Frontiers in Spectral-Spatial Classification of Hyperspectral Images

Pedram Ghamisi, Emmanuel Maggiori, Shutao Li, Roberto Souza, Yuliya Tarabalka, Gabriele Moser, Andrea de Giorgi, Leyuan Fang, Yushi Chen, Mingmin Chi, et al.

Frontiers in Spectral-Spatial Classification of Hyperspectral Images

Pedram Ghamisi, *Senior Member, IEEE,* Emmanuel Maggiori, *Member, IEEE,* Shutao Li, *Senior Member, IEEE,* Roberto Souza, *Member, IEEE,* Yuliya Tarabalka, *Member, IEEE,* Gabriele Moser, *Senior Member, IEEE,* Andrea De Giorgi, *Student Member, IEEE,* Leyuan Fang, *Senior Member, IEEE* Yushi Chen, *Member, IEEE* Mingmin Chi, *Senior Member, IEEE* Sebastiano B. Serpico, *Fellow, IEEE*, and Jón Atli Benediktsson, *Fellow, IEEE*

***Abstract*—This is a preprint, to read the final version please go to IEEE Geoscience and Remote Sensing Magazine on IEEE XPlore.**

Airborne and spaceborne hyperspectral imaging systems have advanced in recent years in terms of spectral and spatial resolution, which makes data sets produced by them a valuable source for land-cover classification. The availability of hyperspectral data with fine spatial resolution has revolutionized hyperspectral image classification techniques by taking advantage of both spectral and spatial information in a single classification framework. The ECHO (Extraction and Classification of Homogeneous Objects) classifier, which was proposed in 1976, might be the first spectral-spatial classification approach of its kind in the remote sensing community. Since then and especially in the latest years, increasing attention has been dedicated to developing sophisticated spectral-spatial classification methods. There is now a rich literature on this particular topic in the remote sensing community, composing of several fast-growing branches. In this paper, the latest advances in spectral-spatial classification of hyperspectral data are critically reviewed. More than 25 approaches based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning are addressed with an emphasis on discussing their methodological foundations. Examples of experimental results on three benchmark hyperspectral data sets, including both well-known long-used data and a recent data set resulting from an international contest, are also presented. Moreover, the utilized training and test sets for the aforementioned data sets as well as several codes and libraries are also shared online with the community.

***Index Terms*—Hyperspectral Data Classification; Spectral-Spatial techniques, Mathematical Morphology-Based Techniques, Extinction Profiles, Markov Random Fields, Segmentation, Sparse Representation-Based Classifiers, Deep learning.**

The work of Pedram Ghamisi is supported by the "High Potential Program" of Helmholtz-Zentrum Dresden-Rossendorf.

Pedram Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology (HIF), Exploration, D-09599 Freiberg, Germany (e-mail: p.ghamisi@gmail.de).

Emmanuel Maggiori and Yuliya Tarabalka are with INRIA Sophia Antipolis, France (email: emmanuel.Maggiori@inria.fr and yuliya.tarabalka@inria.fr)

Roberto Souza is with the Department of Radiology, University of Calgary, Canada (e-mail: roberto.medeiros.souza@gmail.com).

Shutao Li and Leyuan Fang are with the College of Electrical and Information Engineering, Hunan University (email: shutao_li@hnu.edu.cn and leyuan_fang@hnu.edu.cn).

Gabriele Moser, Andrea De Giorgi, and Sebastiano B. Serpico are with the University of Genoa (email: gabriele.moser@unige.it, andrea.degiorgi@edu.unige.it, sebastiano.serpico@unige.it).

Yushi Chen is with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: chenyushi@hit.edu.cn).

Mingmin Chi is with the school of Computer Science, Fudan University, China.

Jón Atli Benediktsson is with the Faculty of Electrical and Computer Engineering, University of Iceland, 107 Reykjavik, Iceland (e-mail: benedikt@hi.is).

I. Introduction

HYPERSPECTRAL IMAGING SENSORS capture data, usually from the visible through the near-infrared wavelength ranges, consisting of hundreds of (narrow) spectral channels with continuous spectral information, which can accurately discriminate diverse materials of interest on the immediate surface of the Earth. Therefore, hyperspectral images (HSIs) are considered to be 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 the analysis of HSIs, and is utilized to differentiate between diverse land-covers of interest available in the scene [1]. A classification technique assigns unknown pixels to one of the available classes, according to a set of representative samples for each class which are known as *training samples*. Detailed information about advanced supervised classifiers for HSI can be found in [2].

The first attempts dedicated to HSI classification were based on techniques developed for multispectral images which only have a few spectral channels, usually less than thirteen. 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. As a matter of fact, in spite of all similarities between HSIs and other optical images (panchromatic, RGB, and multispectral) the analysis of HSI turns out to be more challenging due to 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 uncertainty sources observed.

Hyperspectral imaging often deals with inherently nonlinear relations between captured spectral information and the corresponding material. This nonlinear relation is the result of a wide variety of reasons such as: (1) Undesired scattering from other objects in the acquisition process, (2) different

atmospheric and geometric distortions, and (3) intraclass variability of similar objects. On the other hand, training samples are usually collected by manual labeling of a small number of pixels in an image or based on some field measurements, which is either expensive or time demanding. 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.

In addition, neighborhood pixels in HSIs are highly correlated since remote sensors acquire 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). This fact has triggered the research area of *spectral-spatial classification* since 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 information can provide useful information about the shape of different structures. Moreover, 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.

In order 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 shape. On the other hand, methodologies based on the adaptive neighborhood system are conceptually more flexible and make use of neighborhoods of variable shape. In this context, 2D convolutional neural networks [5] and the Markov random field family [3, 4, 8] are mostly categorized as spectral-spatial classification approaches using the crisp neighborhood system. In contrast, methodologies based on segmentation [9, 10], morphological profiles [11, 12], attribute profiles [6, 13], and extinction profiles [14, 15] can extract spatial and contextual information using adaptive neighborhood systems.

Fig. 1 demonstrates the dynamic of the important subject of hyperspectral image classification in our community. The number of papers is obtained by checking the keywords of "hyperspectral" and "classification" used in the abstract of published journal and conference papers appeared in IEEE Xplore. To highlight the growth in the number of published papers, the time period has been divided into a few equal time slots [i.e., 1998-2001, 2002-2005, 2006-2009, 2010-2013, 2014-2017 (March 1st)]. As can be seen, the number of papers, which demonstrates the popularity of this subject, has been increasing dramatically.

Due to the fast-growth and importance of HSI classification in the remote sensing community, this paper attempts to critically and systematically review the latest advances in spectral-spatial hyperspectral image classification. The focus is on the methodological foundations of the considered families of techniques and on their mutually complementary methodological rationales, in order to provide the reader with a comprehensive picture on the current evolution of HSI spectral-spatial

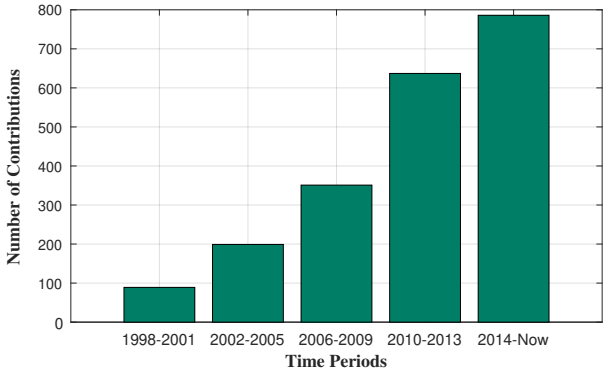


Fig. 1: The number of journal and conference papers available in IEEE Xplore on the subject of hyperspectral image classification within different periods of time. This figure is prepared based on the contributions until March 1, 2017.

classifiers. To this end, computational properties are also recalled and examples of experimental results are discussed for all considered algorithms. Three benchmark data sets, which include both widely known long-used data and a recent data set released within the 2013 IEEE GRSS Data Fusion Contest, are used for this purpose. In this context, we review more than 25 methods categorized into five branches, i.e., mathematical morphology-based techniques, Markov random fields (MRFs), segmentation approaches, sparse representation methods, and deep learning-based classifiers. For each category, the main methodological ideas are recalled and a few key techniques are detailed and exemplified using the aforementioned data sets. Finally, several possible future directions are highlighted. Several codes and libraries as well as the training and test sets used in this paper are shared and made publicly available. It should be noted that this paper dedicates a particular emphasis on methodologies which have been developed since 2013 (after the publication of a previous survey paper on spectral-spatial classification [7]).

It should be noted that HSI classification is the key for a wide variety of real-world applications such as ecological science (e.g., estimating biomass and carbon, studying biodiversity in dense forest zones, and monitoring land-cover changes), geological science (e.g., recovering physico-chemical 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, water quality, monitoring estuarine environments and coastal zones), precision agriculture (e.g., categorizing agricultural classes and extracting nitrogen content for the purpose of precision agriculture), and military applications (e.g., target detection and classification). However, this paper puts emphasis on the methodological aspects of recent publications on spectral-spatial classification.

The rest of the paper is organized as follows. Section II highlights the main notations used in this paper. Section III describes the three studied data sets. Sections IV, V, VI, VII, and VIII are devoted to spectral-spatial classification approaches based on mathematical morphology, MRFs, seg-

mentation, sparse representation, and deep learning, respectively. Section IX wraps up the whole paper and provides potential research directions. Finally, Section X shares the utilized codes, libraries, and training/test samples.

II. Notations

In this paper, matrices are denoted by bold and capital letters. The comma (,) and the semicolon (;) are used for horizontal and vertical concatenation of the elements in a matrix, respectively. $\hat{\mathbf{X}}$ 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 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ 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 $\Omega = \{\omega_1, \omega_2, ..., \omega_C\}$, where C represents the number of classes, using a set of training samples for these classes. Vector $\mathbf{Y} = \{y_1, y_2, ..., y_n\}$ collects the classification labels of all the pixels.

III. Data Sets

Three benchmark data sets have been used to illustrate the considered spectral-spatial methods through examples of experimental results. Two of them are very well-known and have been used for long by the hyperspectral community. The third one is quite recent and was made available to the remote sensing community within 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. The spatial dimensions of this data set are 145 × 145 pixels. The spatial resolution is 20 m. This data set originally includes 220 spectral channels but 20 water absorption bands (104-108, 150-163, 220) have been removed, and the rest (200 bands) has been taken into account for the experiments. The reference data contains 16 classes of interest, which represent mostly different types of crops and are detailed in Table I. Fig. 2 shows a three-band false color image and its corresponding reference samples.

The second data set was captured on the city of Pavia, Italy, by the ROSIS-03 (Reflective Optics Spectrographic Imaging System) airborne instrument. The flight over the city of Pavia, Italy, was operated by the Deutschen Zentrum für Luft- und Raumfahrt (DLR, the German Aerospace Agency) within the context of the HySens project, managed and sponsored by the European Union. The ROSIS-03 sensor has 115 data channels with a spectral coverage ranging from 0.43 to 0.86 μm. Twelve channels have been removed due to noise. The remaining 103 spectral channels are processed. 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:

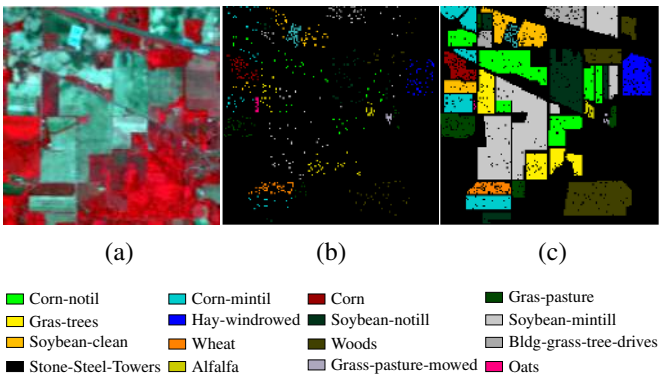


Fig. 2: The AVIRIS Indian Pines hyperspectral data set. (a) Three band false color composite, (b) Reference data and (c) Color code for the classes.

TABLE I: AVIRIS Indian Pines: Number of Training and Test Samples.

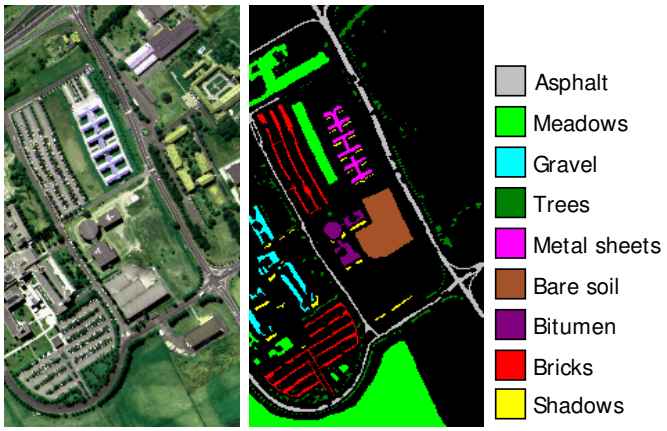
Class		Number of Samples	
No	Name	Training	Test
1	Corn-notill	50	1384
2	Corn-mintill	50	784
3	Corn	50	184
4	Grass-pasture	50	447
5	Grass-trees	50	697
6	Hay-windrowed	50	439
7	Soybean-notill	50	918
8	Soybean-mintill	50	2418
9	Soybean-clean	50	564
10	Wheat	50	162
11	Woods	50	1244
12	Bldg-grass-tree-drives	50	330
13	Stone-Steel-Towers	50	45
14	Alfalfa	50	39
15	Grass-pasture-mowed	50	11
16	Oats	50	5
Total		695	9671

trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow and soil. This data set comprises 640 × 340 pixels. Fig. 3 presents a false color image of the ROSIS-03 Pavia University data and its corresponding reference samples.

The third data set (named "grss_dfc_2013" [17]) was captured by the Compact Airborne Spectrographic Imager (CASI)

TABLE II: ROSIS-03 Pavia University: Number of Training and Test Samples.

Class		Number of Samples	
No	Name	Training	Test
1	Asphalt	548	6304
2	Meadow	540	18146
3	Gravel	392	1815
4	Tree	524	2912
5	Metal Sheet	256	1113
6	Bare Soil	532	4572
7	Bitumen	375	981
8	Brick	514	3364
9	Shadow	231	795
Total		3921	40002



(a) (b) (c)

Fig. 3: The ROSIS-03 Pavia University hyperspectral data set. (a) Three band false color composite, (b) Reference data, and (c) Color code for the classes.

over the University of Houston campus and the neighboring urban area in June, 2012. The size of the data is 349 $\times$ 1905 with the spatial resolution of 2.5 m. This data set is composed of 144 spectral bands ranging 0.38-1.05 μm. This data 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. The "Parking Lot 1" includes parking garages at the ground level and also in elevated areas, while "Parking Lot 2" corresponded to parked vehicles. Table III demonstrates different classes with the corresponding number of training and test samples. Fig. 4 shows a three-band false color image and its corresponding training and test samples.

It is worth noting that, in this paper, we have used a split of the ground truth of each considered data set into the training and the test sets that 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 utilized in this paper can be found at https://pghamisi.wixsite.com/mysite.

TABLE III: CASI Houston University: Number of Training and Test Samples.

Class		Number of Samples	
No	Name	Training	Test
1	Grass Healthy	198	1053
2	Grass Stressed	190	1064
3	Grass Synthetic	192	505
4	Tree	188	1056
5	Soil	186	1056
6	Water	182	143
7	Residential	196	1072
8	Commercial	191	1053
9	Road	193	1059
10	Highway	191	1036
11	Railway	181	1054
12	Parking Lot 1	192	1041
13	Parking Lot 2	184	285
14	Tennis Court	181	247
15	Running Track	187	473
Total		2,832	12,197

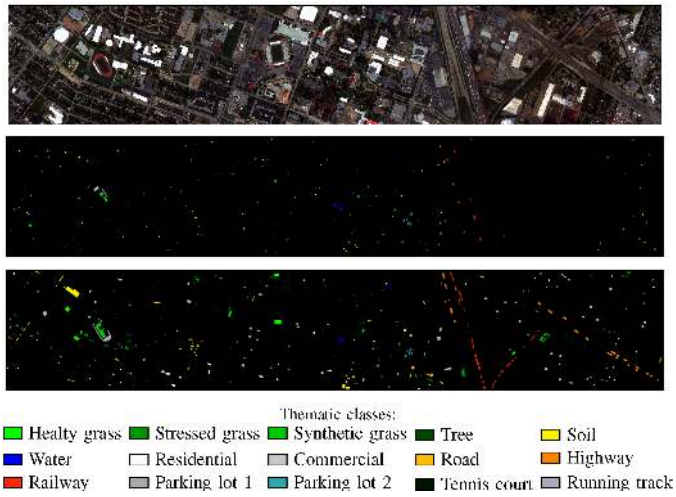


Fig. 4: The CASI Houston University data set - From top to bottom: A color composite representation of the hyperspectral data using bands 70, 50, and 20, as R, G, and B, respectively; Training samples; Test samples; and color code for the classes.

IV. Mathematical Morphology-Based Spectral-Spatial Classifiers

A. Brief Background

The concept of morphological profiles (MPs) was introduced in 2001 [11] and since then, it has been used as a powerful approach to model spatial information (e.g., contextual relations) of the image by extracting structural features (e.g., size, geometry, etc.). MPs are constructed using a successive use of opening/closing operations with a structuring element (SE) of an 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 HSI [i.e., known as extended MPs (EMPs)]. A detailed survey of the MP and its extensions can be found in [1, 7]. Although the MP can improve the discrimination ability of a spectral-spatial classification framework, its concept has a few limitations: (i) the shape of SEs is fixed which make MPs unable to precisely model the shape of different objects and (ii) SEs are only able to extract information w.r.t. the size of existing objects and are unable to characterize information on the gray-level characteristics of the regions.

In order to address the above-mentioned shortcomings of the MP, the morphological attribute profile (AP) was introduced in [13] as a generalization of the MP, which provides a multilevel characterization of an image by using the sequential use of morphological attribute filters (AFs). Compared to MPs, AP is a more flexible tool since it can extract spatial and contextual features based on multiple attributes, which can be purely geometric, or 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 [i.e., known as extended AP (EAP) or extended multi-AP (EMAP) if multiple types

of attributes are taken into account]. A detailed survey about AP and its extensions can be found in [1, 6]. This section takes a closer look to a very recent variant of MPs known as extinction profiles (EPs) [14]. Therefore, in this section, we first briefly discuss the so-called *tree-representation* (max-tree), which is a crucial step for the efficient implementation of the EPs. Then, a brief discussion on attribute and extinction filters is given to highlight the main differences between these two filtering approaches. Furthermore, we discuss extinction profiles and evaluate the performance of different mathematical morphology-based spectral-spatial classifiers through experiments on three widely used hyperspectral data sets.

B. 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 anti-extensive, and extensive by duality[1], connected filters. There are algorithms that allow the max-tree construction in quasi-linear time [21, 22]. The max-tree processing pipeline is depicted by the black path in Fig. 5. The max-tree filtering and image reconstruction processing times are usually negligible compared to the construction time, therefore the max-tree is even more efficient when performing a succession of filtering steps, such as the ones used to construct the extinction profiles [14]. In [23], the principles of the max-tree representation along with the corresponding algorithms and applications were reviewed.

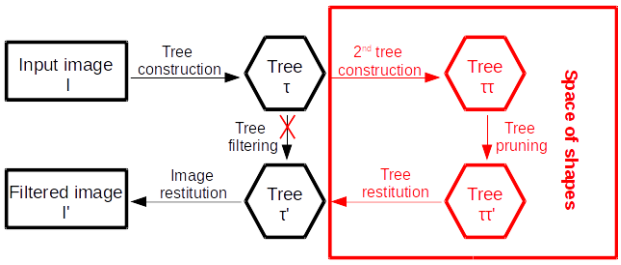


Fig. 5: Max-tree and space of shapes fluxogram.

C. Attribute Filters

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

1) Build the max-tree, if implementing anti-extensive filters, or 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.

Attribute profiles (APs) are constructed by the sequential application of attribute thinning and thickening[2] with a set of progressively stricter threshold values, which were proposed by Dalla Mura et al. [13]. Since then, APs have been investigated intensively for the classification of hyperspectral images. A detailed survey paper on the use of APs for the classification of hyperspectral images can be found in [6].

D. Extinction Values

Extinction values are a measure of persistence of extrema (minima or maxima) proposed by Vachier [25]. The measure of persistence is related to an attribute, which initially (when defined by Vachier) had to be increasing. Extinction values of the height attribute are also known as dynamics [29]. Extinction values can be formally defined. Let M be a regional maximum of a gray scale image $\mathbf{X}$, and $\Psi = (\psi_\lambda)_\lambda$ be a family of decreasing connected anti-extensive 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 still is a regional maxima of $\psi_\lambda(\mathbf{X})$. This definition can be expressed through the following equation:

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

Extinction values of minima can be defined similarly. The height extinction values of maxima of a 1D signal are illustrated in Fig. 6. It is important to emphasize that extinction values are not directly related to the amplitude of the peak, but they also 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].

E. Extinction Filters for Increasing Attributes

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

Natural (real) images are contaminated by noise. Therefore, they contain many irrelevant extrema, *i.e.* extrema with low

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

[2] A filter applied to the min-tree is a thickening operator and a filter applied to the max-tree is a thinning operator.

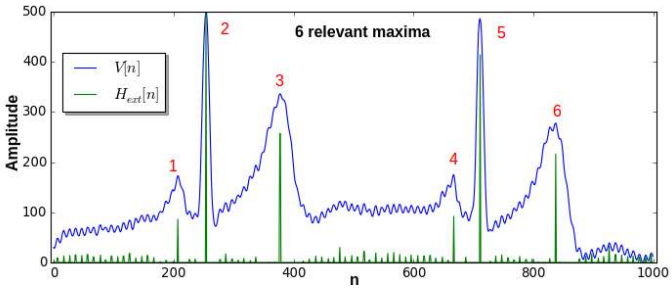
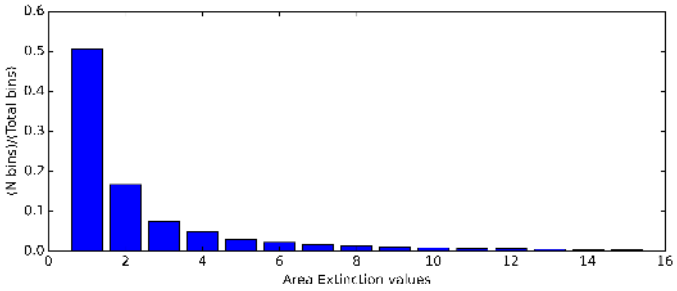


Fig. 6: Height extinction values of maxima of a 1D signal. The six maxima with highest extinction are highlighted.

extinction values. For example, a satellite high resolution panchromatic image of an urban area of the city of Rome, Italy acquired by the QuickBird satellite is depicted in Fig. 7(a). The image is 972×1188 pixels and has 67960 regional maxima. More than 50% of the maxima has an area extinction value of one (Fig. 7(b)), therefore if we apply an area-open [24] filter set to filter structures smaller or equal to one, more than 50% of the image maxima would be filtered.

(a)



(b)

Fig. 7: (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 the following:

1) Build the image max-tree if filtering maxima (anti-extensive) or 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 highest extinction values 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 the following: consider that $Max(\mathbf{X}) = \{\boldsymbol{M_1}, \boldsymbol{M_2}, ..., \boldsymbol{M_N}\}$ denotes the set of regional maxima of the image $\mathbf{X}$. $\boldsymbol{M_i}$ is an image the same size as $\mathbf{X}$ with zero everywhere except in the positions of the pixels that compose the regional maximum $\boldsymbol{M_i}$, where the gray-value is the value of the maximum. Each regional maxima $\boldsymbol{M_i}$ 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 follows:

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

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

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

where $\max$ is the pixel-wise maximum operation. $\boldsymbol{M_1'}$ is the maximum with the highest extinction value, $\boldsymbol{M_2'}$ has the second highest extinction value, and so on.

F. Space of Shapes

Xu et al. [35] proposed to build max-trees of tree-based image representations, i.e. build a max-tree of a max-tree or a max-tree of 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 4-connectivity (vertical and horizontal neighbors of the pixel) or 8-connectivity (all neighbors of the pixel). The second max-tree construction takes us to the space of shapes [35] allowing 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 non-increasing attributes. This methodology was used for blood vessels segmentation, a generalization of constrained connectivity [37], and hierarchical segmentation [38]. The space of shapes fluxogram is depicted in the red path of Fig. 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 Fig. 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 Fig. 9.

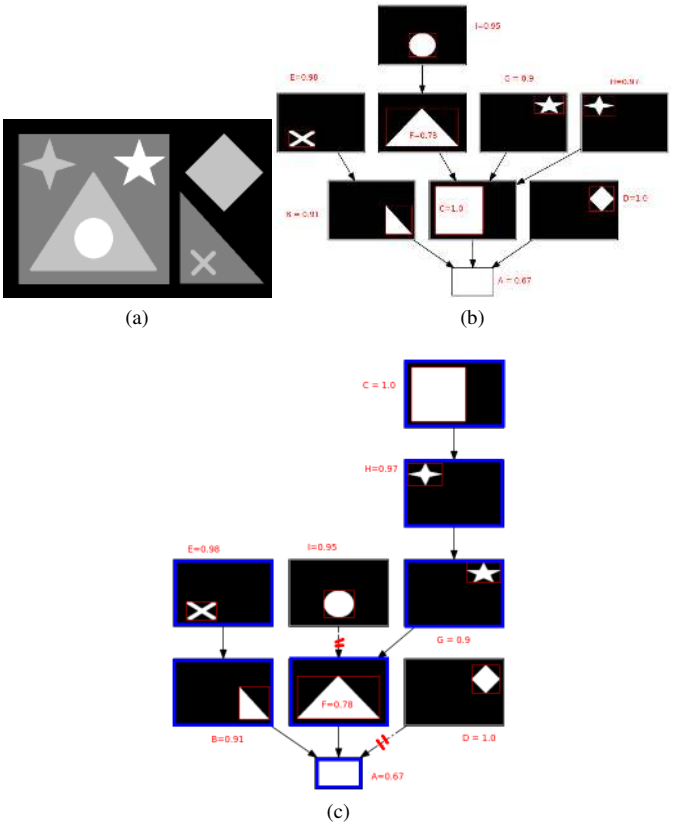


Fig. 8: (a) Synthetic image (b) its max-tree and (c) second max-tree using aspect ratio as the attribute for the second tree construction.

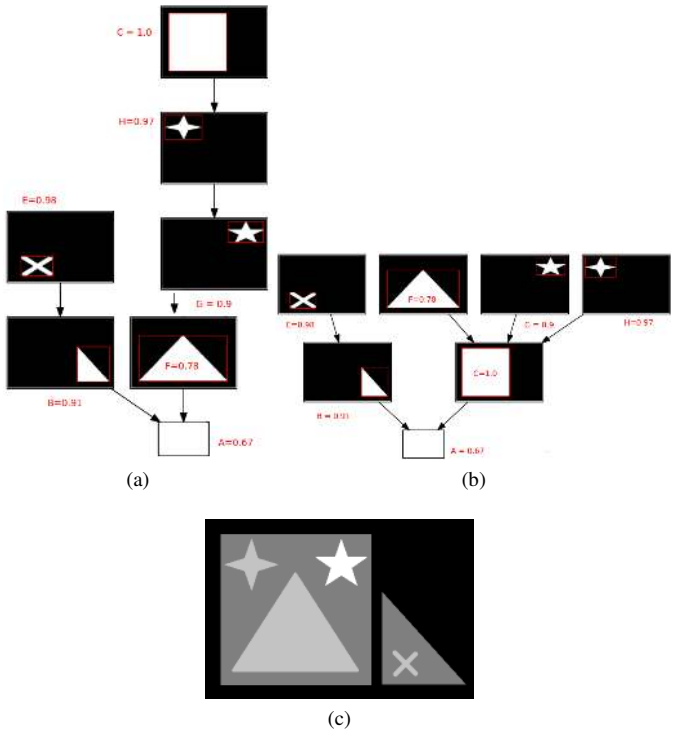


Fig. 9: (a) Second max-tree after filtering step. (b) recovered initial max-tree after filtering step. (c) Resulting image after the filtering procedure.

G. Extinction Filters for Non-increasing Attributes

After building the max-tree of the initial tree representation (max-tree or min-tree in our case), using a non-increasing attribute and, therefore, working on the space of shapes, the height of the attribute used to compute the second max-tree becomes increasing in this space. Therefore, it is possible to compute extinction values and extinction filters for non-increasing attributes. The procedure for computing extinction filters for non-increasing attributes using the max-tree is the following:

1) Build the image max-tree if filtering maxima (anti-extensive) or min-tree if filtering minima (extensive).
2) Compute the second tree (max-tree) of the initial tree representation using the non-increasing attribute chosen.
3) On the second tree, compute the height extinction values for the non-increasing 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.

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

H. Extinction Profiles for Gray-scale Images

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

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

where $\Pi_{\phi^{\lambda}}$ is the thickening extinction profile and $\Pi_{\gamma^{\lambda}}$ is the thinning extinction profile computed with a generic ordered criterion λ (also called threshold or criteria). s is the number of thresholds (i. e., criteria). 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 $\lambda_i \geq \lambda_j$ for thinning.[3] There is a hierarchical relationship between the images generated

[3]Please note that for the EP, the higher value of extrema can provide more detail. This contrasts with the conventional thresholding approach applied on APs, in which 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.

by the EP, i.e., $\Pi_{\boldsymbol{\phi}^{\lambda_1}} \geq \Pi_{\boldsymbol{\phi}^{\lambda_2}} \geq ... \geq \Pi_{\boldsymbol{\phi}^{\lambda_s}} \geq \Pi_{\boldsymbol{\gamma}^{\lambda_s}} \geq \Pi_{\boldsymbol{\gamma}^{\lambda_{s-1}}} \geq ... \geq \Pi_{\boldsymbol{\gamma}^{\lambda_1}}$.

I. Extinction Profiles for Hyperspectral Images

MPs, APs, and EPs, as discussed above, produce several additional features from a single gray scale image (i.e., the input image). It is possible to apply such profiles to all the bands of the hyperspectral data individually and concatenate them. However, this results in producing many redundant features but all the features 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 why the number of bands is first reduced by using a dimensionality reduction approach. Then, a few informative features are fed to MP, AP, or EP to produce spatial and contextual features. In more detail, in order to generalize MPs, APs, and EPs from a gray scale image to HSI, we first need to reduce the dimensionality of the data from $E \subseteq \mathbf{Z}^{d_1}$ to $E' \subseteq \mathbf{Z}^{d_2}$ ($d_2 \leq d_1$) with a generic transformation $\Psi : E \rightarrow E'$ [(e.g., independent component analysis (ICA)]. 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:

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

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

$$\text{EMEP} = \{\text{EEP}_{a_1}, \text{EEP}_{a_2}, ..., \text{EEP}_{a_w}\}, \quad (6)$$

where a_k, $k = \{1, ..., w\}$ denotes different types of attributes. It is easy to understand that due to the fact that different extinction attributes provide complementary spatial and contextual information, the EMEP has a greater capability in extracting spatial information than a single EP [14, 15].

In [40], random forest (RF) ensembles and EMEPs are integrated to shape a spectral-spatial classification framework. In [41, 42], EMEP were used along with composite kernel (CK) learning to perform spectral-spatial classification on HSIs. EMEP was also investigated to fuse spectral and spatial features of hyperpectral and LiDAR data using total variation [43], composite kernel learning [44], deep CNN [45], and sparse and low-rank feature fusion [46].

J. Some Notes on Computational Time

EMEP and EP demand approximately the same computational time since the most time demanding 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 lead to similar processing time. The only difference is that EFs need to compute the extinction values for the attribute, but this can be done simultaneously with the number of nodes, and consequently, it does not add much to the processing time. In terms of non-increasing attributes (*std*), however, EPs need to construct a second max-tree (min-tree), which is not a case for APs. It should be noted that the second max-tree (min-tree) can be constructed much faster since 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]. For detailed analysis of the computational complexity, please see [14].

K. Experimental Results

1) Experimental Setup: For the experiments, RF, with 200 trees, is used to classify input features (see Figs. 14(b), 15(b), and 16(b)). Since EMP only considers attribute area (a), in order to have a fair comparison with EMP, we designed two scenarios: (1) EAP_a and EEP_a which only consider attribute area and (2) EMAP and EMEP which consider five attributes (i.e., area, height, volume, diagonal of the bounding box, and standard deviation) described in [14] and in the previous subsections. EMP is composed of seven opening/closing by reconstruction with a circular SE of size 2, 4, 6, 8, 10, 12, and 14. EMAP is generated using the following attributes and thresholds:

- λ_a = [100, 500, 1000, 2000, 3000, 4000, 5000],
- λ_h = [100, 500, 1000, 2000, 3000, 4000, 5000],
- λ_v = [100, 500, 1000, 2000, 3000, 4000, 5000],
- λ_{bb} = [10, 25, 50, 75, 100, 125, 150], and
- λ_{std} = [10, 20, 30, 40, 50, 60, 70].

For the EMEP (see Fig. 14(f) and Fig. 15(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 EMAP and EMP for each profile.

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 Houston University data, we have also run one extra experiments using a Support Vector Machine (SVM) with CK [48] with weighted summation kernel (see Fig. 16(f)). The weight parameter for both spectral and spatial kernels were simply set to 0.5).

2) Results and Discussions: With reference to Tables IV, V, and VI the following conclusions can be obtained:

- Although EMP, EAP_a, and EEP_a include the same attribute (i.e, area) and number of features, EEP_a leads to 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 EEP_a can improve EMP in terms of classification accuracies is that the shape of the structuring element to produce EMPs is fixed, which imposes a constraint to model spatial structures within a scene.
- As can be seen, one needs to adjust a number of threshold values for EMAP, which is a time consuming procedure.

However, for MPs and EMEPs, one only needs to adjust the number of features.
- As discussed in Section IV-F, the advantage of using AP over EP can be pointed out for non-increasing attributes (standard deviation). In this case, the EP needs to produce the second tree based on the space of the shapes.
- All EMP, EAP_a, and EEP_a can provide results very swiftly.

Table VII schematically compares EMP, EAP, and EEP in terms of classification accuracy, simplicity, and being closer to automatic. The best performance is shown using three bullets while the worst performance is represented by one bullet.

TABLE IV: AVIRIS Indian Pines - Classification accuracies [%] obtained by mathematical morphology-based approaches and the corresponding CPU processing time (in seconds).

Classes	**RF**	**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	100.00	100.00	100.00	100.00	100.00	100.00
14	53.85	94.87	94.87	94.87	94.87	94.87
15	81.82	100.00	100.00	100.00	100.00	100.00
16	100.00	100.00	80.00	100.00	100.00	100.00
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

TABLE V: ROSIS-03 Pavia University - Classification accuracies [%] obtained by mathematical morphology-based approaches and the corresponding CPU processing time (in seconds).

Classes	**RF**	**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	100.00
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
K	0.6498	0.8912	0.8771	0.9332	0.9165	0.9407
Time(s)	8	9	8	8	21	23

V. Markov Random Fields

A. Random fields and probabilistic graphical models

While mathematical morphology captures spatial information within the feature extraction stage of 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 2D 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 dependency properties of the image, usually involving suitable Markovianity conditions [49]. These methodological tools make it possible to capture spatial dependencies in a HSI on a probabilistic basis.

Conventional image classifiers drawn from the pattern recognition literature (e.g., neural networks, RF, or SVM) are usually formalized under the assumption of independent and identically distributed (i.i.d.) pixels [50, 51]. While this non-contextual approach was found effective for remote sensing data at moderate spatial resolutions, it is generally inadequate in the VHR case, including VHR HSI [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 viewpoint, classifiers based on probabilistic graphical models belong to the area of *structured output learning*,

TABLE VI: The CASI Houston University - Classification accuracies [%] obtained by mathematical morphology-based approaches and the corresponding CPU processing time (in seconds).

Classes	**RF**	**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	100.00	100.00	100.00	100.00
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	100.00
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 VII: Performance Evaluation of EMP, EAP, and EEP in Terms of Classification Accuracies, Simplicity, and Being Closer to Automatic. The best performance is shown using three bullets while the worst performance is represented by one bullet.

Techniques	Accuracy	Automation	Speed
EMP	●●	●●	●●●
EAP	●●	●	●●●
EEP	●●●	●●●	●●●

which includes algorithms whose output is supposed to exhibit dependency structures [52].

The main family of probabilistic graphical models that have been extensively applied to HSI classification is given by *Markov random fields (MRF)*, which provide powerful and flexible spatial-contextual models for the prior distribution in Bayesian image analysis [53–55]. They have been recently used for HSI classification in conjunction with SVM [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 (CRF), which model as Markovian the posterior distribution directly [68]. HSI image classification methods have recently been developed using CRFs along with SVM and Mahalanobis distances [8, 69], MLR [70], decision tree ensembles [71], extreme learning machines [72], deep belief networks [73], segmentation and object-based image analysis [74], game theory [75], and adaptive differential evolution for decision fusion with LiDAR data [76]. Here, we shall focus on MRFs, first reviewing the basics and then discussing advanced methods that integrate the MRF and SVM approaches to HSI classification.

B. Key ideas of MRF modeling

MRF models formalize spatial interactions on a local basis using neighborhoods. A *neighborhood system* is defined on the 2D 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 a pixel is neighbor to another, then the vice versa holds as well) and irreflexive (i.e., no pixel is 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 Markovianity property holds ($i = 1, 2, \ldots, n$) [53, 54]:

$$P(y_i|y_j, j \neq i) = P(y_i|y_j, j \in \partial i). \tag{7}$$

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 to only conditioning to the labels of the neighbors – a condition that extends to 2D images the analogous properties of 1D Markov chains [77].

In a 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 Section II), it is possible to prove through the Hammersley-Clifford theorem that, under mild assumptions, the joint posterior distribution $P(\mathbf{Y}|\mathbf{X})$ of all the labels given all image data is a *Gibbs distribution* and is proportional to $\exp[-U(\mathbf{Y}|\mathbf{X})]$, 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 (namely, the MRFs with "only nonzero pairwise clique potential"), the functional form of the energy is [54, 78]:

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

where $D_i(\mathbf{x}_i, y_i)$ is a pixelwise (or *unary*) term associated with the spectral feature vector $\mathbf{x}_i$ and the label of the ith pixel, $V_{ij}(y_i, y_j)$, 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 $U(\mathbf{Y}|\mathbf{X})$ with respect to $\mathbf{Y}$, given the input HSI $\mathbf{X}$. Within this minimization, the pixelwise 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 pixelwise negative class-conditional log-likelihood of the spectral data, estimated through parametric [3, 64, 79, 80] or non-parametric 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, stationary or non-stationary behaviors [53, 54, 65, 80]. More advanced MRFs also allow multiscale, multiresolution, multisensor, and multi-temporal 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 and belief propagation concepts and have attracted increasing interest during the last decade. In the case of binary classification, *graph cuts* make use of a reformulation based on the min-flow/max-cut theorem to reach, with low-order polynomial complexity, a global energy minimum, provided the pairwise potential satisfies a suitable condition [84]. In the multiclass case, graph cut algorithms iteratively define a sequence of suitable binary problems, and under appropriate assumptions on the pairwise potential, converge to local minima with 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 re-weighted message passing (TRW)* 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 numbers of edges in the graph, of classes, and of iterations [91].

Among earlier methods, which have been consolidated since the Eighties, 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 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 HSI can be found, e.g., in [58, 60, 64–66]. For more details on MRF and energy minimization, we refer the reader to [53, 54].

C. Bringing together SVM, MRF, and energy minimization

Among the non-contextual classifiers, SVMs are known for their remarkable generalization capability even in the application to high-dimensional feature spaces – a property that justifies their consolidated use for spectral HSI classification. Hence, the opportunity to combine SVM with contextual MRF models nicely fits the requirements of spectral-spatial HSI classification and has received substantial attention lately [57, 58, 61, 63]. Here, we shall not review the basics of SVM, for which we refer the reader to well-known textbooks such as [51, 93], and we 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, 95], which use parametric modeling, maximum-likelihood, and numerical analysis concepts to approximate pixelwise posteriors. The resulting probabilistic output is plugged into the unary energy. This approach is computationally efficient and has recently led to accurate results with HSI (e.g., [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, 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 SVM and MRF into a unique framework, in which 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 $y_i = -1$ and that associated with $y_i = 1$ can be expressed as an SVM-like kernel expansion ($i = 1, 2, \ldots, n$):

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

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]:

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

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 $\sum_{j \in \partial i} [V_{ij}(-1, y_j) + V_{ij}(1, y_j)]$ is a constant independent on the labels of the neighbors, then the energy $U(\mathbf{Y}|\mathbf{X})$ can be written as $-\sum_{i=1}^{n} y_i \Delta U_i$, or equivalently, in terms of the following unary and pairwise terms ($i = 1, 2, \ldots n; j \in \partial i$) [96]:

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

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, 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 the Markovian kernel merges 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, 96].

Given the integration of 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 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 graph cuts [96], LBP, and TRW. The resulting classifiers iteratively alternate: (i) the update of the additional spatial feature as a function of the current classification map; (ii) the training of

an SVM with the Markovian kernel; and (iii) 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.

The resulting classification methods will be collectively named *Markovian support vector classifiers (MSVC)* in the following. 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].

D. Experimental results

1) Experimental setup: The MSVC framework is experimented with the considered data sets in conjunction with three energy minimization algorithms, i.e., graph cuts, LBP, and TRW. Multiclass labeling is accomplished using the one-vs-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 graph cut (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 [100], 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: (i) the MRF- and kernel-based method in which approximate pixelwise 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 the following); (ii) the MRF- but not kernel-based classifier in which unaries are computed through a Gaussian class-conditional model (MRF-Gauss in the following); and (iii) the kernel- but not MRF-based Contextual SVM (CSVM) technique in [101]. CSVM incorporates spatial information into an SVM for HSI classification using suitable embeddings in a reproducing kernel Hilbert space. A partly similar analytical formulation can be achieved using graph-kernel concepts [102]. Alternately, as discussed in Section IV, if spatial information is characterized in the feature extraction rather than the classification stage, composite kernels can be used to fuse spectral and spatial features [41, 42, 44, 48].

Before applying (ii), 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., $V_{ij}(y_i, y_j) = -1$ if $y_i = y_j$ and $V_{ij}(y_i, y_j) = 0$ otherwise. We recall that, with this choice, both conditions for the applicability of (11) and for the convergence of graph cuts 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.

TABLE VIII: AVIRIS Indian Pines - 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**			**MRF-SVM-Post**	**MRF-Gauss**	**CSVM**
	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

2) Results and discussion: The accuracies obtained by the aforementioned methods on the test samples of the three data sets are collected in Tables VIII-X. For the state-of-the-art MRF-SVM-Post and MRF-Gauss methods, for brevity only the results of the energy minimization algorithm that provides the highest OA are shown in these tables.

The MSVC framework obtains values of OA around 91-92%, 82-87%, and 85-87% in the cases of AVIRIS Indian Pines, ROSIS-03 Pavia University, and CASI Houston University, 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 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 aforementioned nonparametric kernel methods.

On one hand, all the considered Markovian approaches yield improvements over purely spectral classifiers (e.g., see RF and SVM in Tables IV-VI and XI-XIII), an expected conclusion that has been largely demonstrated in the literature (e.g., [7, 105, 106]). 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, from the flexible nonparametric formulation of kernel learning, and from its effectiveness in high-dimensional feature spaces. This comment is also confirmed by previous experiments, which

TABLE IX: ROSIS-03 Pavia University - 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			MRF-SVM-Post	MRF-Gauss	CSVM
	GC	TRW-S	LBP			
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 X: CASI Houston University - 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			MRF-SVM-Post	MRF-Gauss	CSVM
	GC	TRW-S	LBP			
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

indicated 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 (Figs. 14(g), 15(g), and 16(g)).

The three considered energy minimization algorithms overall exhibit similar behaviors. They obtain very similar accuracies in the cases of the AVIRIS Indian Pines and CASI Houston University data sets, while in the case of ROSIS-03 Pavia, TRW-S reaches 3-4% higher OA than graph cuts and LBP. On one hand, the high accuracies achieved confirm the effectiveness of current advanced graph cut and message passing techniques for MRF energy minimization in a HSI classification task – a conclusion that has been drawn in numerous image processing and computer vision applications [100]. On the other hand, the performances obtained using all three methods also confirm the flexibility of the MSVC framework in incorporating arbitrary energy minimization algorithms. This flexibility also comes together with the opportunity to fully automate the resulting classifiers through the aforementioned parameter optimization methods in [58, 98]. We also recall that a previous MSVC formulation using ICM was originally developed in [58] and experimentally validated with various data modalities, including HSI.

VI. 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 non-overlapping homogeneous regions with respect to some criterion of interest or homogeneity criterion (e.g., based on the intensity or on the texture) [107]. Hence, each region in the segmentation map can be seen as a connected spatial neighborhood for all the pixels within this region. One of the pioneering spectral-spatial techniques belongs to this category: the well-known ECHO (Extraction and Classification of Homogeneous Objects) classifier [108], which has been extensively used by the remote sensing community. It is based on region growing to find homogeneous groups of adjacent pixels, which 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) [109–111]. From a segmentation map, any pixelwise classifier and majority voting can be applied to combine spectral and spatial information: for every region in the segmentation map, all the pixels are assigned to the most frequent class within this region, based on pixelwise classification results [111].

It is however a challenging task to perform HSI segmentation automatically. The performance is highly dependent both on the measure of region homogeneity and on the algorithm parameters. Several alternatives have been proposed to deal with this challenge. Tarabalka et al. [10, 112] proposed to perform a marker-controlled segmentation for this purpose. The classification probabilities are used to automatically select the most reliably classified pixels (i.e., pixels belonging with the high probability to the assigned class). The classification map is then obtained by building a minimum spanning forest from the image graph rooted on the selected markers. This method has a similar principle as the widely used MRF-based graph cut approach [113] 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 in building first a hierarchy of segmentations at different 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 to use a binary partition tree (BPT) model for this purpose [114]. In this method, a BPT is first constructed by iteratively clustering similar regions based on a criterion specifically designed for hyperspectral images. Each BPT node is then modeled by its mean spectrum and classified by using an SVM. A *so-called* misclassification rate is computed for each node, which can be understood as the error incurred by assigning the entire node to

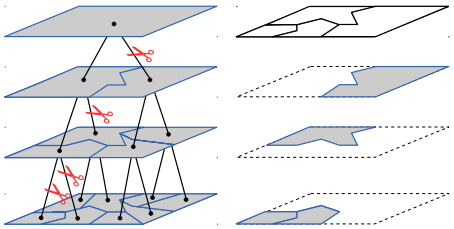


Fig. 10: A binary partition tree (BPT) is a hierarchical subdivision of an image. An exhaustive partitioning can be extracted by "cutting" branches at different scales.

the wrong class. A spectral-spatial classification map is finally built in a bottom-up traversal of the tree by extracting regions with a low misclassification rate. Another BPT-based model has been recently proposed in [115] and extended in [116], where the object-based classification problem is formulated as an energy minimization task. While the graph-cut-based approach has been mentioned in the previous section, we detail in the following the BPT-based segmentation method and demonstrate its performances for the hyperspectral data sets.

A. Binary Partition Tree Model

BPTs were studied by Salembier and Garrido [117] 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 non-overlapping regions, and so on. The construction of a BPT is done 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, though nothing prevents the approach to be used with other inputs too. Every edge in the RAG is labeled with a *dissimilarity* value that compares the two associated regions. BPTs are typically constructed by using a global mutual best fitting region merging approach [118]: at each iteration, the two most similar regions in the current subdivision are merged together. When a merge occurs, a new region is added to the BPT, connected to its two corresponding children. The process finishes 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 Fig. 10). In this procedure, commonly referred to as *pruning*, branches can be selected at different scales, an inherent advantage of such hierarchical structure.

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

Region model. There are essentially two alternatives to represent the spectrum of each region: parametric and non-parametric models. Non-parametric models (*e.g.*, per-band histograms of the pixel values) have proven to be a better approach than the parametric counterpart (*e.g.*, average spectrum), since they represent the real observed distributions and can thus describe the internal variability of a region [114].

In addition to spectral data, the model usually stores the *area* of the region, since 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 [119].

Dissimilarity function. To establish a priority for merging during BPT construction, it is required to provide a means to compare models of two regions. A dissimilarity function $O(R_1, R_2)$ typically used for this purpose comprises two factors as follows:

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

where $|R_i|$ denotes the area of region R_i. The first part of (12), $\min(|R_1|, |R_2|)^{\beta}$, is the so-called *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 [114, 120]. However, using cross-bin measures, which go beyond individual bins, has proven to be more robust [114]. The average of Earth Mover's Distances [121] among 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.* [115] proposed to include within the dissimilarity function an additional force that clusters regions belonging to the same class, despite being spectrally dissimilar:

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

As in (12), there is an area-weighting factor and an unsupervised term $D(R_1, R_2)$, which is computed by comparing spectral histograms of regions. Equation (13) adds a *supervised* term $P(\omega_{R_1} = \omega_{R_2} | R_1, R_2)$, the probability of assigning the same label to both regions. 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 trade-off 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 follows:

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

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 to estimate first per-pixel class probabilities $P(\omega_j | \mathbf{x}_i), j = 1, ..., C$ with an SVM, and then average these individual probabilities within each region:

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

B. Object-Based Classification with Binary Partition Trees

Let $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^d, i = 1, 2, ..., n\}$ be a d-band image seen as a set of n pixel vectors. Object-based classification consists in an exhaustive partitioning of the pixels into a non-overlapping set of regions $\mathrm{R} = (R_j)$, 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 we are given training examples from which we can derive posterior probabilities $P(\omega_j|\mathbf{x}_i)$ of 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 [115], a classification task consists in finding the labeled partitioning (R, Ω) from a BPT that minimizes the energy:

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

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 [119].

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 [122]. Let us denote $C(R)$ 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 $\mathcal{E}(R) = \min_{\omega \in \Omega} E(\{R\}, \{\omega\})$ 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:

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

where R_{left} and R_{right} are the children of R. If the property does not stand, we set $C(R) = C(R_{left}) + C(R_{right})$ 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.

C. Experimental Results

1) Experimental Setup: In these experiments, SVMs are used to classify the samples. We train multi-class one-vs-one SVMs with Gaussian kernels on the CASI Houston University, the AVIRIS Indian Pines, and the ROSIS-03 Pavia University data sets, and derive posterior probabilities from them [95]. The SVM training hyperparameters are set by using 5-fold cross-validation (Houston University: $c = 10, \gamma = 0.1$; 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 Figs. 14(c), 15(c), and 16(c)). Alternatively, we set $\alpha = 0.5$, assigning equal importance to the SVM probabilities and the spectral similarity terms (see Figs. 14(d), 15(d), and 16(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 overall accuracy. The BPTs are constructed with mild area weighting ($\beta = 0.1$) and using a non-parametric 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.

2) Results and Discussions: The numerical results are summarized in Tables XI, XII, and XIII. The unsupervised BPT construction ($\alpha = 0$) 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, which are grouped together by the BPT, enhancing the classification. However, for the much more cluttered scene in the CASI Houston University data set, the BPT fails at clustering semantically significant objects, downgrading the SVM performance in certain individual classes and 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, since BPTs particularly exploit the notion of objects, while these two classes define vague areas without precise boundaries.

TABLE XI: AVIRIS Indian Pines - 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	100.00
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	100.00
14	56.41	97.44	94.87
15	81.82	90.91	100.00
16	100.00	0.0	100.00
OA	65.64	82.46	84.36
AA	76.56	83.79	91.61
Kappa	0.6141	0.8013	0.8224

To illustrate the relevance of BPTs for object-based classification, in Fig. 11 we show visual close-ups of results on the Pavia Center data set [115], under a similar experimental setup. This data set shows a cluttered urban scene, where

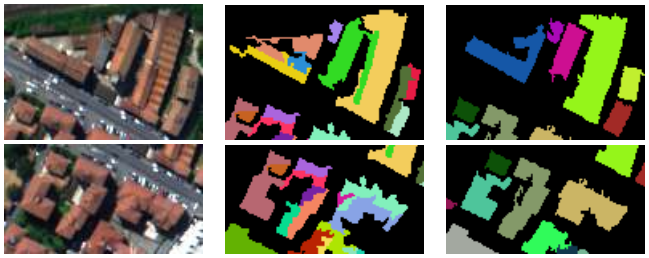
TABLE XII: ROSIS-03 Pavia University - Classification accuracy values obtained by binary partition trees.

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	100.00	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

TABLE XIII: Houston University - 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	100.00	100.00
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	100.00	100.00	100.00
15	98.10	100.00	100.00
OA	81.91	79.69	83.78
AA	83.78	82.44	85.34
Kappa	0.8040	0.7799	0.8242

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.



Color composition BPT ($\alpha = 0$) BPT ($\alpha = 0.5$)

Fig. 11: Supervised BPT construction ($\alpha = 0.5$) clusters significant objects in single tree nodes.

VII. Sparse Representation

A. An Overview of the Sparse Representation-Based Classifiers

Sparse representation (SR) has been demonstrated to be a powerful tool for many computer vision problems (e.g., face recognition, image super-resolution, and data segmentation) [123, 124]. Recently, the SR has also been successfully extended to the hyperspectral image classification [125–127]. In [125], Chen *et al.* firstly proposed a pixel-wise sparse classification model, which is based on the observation that the spectral pixels approximately lie 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 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 j-th 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 $N = \sum_{j=1}^{C} N_j$ is the total number of atoms in $\mathbf{D}$. Given an unknown test pixel $\mathbf{x}^{\text{test}}$, the pixel-wise SR classification model obtains its sparse coefficient $\alpha^{\text{test}} \in \mathbb{R}^N$ by solving the following problem:

$$\hat{\alpha}^{\text{test}} = \operatorname{argmin} \left\|\mathbf{x}^{\text{test}} - \mathbf{D}\alpha^{\text{test}}\right\|_2 \text{ subject to } \left\|\alpha^{\text{test}}\right\|_0 \leqslant S_0 \tag{18}$$

where S_0 is the predefined sparsity level, denoting the number of nonzero coefficients in α^{test}. The above problem can be effectively solved by orthogonal matching pursuit (OMP) [128]. 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:

$$\text{class}(\mathbf{x}^{\text{test}}) = \arg\min_{j=1,\cdots,C} \left\|\mathbf{x}^{\text{test}} - \mathbf{D}_j\alpha_j^{\text{test}}\right\|_2. \tag{19}$$

Since the pixel-wise sparse model does not consider the spatial information of the HSI, the obtained classification map usually appears very noisy. To incorporate the spatial context, a joint sparse model (JSM) was proposed in [125] to utilize 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 $\mathbf{X}^{test}$, while the center pixel is denoted by $\mathbf{x}^{\text{test}}$, JSM aims to obtain the sparse matrix of $\mathbf{X}^{\text{test}}$ by addressing the following problem:

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

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 as simultaneous OMP (SOMP) [129] can be used to solve the above problem. 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:

$$\text{class}(\mathbf{x}^{\text{test}}) = \operatorname*{argmin}_{j=1,\cdots,C} \left\|\mathbf{X}^{\text{test}} - \mathbf{D}_j\mathbf{A}_j^{\text{test}}\right\|_F. \tag{21}$$

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

To sufficiently exploit spectral-spatial information of the HSI, some recent works incorporated different kinds of spatial information into the sparse model [130–133]. In [130], since 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 adaptive sparse representation model to adaptively utilize information among regions of different scales for HSI classification. In [131], Fu *et al*. adaptively 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. The above multiscale and shape adaptive sparse 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 above two sparse models only aim to classify its centered pixel. In [132, 133], instead of classifying each pixel, the HSI was directly segmented into many superpixels and a discriminative sparse model was used to classify the whole superpixel, thus greatly enhancing the efficiency.

On the other hand, some effective spectral-spatial feature extraction methods were combined with the sparse model to improve the classification performance [134–138]. In [134], Song *et al*. first adopted the morphological attribute profiles discussed in Section IV to extract the spatial features and then used the OMP algorithm to classify the extracted features. In [135], 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 [136], Tang *et al*. transformed the original HSI into the high dimensional manifold feature space, and then, the sparse model could be used to effectively reflect the local structures of HSI, which provided promising classification results. In [137], a series of Gabor wavelet filters with different scales and frequencies was first applied on the original HSI to extract the spectral-spatial features, and then, a multi-task sparse model was proposed to exploit the correlations among the features for classification. The above methods only extracted one kind of feature from the HSI. Since different features can reflect the spectral-spectral information of the HSI from different perspectives, Fang *et al*. [138] first extracted multiple features (e.g., Gabor texture, morphological profile, and differential morphological profile) 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.

B. Experimental Results

1) Experimental Setup: In this section, seven well-known sparse representation-based classification methods are used for comparisons. We denote the pixel-wise sparse representation method as SRC [127]. The fixed region based sparse representation method is denoted as the JSRC [127]. The region sizes for the JSRC are set to 5×5, 3×3, and 5×5 for the AVIRIS Indian Pines, the ROSIS-03 Pavia University, and the CASI Houston University images, respectively. The EMAP+SRC [136] performs the SRC classifier on the EMAP extracted features. MASR is the multiscale adaptive sparse representation method [132], which utilizes seven different scales ranging from the 3×3 to 15×15 (see Figs. 14(e), 15(e), and 16(e)). SBSDM [134] stands for the superpixel based sparse classifier and the superpixel numbers are chosen to be 200, 1000, 2000 for the AVIRIS Indian Pines, the ROSIS-03 Pavia University, and the CASI Houston University images, respectively. SAS [133] is the shape adaptive sparse classifier. MFASR [138] is the multiple features-based sparse classifier, where four features (including the spectral pixel, Gabor texture, morphological profile, and differential morphological profile) are used. The sparsity level for the above seven classifiers is set to 3 for the three test images. The classification accuracies for the above seven methods on three test images are tabulated in Table XIV-XVI, respectively.

2) Results and Discussions: From Tables XIV-XVI, the following points can be observed: By only utilizing the spectral information, SRC generally delivers the worst classification result. By further considering the spatial information within a fixed-size region, JSRC can achieve a slight improvement on the three test images. In addition, by adjusting the spatial region according to the HSIs 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, by utilizing the information among multiple features, MFASR generally achieves the best classification results on the AVIRIS Indian Pines and the CASI Houston University images. This shows that combining the sparse classifier with multiple features is an effective way to obtain high accuracy. In addition, to analyze if the unmixing treatment has any effects on the performance, an unmixing-based sparse method (called Unmixing+SRC), which first utilizes a well-known unmixing technique [139] to extract the feature of each pixel, and then, applies SRC to the features for classification, is used. As reported in [140], the dimension for each unmixing feature vector is $2C \times 1$.

Indeed, spectral information is the most important characteristic available in HSI, and with such rich spectral information, HSI 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, and thus, they are hard to be distinguished. On the one hand, utilizing the unmixing technique is an effective way to reduce the spectral mixture problem for HSI classification. As can be observed in Tables XIV and XV, Unmixing+SRC generally outperforms SRC in most of the classes of the Indian Pines and Pavia University images. However, setting the number of endmembers in unmixing for different HSIs is a tricky problem. Since the Houston image is very complex, unmixing-based features may not be effectively extracted, and therefore,

the Unmixing+SRC method cannot deliver very good performance. On the other hand, as can be seen in Tables XIV-XVI, the JSRC, EMAP+SRC, MASR, SBSDM, SAS, and MFASR methods, which jointly utilize spectral and spatial information, may perform better, in terms of OA, than SRC, which only utilizes the spectral information. The main improvement of the spectral-spatial-based methods over the spectral-based method comes from the classes with large homogeneous spatial regions (e.g., classes #1, #8, and #11 in Indian Pines, classes #2 and #8 in Pavia University, and classes #9 and #12 in Houston). For some classes with detailed structures (e.g., class #16 in Indian Pines, class #4 in Pavia University, and classes #6, #14 and #15 in Houston), the spectral-based SRC can perform well and even better than those of spectral-spatial-based methods.

However, the pixel-wise SRC is an efficient classifier, since it only needs to classify one pixel at a time. By utilizing 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 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.

VIII. Deep Learning-Based Spectral-Spatial Classifiers

A. Motivation and Background

Deep learning regards a kind of neural network with two or more hidden layers (the input and output layers are not included). The usage of multiple layers tends to extract abstract, invariant, and discriminant features of inputs, which are very useful for the following processing steps including classification, detection, and segmentation [141]. Indeed, considering the task of classification, linear support vector machine and logistic regression are believed to have one layer, and decision tree or support vector machine with kernels can be attributed as two-layer classifiers [142]. Compared with traditional classification methods, deep learning-based classifiers have great potential to obtain high classification performance when facing complex inputs.

As discussed in Section I, hyperspectral sensors obtain spectral and spatial information simultaneously, and the imaging mechanism makes the data inherently complex. Furthermore, due to the complex atmosphere condition, scattering from neighboring objects, and intra-class variability, it is difficult to extract discriminative and robust features of HSI for accurate classification. On the other hand, it is believed that the 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 the deep networks, each layer can extract the features of the previous layer. In 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.

B. Deep Learning-Based Methods for HSI Classification

1) General framework of deep learning-based methods for HSI 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 (AE), denoising autoencoders (DAE), restricted Boltzmann machines (RBM), convolutional neural networks (CNN), and recurrent layers [143]. Using layer-wise units, various deep models can be established. Deep learning involves a number of models including stack autoencoder (SAE), deep belief network (DBN), deep CNN, and deep recurrent neural network (RNN) [143]. All of the aforementioned deep learning models have been investigated for HSI classification. Deep learning-based methods have shown their capability in the application to HSI [144].

From Fig. 12, 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 3D inputs, which are fed to deep models. Deep learning models hierarchically extract the discriminant features of the inputs, and usually use a softmax classifier to obtain the final classification results.

In general, deep learning models have lots of parameters (i.e., weights) to be tuned in the training procedure, which means a large number of training samples is needed. Without enough training samples, deep models face a problem known as overfitting, which means that the classification performance of test data will be downgraded. This problem becomes serious when fully connected models, including stack autoencoder and deep belief networks, are used for HSI classification. Due to the shared weights and local connections in CNNs, the number of parameters are 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 the review of CNN-based HSI classification methods since its superior performance over other fully connected networks has already been demonstrated in the literature.

2) The core parts and techniques of deep CNNs: A deep CNN usually contains several convolution layers, several non-linear transformation layers, and several pooling layers [143]. The convolution and nonlinear transformation can be defined as follows:

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

where $f(.)$ is a nonlinear function and $*$ is the convolution operation. The matrix $\mathbf{x}_j^l$ is the j-th feature map of the current (l)-th layer, and $\mathbf{x}_i^{l-1}$ is the i-th 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

TABLE XIV: AVIRIS Indian Pines - 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	100.00	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	100.00	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 XV: ROSIS-03 Pavia University - 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

TABLE XVI: Houston University - 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	100.00	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	100.00	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

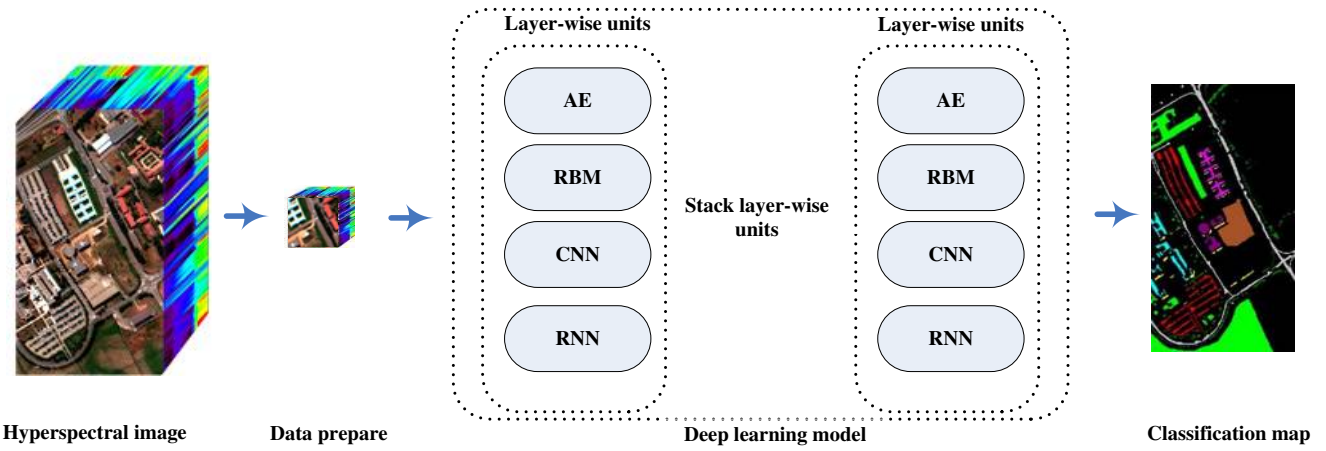


Fig. 12: The general framework of deep learning-based methods for HSI classification.

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 such advantages 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 limited availability of training samples in HSI classification, deep models are facing the serious problem of overfitting. To address the problem in HSI classification, dropout has been widely used. Furthermore, in order to achieve better performance in terms of classification accuracy, batch normalization is adopted in a variety of studies to obtain better model generalization. Below, we describe dropout and batch normalization in more detail.

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 CNN forms a different neural network in each training epoch by dropping neurons randomly. The inherent ensemble method efficiently mitigate the overfitting problem in classification [145].

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. Due to the aforementioned advantages, Batch normalization is a practical tool in CNN training [146].

3) Recent CNN-based methods for HSI classification: CNNs can be used as spectral classifiers. In order to fully use the spatial information provided by HSI, 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 Fig. 13. Typical methods are presented as follows.

In [147], 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 1D vector, which overlooked the spatial patterns. The CNN-based method in [147] takes the voxels in a neighbourhood region into consideration, which obtained good classification 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 3D data, so in [148], 3D CNNs are designed to extract the spectral-spatial features of HSIs. 3D convolution filters reduces the trainable parameters in CNNs, which lead to good classification accuracy.

CNN can be combined with other techniques to further improve the classification performance. In [149], an HSI classification framework is proposed, which was a combination of deep CNN and sparse representation. In the method, the learnt features from CNN were refined by sparse representation, and then followed by a classifier. In [150], a powerful spatial feature extraction, attribute filtering (discussed in Section IV), was combined with deep CNN. The method led to a better performance compared with each involved approaches individually. Furthermore, in [151], Gabor filtering was used to effectively extract spatial information in HSIs, and then, CNN was used for further processing. The methods obtained competitive results even when a limited number of training samples was available.

C. Experimental Results and Discussion

In this experiment, four different CNN-based methods are considered 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 3D images in PCA-CNN, EMP-CNN, and Gabor-CNN. Three principal components have been preserved in all the aforementioned approaches. For the CNN method, however, all bands are used so the input 3D images are with the size of $27 \times 27 \times d$. Similarly, we set

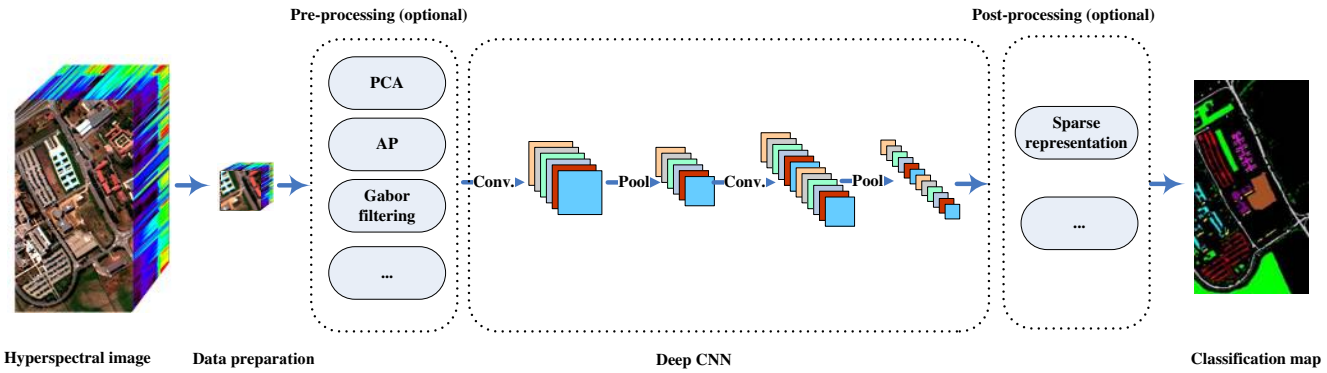


Fig. 13: The general framework of deep CNN-based methods for HSI classification.

the input size to be $11 \times 11 \times 3$ and $11 \times 11 \times d$ in these methods using PCA and without PCA, respectively, on the Houston data set. In the EMP-CNN method, three principal components from HSIs are computed, and then, the opening and closing operations are used to extract spatial information on the first three components. In the experiments, the shape of the structuring element (SE) is set as 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 mini-batch is set to 100 and the number of training epochs for these CNN-based methods is 200.

The classification results of the above-mentioned methods for all the three data sets are shown in Tables XIV, XV, and XVI. 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 Figs. 14(h), 15(h), and 16(h)), followed by the EMP-CNN and PCA-CNN. In addition, 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 CNN, respectively. Besides, it also outperforms the EMP-CNN by 0.25%, 0.58%, and 0.0047 in terms of OA, AA, and Kappa. Furthermore, results shown in Table XIV, XV, and XVI demonstrate similar trend 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 CNN on the Pavia University data set. It also obtains a superior performance compared with the CNN method, and on the Houston data set. The PCA-CNN obtains a better classification performance which is higher than CNN by 0.47%, 1.57%, and 0.0104 in terms of OA, AA, and Kappa, respectively.

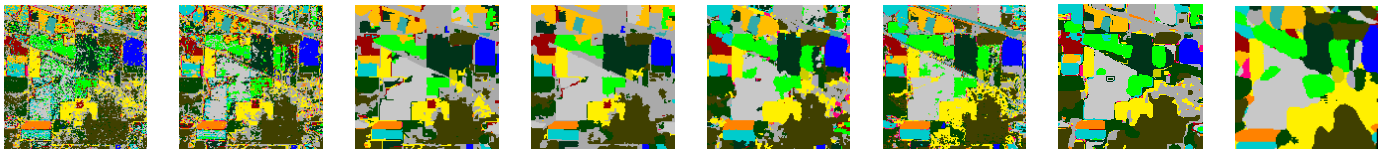
IX. Conclusion and Possible Future Works

In this paper, a closer look has been taken at recent advances in spectral-spatial classification of hyperspectral images. Five branches of spectral-spatial classification techniques based on mathematical morphology, MRFs, segmentation, sparse representation, and deep learning have been reviewed both methodologically and through examples of experimental results, in order to discuss how they address the task of incorporating spatial information into an HSI classification chain. As expected, the results confirm that the inclusion of spatial information in the classification system can significantly improve classification accuracies compared to the situation when spatial information is discarded, significantly contributes to the extraction of the shape of different objects, and addresses the salt and pepper appearance problem typical for spectral classifiers. From this perspective, the families of methods discussed in the paper 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 hand-crafted features and deep features learned from data, respectively. Markovian methods and sparsity-based techniques operate at the classification stage through probabilistic spatial-contextual priors and through a data-representation viewpoint, respectively. Segmentation-based algorithms extract and use information on the regions in the imaged scene.

Consistently with the goal of providing a methodological review and not an experimental comparative study, no special focus was devoted to making model selection or numerical optimization issues homogeneous across the aforementioned families of methods. Nevertheless, the considered approaches overall obtained high and comparable accuracies on the collection of three considered data sets, which included the very well-known Indian Pines and Pavia University data sets and the recent Houston University data set. This scenario confirms the effectiveness of current spatial-spectral approaches to the topical problem of hyperspectral image classification.

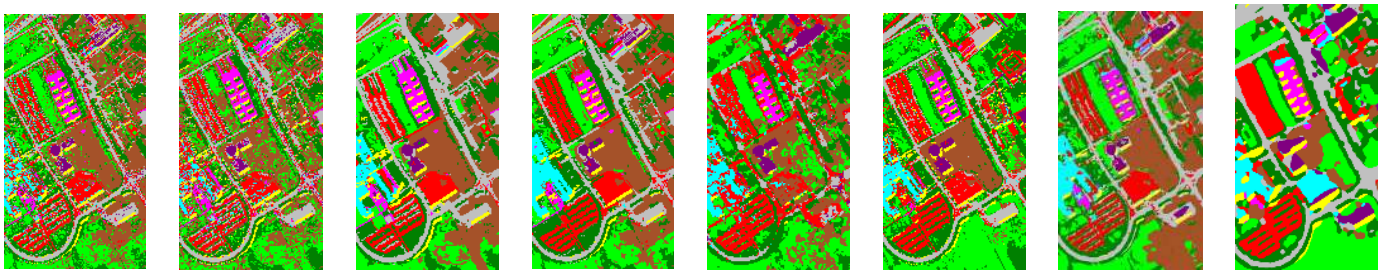
Although the area of spectral-spatial classification of hyperspectral images has been a hot spot in recent years, there are still several aspects worth to be further investigated. Here, we provide readers with pointers to several high potential aspects, which can be followed as possible future works.

1) As shown, extinction profiles 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 in sparse representation for remote



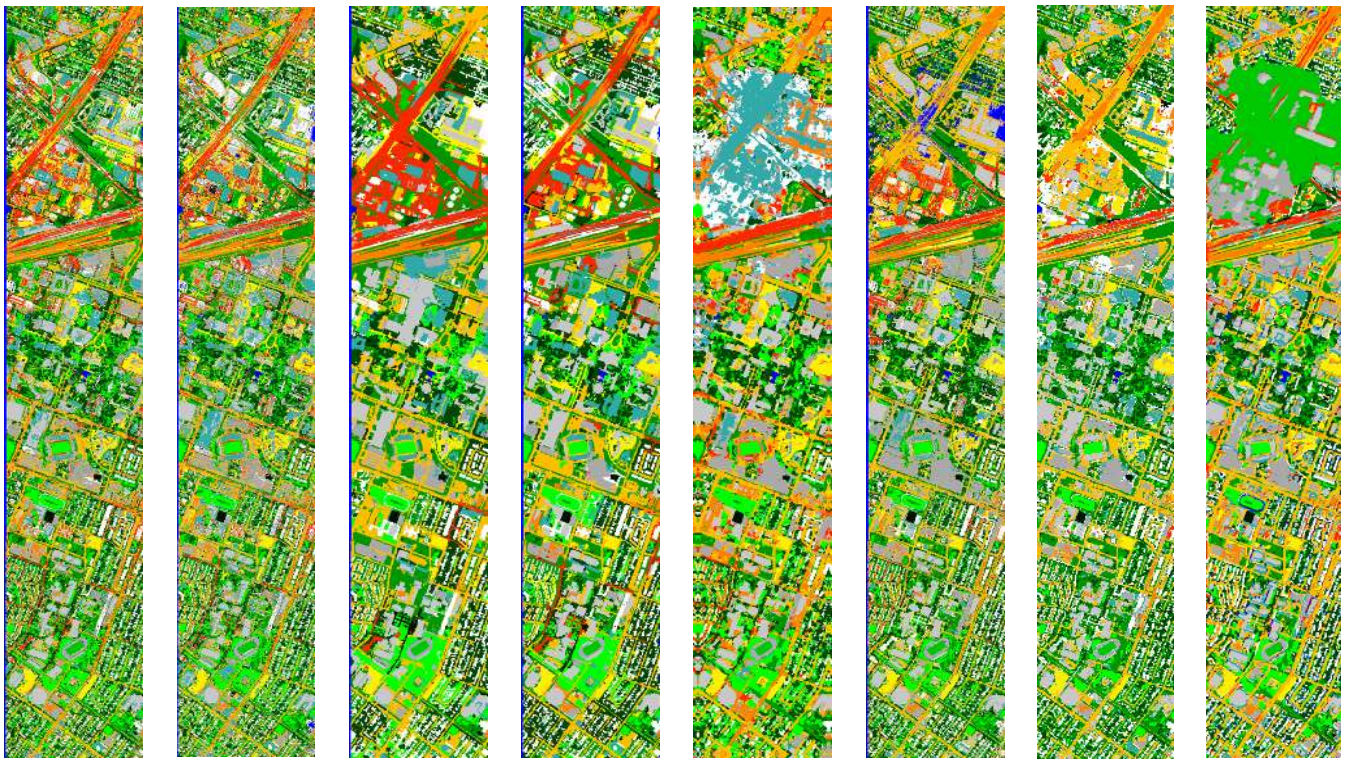
(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 14: Classification maps obtained on AVIRIS Indian Pines: (a) SVM, (b) RF, (c) BPT $\alpha = 0$, (d) BPT $\alpha = 0.5$, (e) MASR, (f) EMEP, (g) MSVC with graph cuts, and (h) Gabor-CNN.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 15: Classification maps obtained on ROSIS-03 Pavia University: (a) SVM, (b) RF, (c) BPT $\alpha = 0$, (d) BPT $\alpha = 0.5$, (e) MASR, (f) EMEP, (g) MSVC with LBP, and (h) Gabor-CNN.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 16: Classification maps obtained on the CASI Houston University data: (a) SVM, (b) RF, (c) BPT $\alpha = 0$, (d) BPT $\alpha = 0.5$, (e) MASR, (f) CK_{EMEP}, (g) MSVC with LBP, and (h) Gabor-CNN.

sensing image classification is to investigate possible solutions for involving spatial information into the model. In addition, the construction of extinction profiles leads to a very sparse feature space. This encourages one to integrate extinction profiles along with sparse and low-rank techniques to further improve classification accuracies and solve the curse of dimensionality at the same time.

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 Section VIII, 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 pixelwise classification, they tend to yield overly unstructured or "blobby" classification maps [152]. 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 hyperspectral image classification.
4) Markov random fields 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, Markov random fields have been found especially effective for HSI classification when integrated with kernels, 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 in developing advanced hierarchical Markov models [154] that allow incorporating multiscale information characterized by segmentation, feature extraction, or CNNs [4], thus possibly bridging to the 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 gaining 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, and many of them 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 in the machine learning and remote sensing communities.
   - Generative adversarial network is an active topic, which has already shown its advantages in the remote sensing community in terms of image translation and data classification [155, 156]. Although the effectiveness of the generative adversarial network has very recently confirmed for spectral-spatial classification of 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.

X. Utilized Codes

Most implementations of the methods described in the paper are made available to the research community. The software and codes for BPT-based classification described in Section VI can be found on http://ooclassif.gforge.inria.fr.

The max-tree and extinction filters 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 MAC operating systems. For Windows users a docker [157] 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 MASR, SAS, and MFASR methods can be found at http://www.escience.cn/people/LeyuanFang/index.html.

The sets of training and test samples utilized in this paper can be found at https://pghamisi.wixsite.com/mysite.

XI. Acknowledgment

The ROSIS-03 Pavia University and AVIRIS Indian Pines data and the corresponding reference information were kindly provided by Prof. Paolo Gamba from the University of Pavia, Italy and Prof. David Landgrebe from Purdue University, respectively. In addition, the authors would like to thank the National Center for Airborne Laser Mapping (NCALM) at the University of Houston for providing the CASI Houston data set, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for organizing the 2013 Data Fusion Contest. The shadow-removed hyperspectral data is provided by Prof. Naoto Yokoya.

Pedram Ghamisi (S'12-M'15-SM'18) received the B.Sc. degree in civil (survey) engineering from the Tehran South Campus of Azad University, Tehran, Iran, the M.Sc. (First Class Hons.) degree in remote sensing from the K. N. Toosi University of Technology, Tehran, in 2012, and the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2015. In 2013 and 2014, he joined the School of Geography, Planning and Environmental Management, University of Queensland, Brisbane, QLD, Australia. He was a Post-Doctoral Research Fellow with the University of Iceland. He was a Post-Doctoral Research Fellow with the Technical University of Munich, Germany, and Heidelberg University, Germany, from 2015 to 2017. He was also a Researcher with the German Aerospace Center (DLR) from 2015 to 2017. He worked as a research scientist at DLR, Remote Sensing Technology Institute (IMF), Germany from October 2017 to May 2018. From June 2018, he works as the head of the machine learning group at Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology (HIF).

His research interests include remote sensing and image analysis, with a special focus on the spectral and spatial techniques for hyperspectral image classification, multisensor data fusion, and machine/deep learning.

Dr. Ghamisi received the Best Researcher Award for M.Sc. students from the K. N. Toosi University of Technology in the academic year of 2010-2011. In 2013, he was awarded the IEEE Mikio Takagi Prize for winning the conference Student Paper Competition at the IEEE International Geoscience and Remote Sensing Symposium, Melbourne, VIC, Australia. In 2016, he was selected as a Talented International Researcher by the Iran's National Elites Foundation. In 2017, he won the Data Fusion Contest 2017 organized by the Image Analysis and Data Fusion Technical Committee of the Geoscience and Remote Sensing Society. He was the winner of the 2017 Best Reviewer Prize of IEEE Geoscience and Remote Sensing Letters (IEEE GRSL). He received the prestigious Alexander von Humboldt Fellowship in 2015. For more info, please see http://pedram-ghamisi.com/.

Emmanuel Maggiori received the Engineering degree in computer science from National University of Central Buenos Aires (Unicen), Tandil, Argentina, in 2014. The same year he joined AYIN and STARS teams at Inria Sophia Antipolis-Mediterranée 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. within TITANE team at Inria, studying machine learning techniques for large-scale processing of satellite imagery.

Shutao Li (M'07-SM'15) received the B.S., M.S., and Ph.D. degrees from Hunan University, Changsha, China, in 1995, 1997, and 2001, respectively. He was a Research Associate with the Department of Computer Science, the Hong Kong University of Science and Technology, Hong Kong, in 2011. From 2002 to 2003, he was a Post-Doctoral Fellow with the Royal Holloway College, University of London, London, U.K., with Prof. John Shawe-Taylor. In 2005, he visited the Department of Computer Science, Hong Kong University of Science and Technology as a Visiting Professor. He joined the College of Electrical and Information Engineering, Hunan University, in 2001. He is currently a Full Professor with the College of Electrical and Information Engineering, Hunan University. He has authored or co-authored over 160 refereed papers. His current research interests include compressive sensing, sparse representation, image processing, and pattern recognition.

He is now an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, and a member of the Editorial Board of the Information Fusion and the Sensing and Imaging. He was a recipient of two 2nd-Grade National Awards at the Science and Technology Progress of China in 2004 and 2006.

Roberto Souza is a dual citizen from Brazil and the United States currently working as a Postdoctoral Fellow at the Seaman Family MR Centre in Calgary. He has a B.Sc. in electrical engineering at the Federal University of Par, a M.Sc. and Ph.D. in computer engineering at the University of Campinas. He has international experience having worked as an intern at the Grenoble Institute of Technology, France, and the University of Pennsylvania, United States. Dr. Souza has extensive expertise in image processing, especially techniques based on mathematical morphology, and machine learning. He is a defender of reproducible research, public data and code. He is the manager and conceiver of the Calgary-Campinas-359 public dataset and the open-source max-tree toolbox (iamxt).

Yuliya Tarabalka (S'08–M'10) received the B.S. degree in computer science from Ternopil Ivan Pul'uj State Technical University, Ukraine, in 2005 and the 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, in 2010.

From July 2007 to January 2008, she was a researcher with the Norwegian Defence Research Establishment, Norway. From September 2010 to December 2011, she was a postdoctoral research fellow with the Computational and Information Sciences and Technology Office, NASA Goddard Space Flight Center, Greenbelt, MD. From January to August 2012 she was a postdoctoral research fellow with the French Space Agency (CNES) and Inria Sophia Antipolis-Méditerranée, France. She is currently a researcher with the TITANE team of Inria Sophia Antipolis-Méditerranée. Her research interests are in the areas of image processing, pattern recognition and development of efficient algorithms. She is Member of the IEEE Society.

Gabriele Moser (S'03M'05SM'14) received the Laurea (M.Sc. equivalent) degree in telecommunications engineering, and the Ph.D. degree in space sciences and engineering from the University of Genoa, Italy, in 2001 and 2005, respectively. Since 2014, he has been an Associate Professor of Telecommunications at the University of Genoa. Since 2001, he has cooperated with the Image Processing and Pattern Recognition for Remote Sensing laboratory of the University of Genoa. Since 2013, he has been the Head of the Remote Sensing for Environment and Sustainability laboratory at the Savona Campus of the University of Genoa. From January to March 2004, he was a Visiting Student with the Institut National de Recherche en Informatique et en Automatique (INRIA), Sophia Antipolis, France. From 2012 to 2016, he was an external collaborator of the Ayin laboratory at INRIA. In 2016, he spent a period as Visiting Professor at the Institut National Polytechnique de Toulouse, France. His research activity is focused on pattern recognition and image processing methodologies for remote sensing and energy applications. He has been an Area Editor of PATTERN RECOGNITION LETTERS (PRL) since 2015, and an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2008. He was an Associate Editor of PRL from 2011 to 2015, and a Guest Co-Editor of the September 2015 special issue of the IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE. He served as Chair of the IEEE GRSS Image Analysis and Data Fusion Technical Committee (IADF TC) from 2013 to 2015, and as IADF TC Co-Chair from 2015 to 2017. He was Publication Co-Chair of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Technical Program Co-Chair of the IEEE GRSS EARTHVISION workshop at the 2015 IEEE/CVF Computer Vision and Pattern Recognition conference (CVPR), and Co-Organizer of the second edition of EARTHVISION at CVPR 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 IGARSS 2016.

Andrea De Giorgi (S'15) received the M.Sc. degree in telecommunications engineering and the 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 Dept. 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 co-recipient of the Interactive Symposium Paper Award at IEEE International Geoscience and Remote Sensing Symposium 2016.

Leyuan Fang (S'10-M'14-SM'17) received the Ph.D. degrees from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2008 and 2015, respectively.

From September 2011 to September 2012, he was a visiting Ph.D. student with the Department of Ophthalmology, Duke University, Durham, NC, USA, supported by the China Scholarship Council. From August 2016 to September 2017, he was a Postdoctoral Research Fellow with the Department of Biomedical Engineering, Duke University, Durham, NC, USA. Since Jan. 2017, he has been an associate professor with the College of Electrical and Information Engineering, Hunan University. His research interests include sparse representation and multi-resolution analysis in remote sensing and medical image processing. He has won the Scholarship Award for Excellent Doctoral Student granted by Chinese Ministry of Education in 2011.

Yushi Chen received the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2008. Currently, he is an Associate Professor in the School of Electronics and Information Engineering, Harbin Institute of Technology, China. His research interests include remote sensing data processing and machine learning. For more details, please refer to http://homepage.hit.edu.cn/chenyushi.

Mingmin Chi (M'05) received the B.S. and M.S. degrees in electrical engineering from Changchun University of Science and Technology, Changchun, China in 1998 and Xiamen University, Xiamen, China in 2002, respectively, and the Ph.D. degree in computer science from University of Trento, Trento, Italy in 2006. Also, she was a student visitor from May 2005 to Mar. 2006 at the Dept. of Schoelkopf in Max-Planck Institute for Biological Cybernetics, Tuebingen, Germany. Currently, she is an associate professor at School of Computer Science in Fudan University, Group leader at Shanghai Key Laboratory of Data Science, Shanghai, China. She is a Guest Editor for the special issue on big data in remote sensing for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) 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, etc..

Sebastiano B. Serpico Fellow of the IEEE, 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 received the Laurea (M.S.) degree in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Italy. He is the Head of the research group on Signal Processing and Recognition Methods and Systems of the Department of Electrical, Electronic, Telecommunications Engineering, and Naval Architecture of the University of Genoa. His current research interests include pattern recognition for remote sensing images. He is the Chairman of the Institute of Advanced Studies in Information and Communication Technologies (ISICT). He has been the project manager of numerous research projects and an evaluator of project proposals for various programs of the European Union, Italian Space Agency, Italian Ministry of Education and Research, etc. He is the author (or coauthor) of over 200 scientific articles published in journals and conference proceedings. He received the Best Paper Award at the 2010 IEEE Workshop on Hyperspectral Image and Signal Processing and the Interactive Symposium Paper Award at IEEE IGARSS 2016. He is an associate editor of the international journal IEEE Transactions on Geoscience and Remote Sensing (TGRS). He was a guest editor of two Special Issues of TGRS on the subject of the analysis of hyperspectral image data (July 2001 issue) and the subject Advances in techniques for the analysis of remote sensing data (March 2005 issue), and of the IEEE Proceedings journal on the subject of Remote Sensing of Natural Disasters (October 2012 issue). From 1998 to 2002, he was the chairman of the SPIE/EUROPTO series of conferences on Signal and Image Processing for Remote Sensing. He was Co-Chair of the IEEE International Geoscience and Remote Sensing Symposium in 2015 (Milan, Italy, July 2015).

Jón Atli Benediktsson received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, in 1984, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1987 and 1990, respectively. On July 1, 2015 he became the President and Rector of the University of Iceland. From 2009 to 2015 he was the Pro Rector of Science and Academic Affairs and Professor of Electrical and Computer Engineering at the University of Iceland. His research interests are in remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in those fields. Prof. Benediktsson was the 2011-2012 President of the IEEE Geoscience and and Remote Sensing Society (GRSS) and has been on the GRSS AdCom since 2000. He was Editor in Chief of the IEEE Transactions on Geoscience and Remote Sensing (TGRS) from 2003 to 2008 and has served as Associate Editor of TGRS since 1999, the IEEE Geoscience and Remote Sensing Letters since 2003 and IEEE Access since 2013. He is on the Editorial Board of the Proceedings of the IEEE, the International Editorial Board of the International Journal of Image and Data Fusion, the Editorial Board of Remote Sensing, and was the Chairman of the Steering Committee of IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS) 2007-2010. Prof. Benediktsson is a co-founder of the biomedical start up company Oxymap (www.oxymap.com). He is a Fellow of the IEEE and a Fellow of SPIE. Prof. Benediktsson was a member of the 2014 IEEE Fellow Committee. He received the Stevan J. Kristof Award from Purdue University in 1991 as outstanding graduate student in remote sensing. In 1997, Dr. Benediktsson was the recipient of the Icelandic Research Council's Outstanding Young Researcher Award, in 2000, he was granted the IEEE Third Millennium Medal, in 2004, he was a co-recipient of the University of Iceland's Technology Innovation Award, in 2006 he received the yearly research award from the Engineering Research Institute of the University of Iceland, and in 2007, he received the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society and the OECE Award from the School of ECE, Purdue University in 2016. He was co-recipient of the 2012 IEEE Transactions on Geoscience and Remote Sensing Paper Award and in 2013 he was co-recipient of the IEEE GRSS Highest Impact Paper Award. In 2013 he received the IEEE/VFI Electrical Engineer of the Year Award. In 2014 he was a co-recipient of the International Journal of Image and Data Fusion Best Paper Award. He is a member of the Association of Chartered Engineers in Iceland (VFI), Societas Scinetiarum Islandica and Tau Beta Pi.

REFERENCES

[1] J. A. Benediktsson and P. Ghamisi, *Spectral-Spatial Classification of Hyperspectral Remote Sensing Images*. Artech House Publishers, INC, Boston, USA, 2015.

[2] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 1, pp. 8–32, March 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. Remote Sens. Geos.*, 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," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 631–651, March 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 Transactions on Geoscience and Remote Sensing*, 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. Geos. 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," *Proceedings of the 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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. Geos. 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. on Geoscience and Remote Sensing*, 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. Geos. 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. Geos. 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. Geos. 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 Geos. Remote Sens. Let.*, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2405–2418, 2014.

[17] "2013 IEEE GRSS Data Fusion Contest," 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. Geos. 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. Jour. 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, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 98–110.

[22] E. Carlinet and T. Géraud, "A comparative review of component tree computation algorithms," *IEEE Transactions on Image Processing*, 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 *2016 29th SIBGRAPI Conf. Graph., Pat. Images Tut.*, Oct 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. Springer Berlin Heidelberg, 1994, vol. 126, pp. 197–208.

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

[26] J. Fabrizio and B. Marcotegui, *Fast Implementation of the Ultimate Opening*. Springer Berlin Heidelberg, 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*. Springer International Publishing, 2015, pp. 157–168.

[28] F. Kiwanuka and M. Wilkinson, *Cluster Based Vector Attribute Filtering*. Springer International Publishing, 2015, pp. 277–288.

[29] M. Grimaud, "New measure of contrast: the dynamics," vol. 1769, 1992, pp. 292–305.

[30] G. Bertrand, "On the dynamics," *Image and Vision Computing*, 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 *XXI Brazilian Symposium on Computer Graphics and Image Processing, 2008. SIBGRAPI '08.*, 2008, pp. 204–211.

[32] E. J. Breen and R. Jones, "Attribute openings, thinnings, and granulometries," *Comput. Vis. Image Underst.*, 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 *ISMM'15, Reykjavik, Iceland, May 27-29, 2015. Proceedings*, pp. 63–74, 2015.

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

[35] Y. Xu, T. Géraud, and L. Najman, "Morphological filtering in shape spaces: Applications using tree-based image representations," in *2012 21st International Conference on Pattern Recognition (ICPR)*, 2012, pp. 485–488.

[36] P. Monasse and F. Guichard, "Fast computation of a contrast-invariant image representation," *IEEE Trans. Image Processs.*, 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*, ser. Lecture Notes in Computer Science Series, vol. 7883. Springer, 2013, pp. 390–401.

[38] Y. Xu, E. Carlinet, T. Geraud, and L. Najman, "Hierarchical segmentation using tree-based shape space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[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. Geos. 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 *2017 IGARSS*, July 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. Geos. Remote Sens.*, vol. PP, no. 99, pp. 1–13, 2017.

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

[44] M. Zhang, P. Ghamisi, and W. Li, "Classification of hyperspectral and lidar data using extinction profiles with feature fusion," *Remote Sens. Let.*, 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 Jour. Sel. Top. App. Earth Obs. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, June 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. Geos. Remote Sens.*, vol. 55, no. 11, pp. 6354–6365, Nov 2017.

[47] R. Souza, L. Rittner, R. Lotufo, and R. Machado, "An array-based node-oriented max-tree representation," in *ICIP'15*, Sept 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 classfication," *IEEE Geos. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, 2006.

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

[50] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[51] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2000.

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

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

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

[55] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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 Transactions on Geoscience and Remote Sensing*, 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*

Sensing, vol. 8, no. 4, 2016.

[58] G. Moser and S. Serpico, "Combining support vector machines and Markov random fields in an integrated framework for contextual image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 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 Journal on Selected Topics in Signal Processing*, 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 Transactions on Geoscience and Remote Sensing*, 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 Transactions on Geoscience and Remote Sensing*, 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 Transactions on Image Processing*, 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, 2017.

[64] P. Chen and J. I. Tourneret, "Toward a sparse Bayesian Markov random field approach to hyperspectral unmixing and classification," *IEEE Transactions on Image Processing*, 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 Transactions on Geoscience and Remote Sensing*, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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 Transactions on Image Processing*, vol. 22, no. 1, pp. 5–16, 2013.

[68] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, 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 International Journal of Geo-Information*, vol. 7, no. 1, 2018.

[70] P. Zhong and R. Wang, "Jointly learning the hybrid CRF and MLR model for simultaneous denoising and classification of hyperspectral imagery," *IEEE Transactions on Neural Networks and Learning Systems*, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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, Learning, and 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," *Statistics, Optimization and Information Computing*, 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 Transactions on Geoscience and Remote Sensing*, 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 Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 31–42, 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 Sensing*, vol. 9, no. 8, 2017.

[77] H. Derin and P. Kelly, "Discrete-index Markov-type random processes," *Proceedings of the 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, 2004.

[79] Q. Jackson and D. Landgrebe, "Adaptive Bayesian contextual classification based on Markov random fields," *IEEE Transactions on Geoscience and Remote Sensing*, 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 Geoscience and Remote Sensing Letters*, 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 Transactions on Geoscience and Remote Sensing*, 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 Transactions on Geoscience and Remote Sensing*, 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 Transactions on Geoscience and Remote Sensing*, 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," *Journal of the Royal Statistical Society. 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 Transactions on Pattern Analysis and Machine Intelligence*, 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 Transactions on Pattern Analysis and Machine Intelligence*, 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," *Journal of Machine Learning Research*, vol. 6, 2005.

[88] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, 2003, pp. 900–907.

[89] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, 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 Transactions on Information Theory*, vol. 51, no. 11, pp. 3697–3717, 2005.

[91] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1568–1583, 2006.

[92] J. Besag, "Statistical analysis of dirty pictures," *Journal of Applied Statistics*, 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 University Press, 2000.

[94] J. Platt, *Advances in Large Margin Classiers*. MIT Press, 2000, ch. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods.

[95] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *The Journal of Machine Learning Research*, 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 *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2014, pp. 3722–3725.

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

[98] S. Serpico and G. Moser, "Weight parameter optimization by the Ho-Kashyap algorithm in MRF models for supervised image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 12, pp. 3695–3705, 2006.

[99] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27:127:27, 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 1068–1080, 2008.

[101] P. Gurram and H. Kwon, "Contextual SVM using Hilbert space embedding for hyperspectral classification," *IEEE Geoscience and Remote Sensing Letters*, 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 Geoscience and Remote Sensing Letters*, vol. 7, no. 4, pp. 741–745, 2010.

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

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

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

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

[107] R. Gonzalez and R. Woods, *Digital Image Processing, Second Edition*. Prentice Hall, 2002.

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

[109] 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, July 2010.

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

[111] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, J. Angulo, and M. Fauvel, "Classification of hyperspectral data using support vector machines and adaptive neigh-

borhoods," *Proc. 6th EARSeL SIG IS Workshop*, 2009.

[112] 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. Systems, Man, and Cybernetics: Part B*, vol. 40, no. 5, pp. 1267–1279, Oct. 2010.

[113] Y. Tarabalka and A. Rana, "Graph-cut-based model for spectral-spatial classification of hyperspectral images," in *IEEE IGARSS*. IEEE, 2014, pp. 3418–3421.

[114] 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.

[115] E. Maggiori, Y. Tarabalka, and G. Charpiat, "Improved partition trees for multi-class segmentation of remote sensing images," in *IEEE IGARSS*. IEEE, 2015, pp. 1016–1019.

[116] ——, "Optimizing partition trees for multi-object segmentation with shape prior," in *26th British Machine Vision Conference (BMVC)*, 2015.

[117] 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.

[118] 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.

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

[120] 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.

[121] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *ICCV*, 1998, pp. 59–66.

[122] P. Salembier, S. Foucher, and C. López-Martínez, "Low-level processing of PolSAR images with binary partition trees," in *IEEE IGARSS*, 2014.

[123] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.

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

[125] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3973–3985, Oct 2011.

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

[127] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification via kernel sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 217–231, Jan 2013.

[128] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, Nov 1993, pp. 40–44 vol.1.

[129] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572 – 588, 2006, sparse Approximations in Signal and Image ProcessingSparse Approximations in Signal and Image Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168405002227

[130] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 12, pp. 7738–7749, Dec 2014.

[131] W. Fu, S. Li, L. Fang, X. Kang, and J. A. Benediktsson, "Hyperspectral image classification via shape-adaptive joint sparse representation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 556–567, Feb 2016.

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

[133] J. Li, H. Zhang, and L. Zhang, "Efficient superpixel-level multitask joint sparse representation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 10, pp. 5338–5351, Oct 2015.

[134] 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 Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 5122–5136, Aug 2014.

[135] R. Roscher and B. Waske, "Shapelet-based sparse representation for landcover classification of hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1623–1634, March 2016.

[136] Y. Y. Tang, H. Yuan, and L. Li, "Manifold-based sparse representation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 12, pp. 7606–7618, Dec 2014.

[137] 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 Transactions on Geoscience and Remote Sensing, vol. 54, no. 6, pp. 3174–3187, June 2016.

[138] L. Fang, C. Wang, S. Li, and J. A. Benediktsson, "Hyperspectral image classification via multiple-feature-based adaptive sparse representation," *IEEE Transactions on Instrumentation and Measurement*, vol. PP, no. 99, pp. 1–12, 2017.

[139] 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, Jun. 2011.

[140] 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, Aug. 2017.

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

[142] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pat. Analysis Machine Intel.*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[143] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

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

[145] G. E. Hinton and et al., "Improving neural networks by preventing co-adaptation of feature detectors," *Comp. Science*, vol. 3, no. 4, pp. 212–223, June 2012.

[146] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

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

[148] 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. Geos. Remote Sens.*, 2016.

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

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

[151] Y. Chen, L. Zhu, P. Ghamisi, X. Jia, G. Li, and L. Tang, "Hyperspectral images classification with gabor filtering and convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2355–2359, Dec 2017.

[152] E. Maggiori, G. Charpiat, Y. Tarabalka, and P. Alliez, "Recurrent neural networks to correct satellite image classification maps," *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

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

[154] 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 Transactions on Geoscience and Remote Sensing (in print)*, 2016.

[155] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–18, 2018.

[156] P. Ghamisi and N. Yokoya, "Img2dsm: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 794–798, May 2018.

[157] D. Merkel, "Docker: Lightweight linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, Mar. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2600239.2600241