a pixel-wise sparse classification model based on the observation that the spectral pixels lie approximately in a low-dimensional subspace spanned by dictionary atoms from the same class. Specifically, let $\mathbf{x} \in \mathbb{R}^{d\times 1}$ be one spectral pixel of the HSI, with d denoting the number of spectral bands. A sparse dictionary can be denoted as $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_C] \in \mathbb{R}^{d\times N}$, where $\mathbf{D}_j \in \mathbb{R}^{d\times N_j}$ is the jth class subdictionary whose columns (atoms) are directly drawn or trained from the training pixels, C is the number of classes, N_j is the number of atoms in subdictionary $\mathbf{D}_j$, and $N = \Sigma_{j=1}^{C} N_j$ is the total number of atoms in $\mathbf{D}$.

Given an unknown test pixel $\mathbf{x}^{\text{test}}$, the pixel-wise SR classification (SRC) model obtains its sparse coefficient $\alpha^{\text{test}} \in \mathbb{R}^N$ by solving the following problem:

$$\hat{\alpha}^{\text{test}} = \text{argmin} \| \mathbf{x}^{\text{test}} - \mathbf{D}\alpha^{\text{test}} \|_2 \text{ subject to} \| \alpha^{\text{test}} \|_0 \leqslant S_0, \quad (18)$$

where S_0 is the predefined sparsity level, denoting the number of nonzero coefficients in α^{test}. Equation (18) can be effectively solved by an orthogonal matching pursuit (OMP) [127]. Finally, the class label of test pixel $\mathbf{x}^{\text{test}}$ can be determined by the minimal residual between $\mathbf{x}^{\text{test}}$ and its approximation from each class subdictionary:

$$\text{class}(\mathbf{x}^{\text{test}}) = \underset{j=1,\cdots,C}{\arg\min} \| \mathbf{x}^{\text{test}} - \mathbf{D}_j \alpha_j^{\text{test}} \|_2. \quad (19)$$

Because the pixel-wise sparse model does not consider the spatial information of the HSI, the obtained classification map usually appears to be very noisy. To incorporate the spatial context, [124] proposed a joint sparse model (JSM) to use the spatial information within a fixed-size region for HSI classification. Assuming the region consists of T pixels and all these pixels can constitute a test matrix $\mathbf{X}^{\text{test}}$, while the center pixel is denoted by $\mathbf{x}^{\text{test}}$, a JSM aims to obtain the sparse matrix of $\mathbf{X}^{\text{test}}$ by addressing the following problem:

$$\begin{aligned} &\hat{\mathbf{A}}^{\text{test}} = \text{argmin} \| \mathbf{X}^{\text{test}} - \mathbf{D}\mathbf{A}^{\text{test}} \|_F \\ &\text{subject to} \| \mathbf{A}^{\text{test}} \|_{row,0} \leqslant S_0, \end{aligned} \quad (20)$$

where $\| \mathbf{A}^{\text{test}} \|_{row,0}$ denotes the joint sparse norm, which can select a number of the most representative nonzero rows in $\mathbf{A}^{\text{test}}$. A variant of the OMP algorithm, termed *simultaneous OMP* (*SOMP*), [128] can be used to solve (20). Then, the class label of the pixel centered on the region T is determined by the minimal total residuals between $\mathbf{X}^{\text{test}}$ and the approximations obtained from each class subdictionary:

$$\text{class}(\mathbf{x}^{\text{test}}) = \underset{j=1,\ldots,C}{\text{argmin}} \| \mathbf{X}^{\text{test}} - \mathrm{D}_j \mathbf{A}_j^{\text{test}} \|_F. \quad (21)$$

Compared with the pixel-wise SRC, the JSM can provide better classification performance. However, the pixels within the fixed-size region may be from a different class; thus, the spatial information of the HSI cannot be effectively exploited using this fixed-scale and fixed-size region.

To sufficiently exploit spectral-spatial information of the HSI, some recent works incorporate different kinds of spatial information into the sparse model [129]–[132]. In [129], because regions of different scales contain complementary yet correlated information (as discussed in the previous section), Fang et al. used multiple scale regions for each pixel and proposed a multiscale SR model to adaptively use information among regions of different scales for HSI classification. In [130], Fu et al. selected the neighboring similar pixels to construct shape-adaptive regions and then used the SOMP algorithm to jointly exploit the correlations within the adaptive region for the classification. These multiscale and shape-adaptive sparse (SAS) models can deliver much higher performance than the original JSM but still require high computational cost. This is because, although the spatial correlations among several scales or shape-adaptive regions are effectively utilized, the two sparse models aim to classify only its centered pixel. In [131] and [132], instead of classifying each pixel, the HSI is directly segmented into many superpixels, and a discriminative sparse model is used to classify the whole superpixel, thus greatly enhancing efficiency.

On the other hand, some effective spectral-spatial feature extraction methods have been combined with the sparse model to improve classification performance [133]–[137]. In [133], Song et al. first adopted the morphological APs discussed in the "Mathematical-Morphology-Based Spectral-Spatial Classifiers" section to extract the spatial features and then used the OMP algorithm to classify the extracted features. In [134], Roscher et al. first introduced a shapelet strategy to extract the spectral-spatial features from the local region and then proposed a shapelet-feature-based sparse model for the classification. In [135], Tang et al. transformed the original HSI into the high-dimensional manifold feature space; then, the sparse model was used to effectively reflect the local structures of the HSI, which provided promising classification results. In [136], a series of Gabor wavelet filters with different scales and frequencies was first applied to the original HSI to extract the spectral-spatial features; then, a multitask sparse model was proposed to exploit the correlations among the features for classification. These methods extracted only one kind of feature from the HSI. Because different features can reflect the spectral-spectral information of the HSI from different perspectives, Fang et al. [137] first extracted multiple features (e.g., the Gabor texture, MP, and differential MP) from the HSI and then proposed a multiple-feature-adaptive sparse classification model to exploit the correlations among different features. This approach achieved excellent classification performance.

EXPERIMENTAL RESULTS

EXPERIMENTAL SETUP

In this section, seven well-known SR-based classification methods are used for comparison. We denote the pixel-wise SR method as *SRC* [124]. The fixed-region-based SR method is denoted as the *JSRC* [124]. The region sizes for the JSRC are set to 5 × 5, 3 × 3, and 5 × 5, respectively, for the AVIRIS Indian Pines, ROSIS-03 Pavia University, and CASI University of Houston images. The EMAP + SRC [133] performs as the SRC classifier on the EMAP-extracted features. A multiscale adaptive SR (MASR) [129] uses seven different scales ranging from the 3 × 3 to 15 × 15 [see Figures 10(e), 11(e), and 12(e)]. In the superpixel-based sparse classifier (SBSDM) [131], the superpixel numbers are chosen to be 200, 1,000, 2,000 for the AVIRIS Indian Pines, ROSIS-03 Pavia University, and CASI University of Houston images, respectively. The multiple-features-based sparse classifier (MFASR) [137] is used when four features—the spectral pixel, Gabor texture, MP, and differential MP—are considered.

The sparsity level for these seven classifiers is set to three for the three test images. The classification accuracies for these seven methods on three test images are tabulated in Tables 14–16, respectively.

TABLE 14. THE AVIRIS INDIAN PINES DATA—THE CLASSIFICATION ACCURACY VALUES OBTAINED BY SPARSE REPRESENTATION- AND DEEP-LEARNING-BASED APPROACHES.

CLASSES	UNMIXING + SRC	SRC	JSRC	EMAP + SRC	MASR	SBSDM	SAS	MFASR	CNN	PCA-CNN	EMP-CNN	GABOR-CNN
1	65.03	40.46	83.82	63.95	87.28	80.85	78.68	86.78	79.25	81.96	85.02	84.44
2	71.17	54.46	82.4	88.9	99.11	95.41	94.26	98.98	90.14	90.99	73.45	91.53
3	73.91	55.98	99.46	80.98	99.46	95.11	92.39	98.91	98.77	100	100	98.77
4	92.39	81.66	84.56	74.5	95.53	95.53	94.41	97.54	90.94	91.32	92.80	94.70
5	93.97	78.48	93.69	63.56	99.57	97.7	93.54	97.7	98.85	98.55	98.70	99.28
6	94.99	89.98	96.13	97.49	100	91.12	97.95	99.54	100	96.46	100	100
7	71.90	61.33	89.65	87.15	96.41	97.71	92.92	94.99	95.10	97.55	93.13	95.84
8	62.03	52.77	85.03	73.9	89.21	81.06	87.47	94.17	91.20	89.74	92.25	90.94
9	75.71	47.52	71.28	85.28	95.21	84.22	88.48	92.55	94.34	93.77	94.85	88.59
10	99.38	94.44	100	92.59	100	100	99.38	99.38	100	100	100	100
11	87.70	82.4	98.95	97.99	99.60	99.20	99.36	99.76	95.54	98.25	99.34	99.34
12	74.85	44.55	96.97	96.97	98.48	96.97	90	98.18	89.66	86.21	89.53	89.66
13	100	95.56	97.78	97.78	100	93.33	100	100	100	100	100	100
14	79.49	56.41	100	66.67	100	97.44	100	97.44	100	94.87	100	97.37
15	90.91	90.91	100	100	100	100	100	100	100	100	100	100
16	100	100	80	100	100	100	100	100	100	100	100	100
OA	75.03	61.10	88.24	80.43	94.44	89.90	90.61	95.22	91.53	91.99	92.40	92.84
AA	83.34	70.43	91.23	85.43	93.63	94.10	94.30	97.25	95.24	94.98	94.94	95.65
Kappa	0.7174	0.5618	0.8659	0.7778	0.9749	0.8846	0.8926	0.9452	0.9008	0.9061	0.9105	0.9161

TABLE 15. THE ROSIS-03 PAVIA UNIVERSITY DATA—THE CLASSIFICATION ACCURACY VALUES OBTAINED BY SPARSE REPRESENTATION- AND DEEP-LEARNING-BASED APPROACHES.

CLASSES	UNMIXING + SRC	SRC	JSRC	EMAP + SRC	MASR	SBSDM	SAS	MFASR	CNN	PCA-CNN	EMP-CNN	GABOR-CNN
1	70.29	57.66	49.73	76.41	43.24	26.68	30.89	82.95	88.43	92.23	95.87	87.75
2	71.08	65.23	71.6	66.83	78.02	76.78	72.23	56.79	91.64	97.72	99.50	97.25
3	67.88	61.27	73.33	65.29	80.99	75.04	71.46	91.07	75.95	52.85	61.12	70.92
4	84.82	96.91	96.81	93.03	96.74	95.60	96.33	96.36	96.53	89.46	94.81	97.09
5	99.46	99.82	99.91	96.5	99.91	92.54	90.75	97.75	98.56	99.46	95.15	98.83
6	85.94	66.32	65.05	44.2	78.22	81.12	70.91	81.87	57.87	57.66	64.84	64.62
7	82.77	84.3	95.11	94.9	99.69	99.9	88.99	99.39	80.43	91.42	80.63	76.66
8	71.08	77.11	82.91	73.75	92.45	85.4	90.81	95.87	98.10	98.06	97.26	99.05
9	94.59	57.66	49.73	76.41	56.48	56.6	30.89	92.33	96.84	98.48	96.08	98.36
OA	75.05	69.05	71.78	69.49	75.98	72.00	68.95	74.39	87.01	88.93	91.37	91.62
AA	80.88	76.76	79.30	74.20	68.96	76.63	71.48	88.26	87.15	86.37	87.25	87.83
Kappa	0.6825	0.6077	0.6379	0.6183	0.8064	0.6393	0.6023	0.6828	0.8308	0.8544	0.8867	0.8914

RESULTS AND DISCUSSIONS

From Tables 14–16, the following points can be observed. By using only the spectral information, the SRC generally delivers the worst classification result. By further considering the spatial information within a fixed-size region, the JSRC can achieve a slight improvement on the three test images. In addition, by adjusting the spatial region according to the HSI structures, the MASR, SBSDM, and SAS methods can perform much better than the SRC and JSRC methods, demonstrating the effectiveness of the adopted multiple scales, superpixel, and shape-adaptive pixel strategy. Furthermore, using the information based on multiple features, the MFASR generally achieves the best classification results on the AVIRIS Indian Pines and the CASI University of Houston images. This shows that combining the sparse classifier with multiple features is an effective way to obtain high accuracy. In addition, to analyze whether the unmixing treatment has any effects on the performance, an unmixing-based sparse method (called *unmixing + SRC*) is used, which first employs a well-known unmixing technique [138] to extract the feature of each pixel and then applies SRC to the features for classification. As reported in [139], the dimension for each unmixing feature vector is $2C \times 1$.

INDEED, SPECTRAL INFORMATION IS THE MOST IMPORTANT CHARACTERISTIC AVAILABLE IN HYPERSPECTRAL IMAGING; WITH SUCH RICH SPECTRAL INFORMATION, HYPERSPECTRAL IMAGING CAN BE EFFECTIVE FOR LAND COVER CLASSIFICATION.

Indeed, spectral information is the most important characteristic available in hyperspectral imaging; with such rich spectral information, hyperspectral imaging can be effective for land cover classification. However, due to the external interferences, the spectral vectors from different classes may be mixed with each other; thus, they are hard to distinguish. On the one hand, the unmixing technique is an effective way to reduce the spectral mixture problem for HSI classification. As can be observed in Tables 14 and 15, unmixing + SRC generally outperforms SRC in most classes of the Indian Pines and Pavia University images.

However, setting the number of endmembers in unmixing for different HSIs is a tricky problem. Because the CASI University of Houston data set is very complex, unmixing-based features may not be effectively extracted, and, therefore, the unmixing + SRC method cannot deliver very good performance. Conversely, as can be seen in Tables 14–16, the JSRC, EMAP + SRC, MASR, SBSDM, SAS, and MFASR methods, which jointly use spectral and spatial information, may perform better in terms of OA than SRC, which uses only the spectral information.

The main improvement of the spectral-spatial-based methods over the spectral-based method appears in the classes with large homogeneous spatial regions (e.g., classes 1, 8, and 11 in the Indian Pines data; classes 2 and 8 in the Pavia University data; and classes 9 and 12 in the Houston data). For some classes with detailed structures (e.g., class 16 in the Indian Pines data, class 4 in the Pavia University data, and classes 6, 14, and 15 in Houston data), the spectral-based SRC can perform well and even better than spectral-spatial-based methods.

TABLE 16. THE UNIVERSITY OF HOUSTON DATA—THE CLASSIFICATION ACCURACY VALUES OBTAINED BY SPARSE REPRESENTATION- AND DEEP-LEARNING-BASED APPROACHES.

CLASSES	UNMIXING + SRC	SRC	JSRC	EMAP + SRC	MASR	SBSDM	SAS	MFASR	CNN	PCA-CNN	EMP-CNN	GABOR-CNN
1	75.97	82.72	83.10	77.40	83.10	82.91	83.10	80.82	82.33	80.43	87.49	87.47
2	77.91	82.61	83.36	83.08	79.7	82.99	77.54	82.52	84.30	84.63	80.99	86.01
3	100	99.8	98.42	100	97.43	100	100	100	95.84	87.78	87.72	78.22
4	72.25	92.42	97.06	74.62	96.12	93.56	82.67	82.77	92.60	89.31	90.43	85.02
5	98.20	97.73	99.53	96.02	93.28	100	100	100	99.90	99.14	100	99.89
6	95.80	99.3	97.90	100	90.21	99.30	97.9	99.30	93.00	95.07	97.90	89.44
7	55.69	71.83	73.04	80.6	69.59	66.04	64.93	86.29	80.39	88.62	90.48	90.19
8	38.37	41.22	43.02	29.63	46.06	43.02	45.49	68.66	70.42	79.69	58.51	74.44
9	62.61	61.38	71.01	56.37	76.02	74.13	77.15	78.75	77.77	79.60	79.77	84.42
10	47.59	47.59	47.2	51.64	45.95	41.89	44.79	66.6	56.08	55.16	64.28	63.61
11	74.67	70.87	76.66	63.76	79.98	79.13	82.35	81.5	75.59	73.21	78.37	80.06
12	68.97	55.43	68.78	63.88	76.95	70.51	78.00	74.35	86.55	88.05	78.29	87.30
13	53.33	60.7	43.86	73.33	61.75	41.75	37.54	63.86	84.21	88.12	76.84	85.06
14	100	98.38	96.76	100	100	100	100	100	93.11	100	99.19	100
15	98.52	96.83	100	98.1	100	98.73	99.79	100	88.37	78.14	77.04	56.95
OA	70.49	73.37	76.35	71.44	77.04	75.66	75.72	82.09	82.75	83.22	84.04	84.12
AA	74.66	77.25	78.35	76.56	79.74	78.26	78.08	84.36	84.04	85.61	83.33	82.94
Kappa	0.6802	0.7128	0.7446	0.6906	0.7520	0.7371	0.7376	0.8058	0.8061	0.8165	0.8254	82.51

However, the pixel-wise SRC is an efficient classifier because it needs to classify only one pixel at a time. By using more spatial information to classify the pixel, the JSRC, MASR, SAS, and MFASR methods usually require much higher computational cost. Also, the feature-based classifiers (e.g., EMAP + SRC and MFASR) consume a large amount of the computational cost. By contrast, instead of classifying the HSI in a pixel way, the superpixel-based SBSDM method can classify the whole superpixel (containing multiple spectral pixels) at once and thus greatly enhances classification efficiency.

DEEP-LEARNING-BASED SPECTRAL-SPATIAL CLASSIFIERS

MOTIVATION AND BACKGROUND

Deep learning involves a kind of neural network with two or more hidden layers (the input and output layers are not included). The use of multiple layers tends to extract abstract, invariant, and discriminant features from inputs, which are very useful for the processing steps discussed in this section, including classification, detection, and segmentation [140]. Indeed, considering the task of classification, a linear SVM and logistic regression are believed to have one layer, and a decision tree or SVM with kernels can be considered a two-layer classifier [141]. Compared with traditional classification methods, deep-learning-based classifiers have great potential to obtain high classification performance when facing complex inputs.

As discussed in the "Hyperspectral Imaging Classification" section, hyperspectral sensors obtain spectral and spatial information simultaneously, and the imaging mechanism makes the data inherently complex. Furthermore, due to complex atmospheric conditions, scattering from neighboring objects, and intraclass variability, it is difficult to extract robust, discriminative features of hyperspectral imaging for accurate classification. As an alternative, deep-learning methods can progressively learn invariant and discriminative features. Therefore, it is not surprising that deep learning is widely used for HSI classification.

In deep networks, each layer can extract the features of the previous layer. Under this scheme, high-level features can be learned from low-level ones, while the proper features can be useful for the subsequent classification task. Deep-learning models can potentially lead to abstract and complex features at higher layers, and more abstract features are generally invariant to most local changes of the inputs. With proper training data, advanced learning methods, and powerful computing devices, deep-learning methods can achieve better performance in terms of classification accuracy compared with shallow models.

COMPARED WITH TRADITIONAL CLASSIFICATION METHODS, DEEP-LEARNING-BASED CLASSIFIERS HAVE GREAT POTENTIAL TO OBTAIN HIGH CLASSIFICATION PERFORMANCE WHEN FACING COMPLEX INPUTS.

DEEP-LEARNING-BASED METHODS FOR HYPERSPECTRAL IMAGING CLASSIFICATION

GENERAL FRAMEWORK OF DEEP-LEARNING-BASED METHODS FOR HYPERSPECTRAL IMAGING CLASSIFICATION

Typical deep neural networks stack layer-wise units to formulate the deep models. The layer-wise units have a number of alternatives such as autoencoders (AEs), denoising AEs (DAEs), restricted Boltzmann machines (RBMs), CNNs, and recurrent layers [142]. By using layer-wise units, various deep models can be established. Deep learning involves a number of models including stack AEs (SAEs), DBNs, deep CNNs, and deep recurrent neural networks (RNNs) [142]. All of these deep-learning models have been investigated for HSI classification. Deep-learning-based methods have shown their capabilities with respect to hyperspectral imaging applications [143].

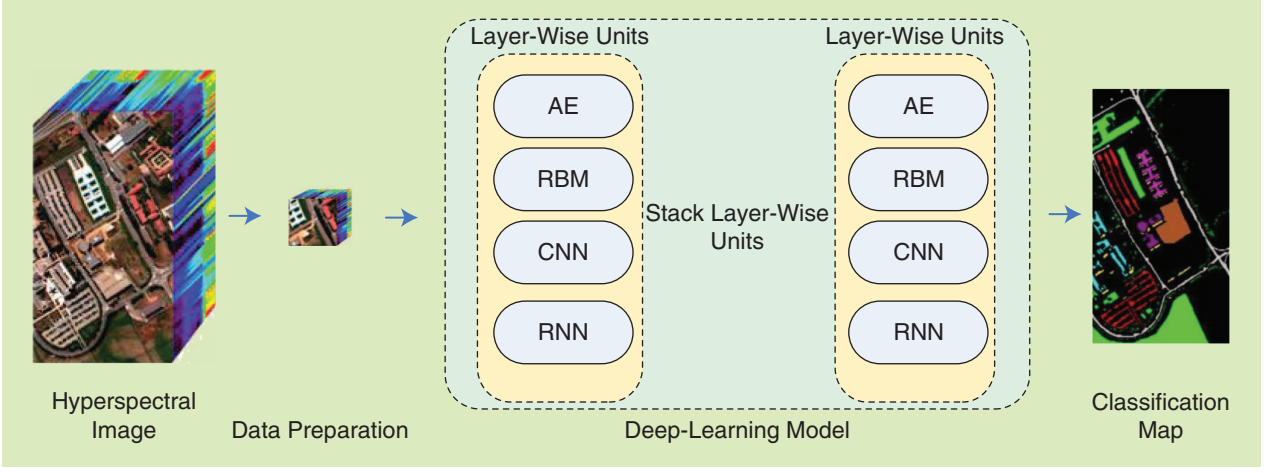


FIGURE 15. The general framework of deep-learning-based methods for HSI classification.

From Figure 15, one can see the general framework of deep-learning-based methods for HSI classification. For spectral-spatial HSI classification, the neighboring pixel vectors of the pixel to be classified are selected to form three-dimensional (3-D) inputs, which are fed to deep models. Deep-learning models hierarchically extract the discriminant features of the inputs and typically use a softmax classifier to obtain the final classification results.

In general, deep-learning models have many parameters (i.e., weights) to be tuned in the training procedure, which means a large number of training samples is needed. Without sufficient training samples, deep models face a problem known as *overfitting*, i.e., the classification performance of test data will be downgraded. This problem becomes serious when fully connected models, including SAEs and DBNs, are used for HSI classification. Due to the shared weights and local connections in CNNs, the number of parameters is dramatically reduced, so CNNs are widely used for HSI classification when only a limited number of training samples is available. In this study, we focus on reviewing CNN-based HSI classification methods because their superior performance over other fully connected networks has already been demonstrated in the literature.

WITHOUT SUFFICIENT TRAINING SAMPLES, DEEP MODELS FACE A PROBLEM KNOWN AS *OVERFITTING*, I.E., THE CLASSIFICATION PERFORMANCE OF TEST DATA WILL BE DOWNGRADED.

THE CORE PARTS AND TECHNIQUES OF DEEP CONVOLUTIONAL NEURAL NETWORKS

A deep CNN usually contains several convolution layers, several nonlinear transformation layers, and several pooling layers [142]. The convolution and nonlinear transformation can be defined as

$$\mathbf{x}_j^l = f\left(\sum_{i=1}^{M} \mathbf{x}_i^{l-1} * \mathbf{k}_{ij}^l + \mathbf{b}_j^l\right), \tag{22}$$

where $f(.)$ is a nonlinear function and * is the convolution operation. The matrix $\mathbf{x}_j^l$ is the jth feature map of the current (l)th layer, and $\mathbf{x}_i^{l-1}$ is the ith feature map of the previous $(l-1)$th layer. M is the number of input feature maps of the current (l)th layer. Furthermore, $\mathbf{k}_{ij}^l$ and $\mathbf{b}_j^l$ are learnable parameters. In the initialization, $\mathbf{k}_{ij}^l$ and $\mathbf{b}_j^l$ are randomly drawn and set to zero, respectively. Then, they are fine-tuned through a back-propagation algorithm.

The rectified linear unit (ReLU) is a relatively new but useful nonlinear operation. It accepts the output of a neuron if it is positive, while it returns 0 if the output is negative. The ReLU operation has advantages such as sparse activation, efficient gradient propagation, and low computation load. Pooling is an operation that combines a small $N \times N$ (e.g., $N = 2$) patch of the previous layer. Pooling usually offers invariance to the deep model by reducing the spatial resolution of the feature maps.

Because of high dimensionality and the limited availability of training samples in HSI classification, deep models are facing the serious problem of overfitting. To address this problem in HSI classification, dropout has been widely used. Furthermore, to achieve better performance in terms of classification accuracy, batch normalization is adopted in a variety of studies to obtain better model generalization.

Dropout is based on setting the output of some hidden neurons to zero, i.e., to drop them. Consequently, the dropped neurons do not contribute in the forward pass and are not used in the back-propagation procedure. Deep CNNs form a different neural network in each training epoch by dropping neurons randomly. The inherent ensemble method efficiently mitigates the overfitting problem in HSI classification [144].

Another useful method for performance improvement is batch normalization. Batch normalization explicitly forces the activations of each layer to have zero means and unit variants. Batch normalization alleviates the problem caused by improper network initialization, and it efficiently speeds up the training procedure by preventing gradient vanishing. Because of these advantages, batch normalization is a practical tool in CNN training [145].

RECENT CONVOLUTIONAL-NEURAL-NETWORK-BASED METHODS FOR HYPERSPECTRAL IMAGING CLASSIFICATION

CNNs can be used as spectral classifiers. To fully use the spatial information provided by HSIs, some spectral-spatial CNN-based methods have been proposed in recent years. The general framework of deep CNN-based methods for HSI classification is shown in Figure 16. Typical methods are presented in the following.

In [146], a classification framework based on principal component analysis (PCA), deep CNN, and logistic regression was proposed. Traditional SAE-based and DBN-based methods usually flattened the spatial map to a 1-D vector, which overlooked the spatial patterns. The CNN-based method in [146] takes the voxels in a neighborhood region into consideration, which allowed obtaining good performance in terms of classification accuracy. Furthermore, the method investigates PCA to reduce the redundancy of spectral information and potentially mitigate the overfitting problem in HSI classification. HSIs are inherently 3-D data, so, in [147], 3-D CNNs are designed to extract the spectral-spatial features of HSIs. Three-dimensional convolution filters reduce the trainable parameters in CNNs, which leads to good classification accuracy.

CNNs can be combined with other techniques to further improve classification performance. In [148], an HSI classification framework is proposed that combines deep CNNs and SR. The features learned from CNNs were refined by SR and then followed by a classifier. In [149], a powerful spatial feature extraction method, i.e., attribute filtering

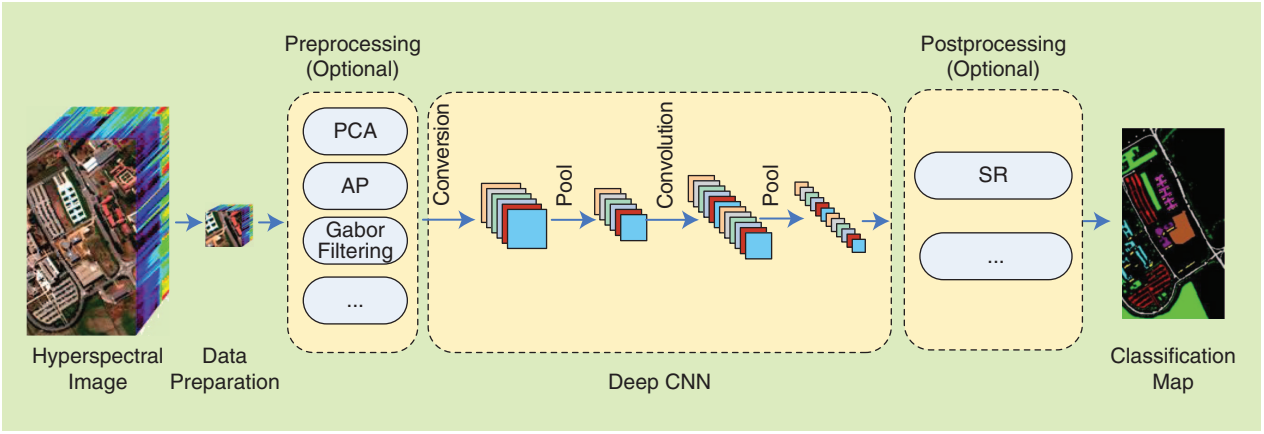


FIGURE 16. The general framework of deep CNN-based methods for HSI classification.

(discussed in the "Mathematical-Morphology-Based Spectral-Spatial Classifiers" section), was combined with deep CNNs. This method led to a better performance compared to individual approaches. Furthermore, in [150], Gabor filtering was used to effectively extract spatial information in HSIs, and then, a CNN was used for further processing. The methods obtained competitive results even when a limited number of training samples was available.

EXPERIMENTAL RESULTS AND DISCUSSION

In this experiment, we considered four different CNN-based methods to provide a comprehensive comparison. For the Indian Pines and Pavia University data sets, we use $27 \times 27 \times 3$ neighbors of each pixel as the input 3-D images in PCA-CNNs, EMP-CNNs, and Gabor-CNNs. Three principal components have been preserved in all of the approaches. For the CNN method, however, all bands are used, so the input 3-D images have the size of $27 \times 27 \times d$. Similarly, we set the input size to be $11 \times 11 \times 3$ and $11 \times 11 \times d$ in these methods with PCA and without PCA, respectively, on the University of Houston data set. In the EMP-CNN method, three principal components from HSIs are computed; then, the opening and closing operations are used to extract spatial information on the first three components. In the experiments, the shape of the SE is set as a disk with an increasing size from 1 to 4. Therefore, 24 spatial features are used for classification. Furthermore, on all the three data sets, the input images are normalized into the range of $[-0.5\ 0.5]$, the size of the minibatch is set to 100, and the number of training epochs for these CNN-based methods is 200.

The classification results of these methods for all three data sets are shown in Tables 14–16. For the PCA-CNN, the CNN is conducted on the three principal components, which is useful when the training samples are limited. From the results, one can see that for all three data sets the Gabor-CNN shows the best performance [see Figures 10(h), 11(h), and 12(h)], followed by the EMP-CNN and PCA-CNN. Additionally, the CNN achieves inferior results compared to the other three deep methods. On the Indian Pines data set, the Gabor-CNN exhibits the highest OA, AA, and Kappa, with an improvement of 1.31%, 0.41%, and 0.0153 over the CNN, respectively. Moreover, it also outperforms the EMP-CNN by 0.25%, 0.58%, and 0.0047 in terms of OA, AA, and Kappa, respectively. The results shown in Tables 14–16 demonstrate similar trends on the other two data sets. For example, the EMP-CNN increases OA, AA, and Kappa by 4.36%, 0.1%, and 0.0559, respectively, compared with the CNN on the Pavia University data set. It also obtains a superior performance compared with the CNN method on the University of Houston data set. The PCA-CNN obtains a classification performance that is higher than the CNN by 0.47%, 1.57%, and 0.0104 in terms of OA, AA, and Kappa, respectively.

BATCH NORMALIZATION ALLEVIATES THE PROBLEM CAUSED BY IMPROPER NETWORK INITIALIZATION, AND IT EFFICIENTLY SPEEDS UP THE TRAINING PROCEDURE BY PREVENTING GRADIENT VANISHING.

CONCLUSIONS AND POSSIBLE FUTURE WORKS

In this article, we have taken a closer look at recent advances in spectral-spatial classification of hyperspectral images. We reviewed five branches of spectral-spatial classification techniques based on mathematical morphology, MRFs, segmentation, SR, and deep learning, both methodologically and through examples of experimental results, to discuss how they address the task of incorporating spatial information into an HSI classification chain. As expected, the results confirm that including spatial information in the classification system can significantly improve classification accuracies compared to cases in which spatial information is discarded; this contributes to the extraction of different objects' shape to address the "salt and pepper" appearance problem that occurs with spectral classifiers. From this perspective, the families of methods discussed in this article benefit from spatial information in different

and complementary ways. Mathematical morphology and deep-learning approaches characterize the desired spatial information at the feature extraction stage through shallow handcrafted features and deep features learned from data. Markovian methods and sparsity-based techniques operate at the classification stage through, respectively, probabilistic spatial-contextual priors and a data-representation viewpoint. Segmentation-based algorithms extract and use information on the regions in the imaged scene.

Consistent with the goal of providing a methodological review and not a comparative experimental study, we did not focus specifically on making model selection or numerical optimization issues homogeneous across the discussed families of methods. Nevertheless, the considered approaches achieved high and comparable overall accuracies in the application to the three considered data sets, which included the very well-known Indian Pines and Pavia University data sets, as well as the recent University of Houston data set. This scenario confirms the effectiveness of current spatial-spectral approaches to the topical problem of HSI classification.

Although the area of spectral-spatial classification of HSIs has been a hot topic in recent years, there are still several aspects worthy of further investigation. Here, we provide pointers to several high-potential areas that may be followed for possible future work.

1) As shown, EPs can provide accurate classification results swiftly in an unsupervised manner. These capabilities encourage one to investigate the performance of this filtering approach for applications related to Earth observation big data processing.
2) An important aspect to investigate in SR for remote-sensing image classification is possible ways to involve spatial information in the model. Additionally, the construction of EPs leads to a very sparse feature space. This encourages one to integrate EPs along with sparse and low-rank techniques to further improve classification accuracies and, at the same time, solve the curse of dimensionality.
3) In the context of segmentation-based methods, possibly one of the most promising directions of work is their combination with classifiers based on CNNs. As discussed in the "Deep-Learning-Based Spectral-Spatial Classifiers" section, CNNs are becoming increasingly popular because of their outstanding recognition capabilities and the automatic learning of hierarchical image features. However, when the goal is to perform a pixel-wise classification, they tend to yield overly unstructured or "blobby" classification maps [151]. The use of a segmentation method coupled with CNNs has proven effective to improve such results [153] and is certainly an interesting topic to be studied in the context of HSI classification.
4) MRFs have proven to be flexible and powerful tools for characterizing contextual information within a Bayesian spectral-spatial classification task. In the case of HSI classification, MRFs have been found especially effective when integrated with kernels, an SVM, and recent energy minimization algorithms. A remarkable property of this integrated framework is that it can be used in conjunction with a wide variety of kernels, MRF models, and energy minimization techniques. In this respect, a promising extension consists of developing advanced hierarchical Markov models [153] that allow the incorporation of multiscale information characterized by segmentation, feature extraction, or CNNs [4], thus possibly leading to morphological, region-based, and deep-learning approaches. A further topical generalization would be to extend the described integrated framework by means of CRFs, which allow additional flexibility in characterizing spatial information and its relationship to the spectral data.
5) Although there are several deep-learning-based spectral-spatial classifiers, deep learning is still in the early stage for HSI classification. Deep learning embraces a wide range of models, many of which have the potential to fulfill the classification task with high accuracy:

- The design of a proper architecture is the core part of a deep model. How to design a proper deep network is still an open area of discussion in the machine-learning and remote-sensing communities.
- A generative adversarial network is an active topic and has already demonstrated its advantages in the remote-sensing community in terms of image translation and data classification [154], [155]. Although the effectiveness of the generative adversarial network has very recently been confirmed for the spectral-spatial classification of an HSI, its concept can be further adapted and modified, making it suitable for large-scale classification problems with a limited number of training samples.
- Deep learning can be combined with other machine-learning or image-processing methods, such as ensemble learning and graph models to achieve better classification performance.

ACKNOWLEDGMENTS

The ROSIS-03 Pavia University and AVIRIS Indian Pines data, and their corresponding reference information, were kindly provided by Prof. Paolo Gamba from the University of Pavia, Italy, and Prof. David Landgrebe from Purdue University, West Lafayette, Indiana, respectively. Additionally, we would like to thank the National Center for Airborne Laser Mapping at the University of Houston, Texas, for providing the CASI Houston data set and the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Technical Committee for organizing the 2013 Data Fusion Contest. The shadow-removed hyperspectral data are provided by Prof. Naoto Yokoya. The work of Pedram Ghamisi is supported by the High Potential Program of Helmholtz-Zentrum Dresden-Rossendorf. Most implementations of the methods described in the article are made available to the research community. The software

and codes for binary partition tree-based classification described in the "Segmentation" section can be found at http://ooclassif.gforge.inria.fr. The max-tree and extinction filter implementation are available at https://github.com/rmsouza01/siamxt. The attribute profile and extinction profile executables can be found on http://pedram-ghamisi.com/index_sub2.html. This distribution is compatible with Linux and Macintosh operating systems. For Windows users, a docker [156] is available at https://hub.docker.com/r/marianapbento/siamxt-1.0/, and the corresponding documentation can be found at http://adessowiki.fee.unicamp.br/adesso/wiki/iamxt/view/. The source codes of the multiscale adaptive sparse representation, space-adaptive sparse models, and multiple-features-based sparse classifier methods can be found at http://www.escience.cn/people/LeyuanFang/index.html. The sets of training and test samples used in this article can be found at https://pghamisi.wixsite.com/mysite.

AUTHOR INFORMATION

Pedram Ghamisi (p.ghamisi@gmail.de) received his B.Sc. degree in civil (survey) engineering from the Tehran South Campus of Azad University, Iran, in 2008; his M.Sc. (first-class honors) degree in remote sensing from the K.N. Toosi University of Technology, Tehran, Iran, in 2012; and his Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, in 2015. Since early 2018, he has worked as the head of the machine-learning group at Helmholtz-Zentrum Dresden-Rossendorf. He received the Best Researcher Award for M.Sc. students from the K.N. Toosi University of Technology during the 2010–2011 academic year. In 2017, he won the Data Fusion Contest 2017 organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. He was the winner of the 2017 Best Reviewer Prize of *IEEE Geoscience and Remote Sensing Letters*. He received the prestigious Alexander von Humboldt Fellowship in 2015. He is an associate editor of *IEEE Geoscience and Remote Sensing Letters*.

Emmanuel Maggiori (emmanuel.maggiori@inria.fr) received his B.Sc. degree in computer science from the National University of Central Buenos Aires (Unicen), Tandil, Argentina, in 2014. That same year, he joined the AYIN and STARS teams at Inria Sophia Antipolis-Mediterranée, Valbonne, France, as a research intern, where he studied the use of hierarchical data structures and optimization algorithms for remote-sensing image analysis. Between 2015 and 2017, he worked on his Ph.D. degree within the TITANE team at Inria, studying machine-learning techniques for the large-scale processing of satellite imagery. He is a Member of the IEEE.

Shutao Li (shutao_li@hnu.edu.cn) received his B.S., M.S., and Ph.D. degrees from Hunan University, Changsha, China, in 1995, 1997, and 2001, respectively. He was a research associate in the Department of Computer Science at the Hong Kong University of Science and Technology, China, in 2011. He joined the College of Electrical and Information Engineering, Hunan University, in 2001, where he is currently a full professor. He has authored or coauthored more than 160 referenced papers. He is currently an associate editor of *IEEE Transactions on Geoscience and Remote Sensing* and *IEEE Transactions on Instrumentation and Measurement*, and he is also a member of the editorial boards of *Information Fusion and Sensing and Imaging*. He was a recipient of two Second Grade National Awards at the Science and Technology Progress of China in 2004 and 2006, respectively. His current research interests include compressive sensing, sparse representation, image processing, and pattern recognition. He is a Senior Member of the IEEE.

Roberto Souza (roberto.medeiros.souza@gmail.com) received his B.Sc. degree in electrical engineering from the Federal University of Par, Brazil, and his M.Sc. and Ph.D. degrees in computer engineering from the University of Campinas, Brazil. He is currently working as a postdoctoral fellow at the Seaman Family MR Centre in Calgary, Alberta, Canada. He has worked as an intern at the Grenoble Institute of Technology, France, and the University of Pennsylvania, Philadelphia. He is the manager and conceiver of the Calgary-Campinas 359 public data set and the open-source max-tree toolbox. He has extensive expertise in image processing, especially techniques based on mathematical morphology and machine learning. His research interests include reproducible research and public data and code. He is a Member of the IEEE.

Yuliya Tarabalka (yuliya.tarabalka@inria.fr) received her B.S. degree in computer science from Ternopil Ivan Pul'uj State Technical University, Ukraine, in 2005 and her M.Sc. degree in signal and image processing from the Grenoble Institute of Technology (INPG), France, in 2007. She received a joint Ph.D. degree in signal and image processing from INPG and in electrical engineering from the University of Iceland, Reykjavik, in 2010. From 2007 to 2008, she was a researcher with the Norwegian Defence Research Establishment, Norway. From 2010 to 2011, she was a postdoctoral research fellow with the Computational and Information Sciences and Technology Office, NASA Goddard Space Flight Center, Greenbelt, Maryland. In 2012, she was a postdoctoral research fellow with the French Space Agency and Inria Sophia Antipolis-Méditerranée, France. She is currently a researcher with the TITANE team of Inria Sophia Antipolis-Méditerranée, Valbonne, France. Her research interests include image processing, pattern recognition, and the development of efficient algorithms. She is a Member of the IEEE.

Gabriele Moser (gabriele.moser@unige.it) received his laurea degree in telecommunications engineering and his Ph.D. degree in space sciences and engineering from the University of Genoa, Italy, in 2001 and 2005, respectively; he has been an associate professor of telecommunications there since 2014. He has been an area editor of *Pattern Recognition Letters* since 2015, and an associate editor of *IEEE Geoscience and Remote Sensing Letters* since 2008. He served

as chair of the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Technical Committee (IADF TC) from 2013 to 2015 and as IADF TC cochair from 2015 to 2017. He received the Best Paper Award at the 2010 IEEE Workshop on Hyperspectral Image and Signal Processing and the Interactive Symposium Paper Award at the IEEE International Geoscience and Remote Sensing Symposium in 2016. He is a Senior Member of the IEEE.

Andrea De Giorgi (andrea.degiorgi@edu.unige.it) received his M.Sc. degree in telecommunications engineering and his Ph.D. degree in telecommunication and electronic engineering from the University of Genoa, Italy, in 2012 and 2018, respectively. Since 2012, he has been with the Methods and Systems for Signal Processing and Recognition group at the Department of Electrical, Electronic, and Telecommunication Engineering and Naval Architecture of the University of Genoa. His research interests span aspects of signal and image processing and pattern recognition for remote sensing and industrial applications. He was a corecipient of the Interactive Symposium Paper Award at the IEEE International Geoscience and Remote Sensing Symposium in 2016. He is a Senior Member of the IEEE.

Leyuan Fang (leyuan_fang@hnu.edu.cn) received his Ph.D. degrees from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2008 and 2015, respectively. Since 2017, he has been an associate professor with the College of Electrical and Information Engineering at Hunan University. From 2011 to 2012, with the support of the China Scholarship Council, he was a visiting Ph.D. student with the Department of Ophthalmology, Duke University, Durham, North Carolina. From 2016 to 2017, he was a postdoctoral research fellow with the Department of Biomedical Engineering at Duke University. In 2011, he won the Scholarship Award for Excellent Doctoral Student granted by Chinese Ministry of Education. His research interests include sparse representation and multiresolution analysis in remote sensing and medical image processing. He is a Senior Member of the IEEE.

Yushi Chen (chenyushi@hit.edu.cn) received his B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, China, in 2001, 2003, and 2008, respectively. Currently, he is an associate professor in the School of Electronics and Information Engineering at the Harbin Institute of Technology. His research interests include remote sensing data processing and machine learning. He is a Member of the IEEE.

Mingmin Chi (mmchi@fudan.edu.cn) received her B.S. and M.S. degrees in electrical engineering from Changchun University of Science and Technology, China, and Xiamen University, China, in 1998 and 2002, respectively. She received her Ph.D. degree in computer science from the University of Trento, Italy, in 2006. Currently, she is an associate professor in the School of Computer Science at Fudan University and group leader at the Shanghai Key Laboratory of Data Science, China. She is a guest editor for the special issue on big data in remote sensing for *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* and for the special issue on analysis of big data in remote sensing for *Remote Sensing*. Her research interests include data science, big data, and machine learning with applications to astronomy, remote sensing, computer vision, natural language processing, and so on. She is a Senior Member of the IEEE.

Sebastiano B. Serpico (sebastiano.serpico@unige.it) received his laurea degree in electronic engineering and his Ph.D. degree in telecommunications from the University of Genoa, Italy. He is a full professor of telecommunications at the Polytechnic School of the University of Genoa, where he teaches courses in the areas of telecommunications, signal processing, pattern recognition, and remote sensing. He heads the research group on signal processing and recognition methods and systems in the Department of Electrical, Electronic, and Telecommunications Engineering, and Naval Architecture at the University of Genoa. He is the chair of the Institute of Advanced Studies in Information and Communication Technologies. He received the Best Paper Award at the 2010 IEEE Workshop on Hyperspectral Image and Signal Processing and the Interactive Symposium Paper Award at the IEEE International Geoscience and Remote Sensing Symposium in 2016. He is an associate editor of *IEEE Transactions on Geoscience and Remote Sensing*. He is a Fellow of the IEEE.

Jón Atli Benediktsson (benedikt@hi.is) received his Cand. Sci. degree in electrical engineering from the University of Iceland, Reykjavik, in 1984 and his M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, Indiana, in 1987 and 1990, respectively. He is the president and rector of the University of Iceland, Reykjavik, where he is also a professor of electrical and computer engineering. He is a cofounder of the biomedical start-up company Oxymap (www.oxymap.com). In 2007, he received the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society (GRSS) and the Outstanding Electrical and Computer Engineering Award from the School of Electrical and Computer Engineering, Purdue University in 2016. He was corecipient of the 2012 IEEE Transactions on Geoscience and Remote Sensing Paper Award, and in 2013 he was corecipient of the GRSS Highest Impact Paper Award. In 2013, he received the IEEE/Association of Chartered Engineers in Iceland (VFI) Electrical Engineer of the Year Award. He is a fellow of the International Society for Optics and Photonics and a Fellow of the IEEE.

REFERENCES

[1] J. A. Benediktsson and P. Ghamisi, *Spectral-Spatial Classification of Hyperspectral Remote Sensing Images*. Norwood, MA: Artech House, 2015.

[2] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

[3] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral-spatial classification of hyperspectral images based on hidden Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2565–2574, 2014.

[4] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial-contextual information in very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013.

[5] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, 2016.

[6] P. Ghamisi, M. Dalla Mura, and J. A. Benediktsson, "A survey on spectral-spatial classification techniques based on attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2335–2353, 2015.

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

[8] Y. Zhong, X. Lin, and L. Zhang, "A support vector conditional random fields classifier with a Mahalanobis distance boundary constraint for high spatial resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. and Remote Sens.*, vol. 7, no. 4, pp. 1314–1330, 2014.

[9] P. Ghamisi, M. S. Couceiro, F. M. Martins, and J. A. Benediktsson, "Multilevel image segmentation approach for remote sensing images based on fractional-order Darwinian particle swarm optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2382–2394, 2014.

[10] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Multiple spectral-spatial classification approach for hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4122–4132, Nov. 2010.

[11] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, 2001.

[12] 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.

[13] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high-resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, 2010.

[14] P. Ghamisi, R. Souza, J. A. Beneiktsson, X. X. Zhu, L. Rittner, and R. Lotufo, "Extinction profiles for the classification of remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5631–5645, 2016.

[15] P. Ghamisi, R. Souza, J. A. Benediktsson, L. Rittner, R. Lotufo, and X. X. Zhu, "Hyperspectral data classification using extended extinction profiles," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1641–1645, 2016.

[16] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. Van Kasteren, W. Liao, R. Bellens, A. Pizurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. and Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, 2014.

[17] IEEE GRSS. (2013). Image analysis and data fusion. [Online]. Available: http://www.grss-ieee.org/community/technical-committees/data-fusion/

[18] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, 2005.

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

[20] P. Salembier, A. Oliveras, and L. Garrido, "Antiextensive connected operators for image and sequence processing," *IEEE Trans. Image Process.*, vol. 7, no. 4, pp. 555–570, 1998.

[21] T. Géraud, E. Carlinet, S. Crozet, and L. Najman, *A Quasi-Linear Algorithm to Compute the Tree of Shapes of nD Images*. Berlin, Germany: Springer-Verlag, 2013, pp. 98–110.

[22] E. Carlinet and T. Géraud, "A comparative review of component tree computation algorithms," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3885–3895, Sept. 2014.

[23] R. Souza, L. Tavares, L. Rittner, and R. Lotufo, "An overview of max-tree principles, algorithms and applications," in *Proc. 29th Conf. Graphics, Patterns and Images (SIBGRAPI)*, 2016, pp. 15–23.

[24] L. Vincent, "Morphological area openings and closings for grey-scale images," in *Shape in Picture*, Y.-L. O, A. Toet, D. Foster, H. Heijmans, and P. Meer, Eds. Berlin, Germany: Springer-Verlag, 1994, pp. 197–208.

[25] C. Vachier, "Extinction value: A new measurement of persistence," in *Proc. IEEE Workshop Nonlinear Signal and Image Processing*, 1995, pp. 254–257.

[26] J. Fabrizio and B. Marcotegui, *Fast Implementation of the Ultimate Opening*. Berlin, Germany: Springer-Verlag, 2009, pp. 272–281.

[27] P. Teeninga, U. Moschini, S. Trager, and M. Wilkinson, "Improved detection of faint extended astronomical objects through statistical attribute filtering," in *Mathematical Morphology and Its Applications to Signal and Image Processing*, J. Benediktsson, J. Chanussot, L. Najman, and H. Talbot, Eds. New York: Springer Int. Publishing, 2015, pp. 157–168.

[28] F. Kiwanuka and M. Wilkinson, "Cluster based vector attribute filtering," in *Mathematical Morphology and Its Applications to Signal and Image Processing*, J. Benediktsson, J. Chanussot, L. Najman, and H. Talbot, Eds. New York: Springer Int. Publishing, 2015, pp. 277–288.

[29] M. Grimaud, "New measure of contrast: The dynamics," in *Proc. SPIE Image Algebra and Morphological Image Processing III*, 1992. doi: 10.1117/12.60650.

[30] G. Bertrand, "On the dynamics," *Image Vision Comput.*, vol. 25, no. 4, pp. 447–454, Apr. 2007.

[31] A. Silva and R. Lotufo, "New extinction values from efficient construction and analysis of extended attribute component tree," in *Proc. XXIst Brazilian Symp. Computer Graphics and Image Processing (SIBGRAPI)*, 2008, pp. 204–211.

[32] E. J. Breen and R. Jones, "Attribute openings, thinnings, and granulometries," *Comput. Vision Image Understanding*, vol. 64, no. 3, pp. 377–389, 1996.

[33] R. Souza, L. Rittner, R. Machado, and R. Lotufo, "A comparison between extinction filters and attribute filters," in *Proc. Int. Symp. Memory Management (ISMM)*, 2015, pp. 63–74.

[34] P. Soille, *Morphological Image Analysis: Principles and Applications*, 2nd ed. New York: Springer-Verlag, 2003.

[35] Y. Xu, T. Géraud, and L. Najman, "Morphological filtering in shape spaces: Applications using tree-based image representations," in *Proc. Int. Conf. Pattern Recognition (ICPR)*, 2012, pp. 485–488.

[36] P. Monasse and F. Guichard, "Fast computation of a contrast-invariant image representation," *IEEE Trans. Image Process.*, vol. 9, no. 5, pp. 860–872, 2000.

[37] Y. Xu, T. Géraud, and L. Najman, "Two applications of shape-based morphology: Blood vessels segmentation and a generalization of constrained connectivity," in *Mathematical Morphology and Its Application to Signal and Image Processing* (Lecture Notes in Computer Science Series), C. L. L. Hendriks, G. Borgefors, R. Strand, Eds. New York: Springer-Verlag, 2013, pp. 390–401.

[38] Y. Xu, E. Carlinet, T. Geraud, and L. Najman, "Hierarchical segmentation using tree-based shape space," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 457–469, Mar. 2017.

[39] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, pp. 55– 63, 1968.

[40] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, "Random forest ensembles and extended multiextinction profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 202–216, Jan. 2018.

[41] P. Du, J. Xia, P. Ghamisi, A. Iwasaki, and J. A. Benediktsson, "Multiple composite kernel learning for hyperspectral image classification," in *Proc. IEEE Int. Geoscience Remote Sensing Symp. (IGARSS)*, 2017, pp. 2223–2226.

[42] L. Fang, N. He, S. Li, P. Ghamisi, and J. A. Benediktsson, "Extinction profiles fusion for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1803–1815, Mar. 2018.

[43] B. Rasti, P. Ghamisi, and R. Gloaguen, "Hyperspectral and lidar fusion using extinction profiles and total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3997–4007, 2017.

[44] M. Zhang, P. Ghamisi, and W. Li, "Classification of hyperspectral and lidar data using extinction profiles with feature fusion," *Remote Sens. Lett.*, vol. 8, no. 10, pp. 957–966, 2017.

[45] P. Ghamisi, B. Hfle, and X. X. Zhu, "Hyperspectral and lidar data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. and Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, 2017.

[46] B. Rasti, P. Ghamisi, J. Plaza, and A. Plaza, "Fusion of hyperspectral and lidar data using sparse and low-rank component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6354–6365, 2017.

[47] R. Souza, L. Rittner, R. Lotufo, and R. Machado, "An array-based node-oriented max-tree representation," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2015, pp. 3620–3624.

[48] 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, 2006.

[49] D. Koller and N. Friedman, *Probabilistic Graphical Models*. Cambridge, MA: MIT Press, 2009.

[50] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[51] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 2000.

[52] S. Nowozin and C. Lampert, "Structured learning and prediction in computer vision," *Found. Trends Comput. Graph. Vision*, vol. 6, no. 3-4, pp. 185–365, 2010.

[53] Z. Kato and J. Zerubia, "Markov random fields in image segmentation," *Found. Trends Signal Process.*, vol. 5, no. 1-2, pp. 1–155, 2011.

[54] S. Li, *Markov random field modeling in image analysis*. London: Springer-Verlag, 2009.

[55] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, 1984.

[56] M. Khodadadzadeh, J. Li, A. Plaza, H. Ghassemian, J. Bioucas-Dias, and X. Li, "Spectral-spatial classification of hyperspectral data using local and global probabilities for mixed pixel characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6298–6314, 2014.

[57] H. Yu, L. Gao, J. Li, S. S. Li, B. Zhang, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification using subspace-based support vector machines and adaptive Markov random fields," *Remote Sens.*, vol. 8, no. 4, pp. 1–21, 2016.

[58] G. Moser and S. Serpico, Combining support vector machines and Markov random fields in an integrated framework for contextual image classification, *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2734–2752, 2013.

[59] S. Sun, P. Zhong, H. Xiao, and R. Wang, "An MRF model-based active learning framework for the spectral-spatial classification of hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1074–1088, 2015.

[60] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 844–856, 2013.

[61] M. Golipour, H. Ghassemian, and F. Mirzapour, "Integrating hierarchical segmentation maps with MRF prior for classification of hyperspectral images in a bayesian framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 805–816, 2016.

[62] N. Bali and A. Mohammad-Djafari, "Bayesian approach with hidden Markov modeling and mean field approximation for hyperspectral data analysis," *IEEE Trans. Image Process.*, vol. 17, no. 2, pp. 217–225, 2008.

[63] X. Cao, L. Xu, D. Meng, Q. Zhao, and Z. Xu, "Integration of 3-dimensional discrete wavelet transform and Markov random field for hyperspectral image classification," *Neurocomputing*, vol. 226, pp. 90–100, Feb. 2017.

[64] P. Chen and J. I. Tourneret, "Toward a sparse Bayesian Markov random field approach to hyperspectral unmixing and

classification," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 426–438, 2017.

[65] J. Xia, J. Chanussot, P. Du, and X. He, "Spectral-spatial classification for hyperspectral data using rotation forests with local feature extraction and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2532–2546, 2015.

[66] B. B. Damodaran, R. R. Nidamanuri, and Y. Tarabalka, "Dynamic ensemble selection approach for hyperspectral image classification with joint spectral and spatial information," *IEEE J. Sel. Topics Appl. Earth Observ. and Remote Sens.*, vol. 8, no. 6, pp. 2405–2417, 2015.

[67] O. Eches, J. A. Benediktsson, N. Dobigeon, and J. Tourneret, "Adaptive Markov random fields for joint unmixing and segmentation of hyperspectral images," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 5–16, 2013.

[68] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, 2011.

[69] J. Yang, Z. Jiang, S. Hao, and H. Zhang, "Higher order support vector random fields for hyperspectral image classification," *ISPRS Int. J. Geo-Inform.*, vol. 7, no. 1, p. 19, 2018.

[70] P. Zhong and R. Wang, "Jointly learning the hybrid CRF and MLR model for simultaneous denoising and classification of hyperspectral imagery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1319–1334, 2014.

[71] F. Li, L. Xu, P. Siva, A. Wong, and D. Clausi, "Hyperspectral image classification with limited labeled training samples using enhanced ensemble learning and conditional random fields," *IEEE J. Sel. Topics Appl. Earth Observ. and Remote Sens.*, vol. 8, no. 6, pp. 2427–2438, 2015.

[72] Y. Zhang, L. Yu, D. Li, and Z. Pan, "Hyperspectral image classification using extreme learning machine and conditional random field," *Adaptation, Learn., Optimization*, vol. 16, pp. 167–178, 2014.

[73] P. Zhong and Z. Gong, "A hybrid DBN and CRF model for spectral-spatial classification of hyperspectral images," *Statist. Optimization Inform. Comput.*, vol. 5, no. 2, pp. 75–98, 2017.

[74] Y. Zhong, J. Zhao, and L. Zhang, "A hybrid object-oriented conditional random field classification framework for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7023–7037, 2014.

[75] J. Zhao, Y. Zhong, T. Jia, X. Wang, Y. Xu, H. Shu, and L. Zhang, "Spectral-spatial classification of hyperspectral imagery with cooperative game," *ISPRS J. Photogrammetry Remote Sens.*, vol. 135, pp. 31–42, Jan. 2018.

[76] Y. Zhong, Q. Cao, J. Zhao, A. Ma, B. Zhao, and L. Zhang, "Optimal decision fusion for urban land-use/land-cover classification based on adaptive differential evolution using hyperspectral and lidar data," *Remote Sens.*, vol. 9, no. 8, 2017.

[77] H. Derin and P. Kelly, "Discrete-index Markov-type random processes," *Proc. IEEE*, vol. 77, no. 10, pp. 1485–1510, 1989.

[78] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, 2004.

[79] 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, 2002.

[80] W. Li, S. Prasad, and J. E. Fowler, "Hyperspectral image classification using Gaussian mixture models and Markov random fields," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 153–157, 2014.

[81] G. Rallier, X. Descombes, F. Falzon, and J. Zerubia, "Texture feature analysis using a Gauss-Markov model in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1543–1551, 2004.

[82] A. Schistad Solberg, T. Taxt, and A. Jain, "A Markov random field model for classification of multisource satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 100–113, 1996.

[83] F. Melgani and S. Serpico, "A Markov random field approach to spatio-temporal contextual image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 11 PART I, pp. 2478–2487, 2003.

[84] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *J. Roy. Statistical Soc. Series B (Methodological)*, vol. 51, no. 2, pp. 271–279, 1989.

[85] 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, 2001.

[86] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, 2004.

[87] A. Ihler, J. Fisher, III, and A. Willsky, "Loopy belief propagation: Convergence and effects of message errors," *J. Mach. Learn. Res.*, vol. 6, pp. 905–936, 2005.

[88] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," in *Proc. IEEE Int. Conf. Computer Vision*, 2003, pp. 900–907.

[89] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[90] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "Map estimation via agreement on trees: Message-passing and linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 11, pp. 3697–3717, 2005.

[91] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, 2006.

[92] J. Besag, "Statistical analysis of dirty pictures," *Appl. Stat.*, vol. 20, no. 5-6, pp. 63–87, 1993.

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

[94] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classiers,* A. J. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000, pp. 61–74.

[95] 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, 2004.

[96] A. De Giorgi, G. Moser, and S. Serpico, "Contextual remote-sensing image classification through support vector machines, Markov random fields and graph cuts," in *Proc. Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2014, pp. 3722–3725.

[97] E. M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton, NJ: Princeton Univ. Press, 2005.

[98] S. 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, 2006.

[99] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelligent Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[100] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, 2008.

[101] P. Gurram and H. Kwon, "Contextual SVM using Hilbert space embedding for hyperspectral classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1031–1035, 2013.

[102] 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, 2010.

[103] B.-C. Kuo and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 5, pp. 1096–1105, 2004.

[104] V. Vapnik and O. Chapelle, "Bounds on error expectation for support vector machines," *Neural Comput.*, vol. 12, no. 9, pp. 2013–2036, 2000.

[105] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, 2014.

[106] R. Gonzalez and R. Woods, *Digital Image Processing*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 2002.

[107] R. L. Kettig and D. A. Landgrebe, "Classification of multispectral image data by extraction and classification of homogeneous objects," *IEEE Trans. Geosci. Electron.*, vol. 14, no. 1, pp. 19–26, 1976.

[108] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson, "Segmentation and classification of hyperspectral images using watershed transformation," *Pattern Recognition*, vol. 43, no. 7, pp. 2367–2379, 2010.

[109] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 9, pp. 2973–2987, 2009.

[110] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, J. Angulo, and M. Fauvel, "Classification of hyperspectral data using support vector machines and adaptive neighborhoods," in *Proc. 6th European Association of Remote Sensing Laboratories Special Interest Group (EARSeL SIG IS) Workshop*, 2009.

[111] 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, 2010.

[112] Y. Tarabalka and A. Rana, "Graph-cut-based model for spectral-spatial classification of hyperspectral images," in *Proc. Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2014, pp. 3418–3421.

[113] S. Valero, P. Salembier, and J. Chanussot, "Hyperspectral image representation and processing with binary partition trees," *IEEE Trans. Image Process*, vol. 22, no. 4, pp. 1430–1443, 2013.

[114] E. Maggiori, Y. Tarabalka, and G. Charpiat, "Improved partition trees for multi-class segmentation of remote sensing images," in *Proc. Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2015, pp. 1016–1019.

[115] E. Maggiori, Y. Tarabalka, and G. Charpiat, "Optimizing partition trees for multi-object segmentation with shape prior," in *Proc. 26th British Mach. Vision Conf. (BMVC)*, 2015.

[116] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 561–576, 2000.

[117] P. Lassalle, J. Inglada, J. Michel, M. Grizonnet, and J. Malik, "A scalable tile-based framework for region-merging segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5473–5485, 2015.

[118] E. Maggiori, Y. Tarabalka, and G. Charpiat, "Optimizing partition trees for multi-object segmentation with shape prior," in *Proc. 26th British Mach. Vision Conf.*, 2015.

[119] F. Calderero and F. Marques, "Region merging techniques using information theory statistical measures," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1567–1586, 2010.

[120] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proc. Int. Conf. Comput. Vision (ICCV)*, 1998, pp. 59–66.

[121] P. Salembier, S. Foucher, and C. López-Martínez, "Low-level processing of PolSAR images with binary partition trees," in *Proc. Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2014.

[122] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[123] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[124] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, 2011.

[125] U. Srinivas, Y. Chen, V. Monga, N. M. Nasrabadi, and T. D. Tran, "Exploiting sparsity in hyperspectral image classification via graphical models," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 505–509, 2013.

[126] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification via kernel sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 217–231, 2013.

[127] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with

applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Systems and Computers*, 1993, pp. 40–44.

[128] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. (2006, Mar.). Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Process*. [Online]. *86(3)*, pp. 572–588. Available: http://www.sciencedirect.com/science/article/pii/S0165168405002227

[129] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, 2014.

[130] W. Fu, S. Li, L. Fang, X. Kang, and J. A. Benediktsson, "Hyperspectral image classification via shape-adaptive joint sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. and Remote Sens.*, vol. 9, no. 2, pp. 556–567, 2016.

[131] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spatial classification of hyperspectral images with a superpixel-based discriminative sparse model," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4186–4201, 2015.

[132] J. Li, H. Zhang, and L. Zhang, "Efficient superpixel-level multitask joint sparse representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5338–5351, 2015.

[133] B. Song, J. Li, M. D. Mura, P. Li, A. Plaza, J. M. Bioucas-Dias, J. A. Benediktsson, and J. Chanussot, "Remotely sensed image classification using sparse representations of morphological attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5122–5136, 2014.

[134] R. Roscher and B. Waske, "Shapelet-based sparse representation for landcover classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1623–1634, 2016.

[135] Y. Y. Tang, H. Yuan, and L. Li, "Manifold-based sparse representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7606–7618, 2014.

[136] S. Jia, J. Hu, Y. Xie, L. Shen, X. Jia, and Q. Li, "Gabor cube selection based multitask joint sparse representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3174–3187, 2016.

[137] L. Fang, C. Wang, S. Li, and J. A. Benediktsson, "Hyperspectral image classification via multiple-feature-based adaptive sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1646–1657, 2017.

[138] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, 2011.

[139] T. Lu, S. Li, L. Fang, X. Jia, and J. A. Benediktsson, "From subpixel to superpixel: A novel fusion framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4398–4411, 2017.

[140] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[141] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[142] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[143] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, 2016.

[144] G. E. Hinton, G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *Comp. Sci.*, vol. 3, no. 4, pp. 212–223, 2012.

[145] S. Ioffe and C. Szegedy. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. ArXiv. [Online]. Available: http://arxiv.org/abs/1502.03167

[146] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectralspatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, 2015.

[147] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, 2016.

[148] H. Liang and Q. Li, "Hyperspectral imagery classification using sparse representations of convolutional neural network features," *Remote Sens.*, vol. 8, no. 2, 2016.

[149] E. Aptoula, M. C. Ozdemir, and B. Yanikoglu, "Deep learning with attribute profiles for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1970–1974, 2016.

[150] Y. Chen, L. Zhu, P. Ghamisi, X. Jia, G. Li, and L. Tang, "Hyperspectral images classification with Gabor filtering and convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2355–2359, 2017.

[151] E. Maggiori, G. Charpiat, Y. Tarabalka, and P. Alliez, "Recurrent neural networks to correct satellite image classification maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 4962–4971, 2017.

[152] N. Audebert, B. Le Saux, and S. Lefevre, "How useful is region-based classification of remote sensing images in a deep learning framework?" in *Proc. Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2016, pp. 5091–5094.

[153] I. Hedhli, G. Moser, S. Serpico, and J. Zerubia, "A new cascade model for the hierarchical joint classification of multitemporal and multiresolution remote sensing data," *IEEE Trans. Geosci. and Remote Sens.*, vol. 54, no. 11, pp. 6333–6348, 2016.

[154] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification,' *IEEE Trans. Geosci. Remote Sens.*, to be published.

[155] P. Ghamisi and N. Yokoya, "Img2dsm: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 794–798, 2018.

[156] D. Merkel. (2014, Mar.). Docker: Lightweight Linux containers for consistent development and deployment. *Linux J.* [Online]. *2014(239)*. Available: http://dl.acm.org/citation.cfm?id=2600239.2600241

GRS