Hyperspectral Image Representation and Processing With Binary Partition Trees

Silvia Valero, *Student Member, IEEE*, Philippe Salembier, *Fellow, IEEE*,
and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—The optimal exploitation of the information provided by hyperspectral images requires the development of advanced image-processing tools. This paper proposes the construction and the processing of a new region-based hierarchical hyperspectral image representation relying on the binary partition tree (BPT). This hierarchical region-based representation can be interpreted as a set of hierarchical regions stored in a tree structure. Hence, the BPT succeeds in presenting: 1) the decomposition of the image in terms of coherent regions, and 2) the inclusion relations of the regions in the scene. Based on region-merging techniques, the BPT construction is investigated by studying the hyperspectral region models and the associated similarity metrics. Once the BPT is constructed, the fixed tree structure allows implementing efficient and advanced application-dependent techniques on it. The application-dependent processing of BPT is generally implemented through a specific pruning of the tree. In this paper, a pruning strategy is proposed and discussed in a classification context. Experimental results on various hyperspectral data sets demonstrate the interest and the good performances of the BPT representation.**

***Index Terms*—Binary partition tree, classification, hyperspectral imaging, segmentation.**

I. Introduction

HYPERSPECTRAL sensors collect multivariate discrete images in a series of narrow and contiguous wavelength bands. The resulting datasets contain numerous image bands, each of them depicting the scene as viewed within a given wavelength λ. The entire data $\mathbf{I}_\lambda$ can be seen as a three dimensional data cube formed by a set of N_z discrete 2D images $\mathbf{I}_\lambda = \{I_{\lambda_j},\ j = 1, \ldots, N_Z\}$. Each I_{λ_j} is formed by a set of N_p pixels where each pixel p represents the spatial coordinates in the image. Consequently, given a specific wavelength λ_j, $I_{\lambda_j}(p)$ is the radiance value of the pixel p on the waveband I_{λ_j}. The spectrum of a pixel as a function of wavelength λ is called the spectral radiance curve and it provides insightful characteristics of the material represented by the pixel.

Hyperspectral imaging enables the characterization of regions based on their spectral properties which provides a rich amount of information. This new source of information has led to the use of such images in a growing number of real-life applications, such as remote sensing, food safety, and healthcare or medical research. However, the price of this wealth of information is a huge amount of data that cannot be fully exploited using traditional imagery analysis tools. Hence, given the wide range of real-life applications, a great deal of research is invested in the field of hyperspectral data processing [1].

A hyperspectral image can be considered as a mapping from a $2D$ spatial space to a spectral space of dimension N_z. The spectral space is important because it contains much more information about the surface of target objects than what can be perceived by human vision. Accordingly, conventional analysis techniques have traditionally focused on the spectral properties of the hyperspectral data by only using the spectral space. These pixel-based procedures analyze the spectral properties of every pixel, without taking into account the spatial or contextual information related to the pixel of interest. In this framework, many different supervised and semi-supervised techniques have been proposed to perfom pixelwise classification [5], [4], [2], [6], [3]. Without taking into consideration the spatial location of the pixels, these techniques assign to each pixel the label corresponding to its predicted class.

In the last few years, the importance of the spatial space and, in particular, of taking into account the spatial correlation has been demonstrated in different contexts such as classification [7], [8]–[10], image segmentation [11], [12], [13] or unmixing [14]. In these techniques, the spatial information is combined with the spectral information. For instance, in a classification context, pixels are classified by their spectral information and also by the information provided by their spatial neighborhood. These approaches have corroborated how essential are the spatial variations and correlation in order to interpret objects in natural scenes.

For this reason, optimal hyperspectral analysis tools should take into account both the spatial and the spectral spaces in order to be robust and efficient. However, the number of wavelengths per pixel and the number of pixels per image, as well as the complexity of jointly handling spatial and spectral correlation explain why this approach is still a largely open

Manuscript received July 28, 2011; revised September 25, 2012; accepted November 13, 2012. Date of publication December 4, 2012; date of current version February 6, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Farhan A. Baqai.

J. Chanussot is with the GIPSA-Laboratory, Signal & Image Department, Grenoble Institute of Technology, Grenoble 38000, France (e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr).

P. Salembier is with the Technical University of Catalonia, Barcelona 08034, Spain (e-mail: philippe.salembier@upc.edu).

S. Valero is with the Centre d'Etudes Spatiales de la BIOSphère, Unité Mixte CNES-CNRS-UPS-IRD, Toulouse 31401 Cedex 9, France (email: silvia.valero@cesbio.cnes.fr).

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

Digital Object Identifier 10.1109/TIP.2012.2231687

research issue for effective and efficient hyperspectral data processing.

The inclusion of the spatial information in hyperspectral analysis is directly related to the definition of a pixel neighborhood. In this context, the work in [15] defines local fixed square neighborhood around each pixel in order to introduce the contextual information. To solve the limitations of the fixed neighborhood window, morphological filters are proposed in [16] to define an adaptive spatial neighborhood having similar characteristics. One problem of such approaches is that the spectral-spatial analysis of hyperspectral images is done at the pixel level. This representation has major drawbacks given that a pixel is the most elementary unit of the image. As a result, hyperspectral image processing at the pixel level has to face major difficulties in terms of scale: the scale of representation is, most of the time, far too low with respect to the interpretation or decision scale.

One the other hand, the definition of the best similar pixel neighborhood (if there is any) is not straightforward. One of the main difficulties is the huge number of applications potentially considered for one given image. Hence, the interpretation of an image at different scales of analysis has led some authors to deal with hierarchical image segmentations. This approach provides a hierarchy of partitions at different levels of resolution through iterative merging steps. In this framework, different hierarchical segmentation techniques previously proposed for mutlispectral data, such as ECHO [17] or e-Cognition [18], have been used in a hyperspectral context. The important difference between the number of spectral bands in multi and hyperspectral data has led to nonoptimal solutions for such approaches. Recently, some hierarchical segmentation techniques working directly with hyperspectral imagery have been presented [19], [20]. The result of these techniques is a final image partition defining a pixel neighborhood. This is obtained by stopping the merging process at some point to reach one single hierarchical level. The stop criterion can depend on different parameters: 1) the number of regions in the case of [20] or 2) an intra-variance statistical criterion [19].

Despite hierarchical segmentations have introduced the importance of the interpretation of an image depending on the scale of observation [9], [10], they suffer from an important drawback. The main problem of such strategy is that they assume that the optimal partition corresponds to one actual level in the hierarchy of segmentations. However, this assumption is rarely true and the techniques following this assumption are unable to deal with situation where coherent objects are found at different levels of the hierarchy. By contrast, as will be seen in the sequel, the technique proposed here does not make this assumption and constructs the final partition by selecting regions at different levels of the hierarchy.

The attractive solution consists in relying on region-based image representations [21], [42]. These representations can be considered as a first abstraction from the pixel-based representation, providing a multiscale hierarchy of regions at different resolution levels. One example of such representations corresponds to Binary Partition Tree (BPT) which was proposed in [22]. This image representation has been successfully used in the past for various applications dealing with color images or video sequences. The BPT is a hierarchical region-based representation having a rather generic construction (to a large extend, application independent). A BPT can be interpreted as a set of hierarchical regions stored in a tree structure. Fig. 1 is an illustration of a BPT where the tree nodes represent image regions and the branches represent the inclusion relationship among the nodes. In this tree representation, three types of nodes can be found. Firstly, leaf nodes representing the regions of an initial partition; secondly, the root node representing the entire image support and finally, the remaining tree nodes representing regions formed by the merging of their two child nodes corresponding to two adjacent regions.

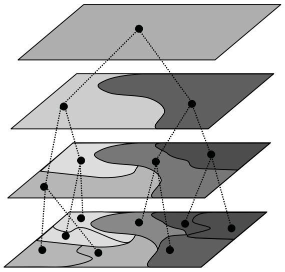


Fig. 1. Example of hierarchical region-based representation using BPT.

The BPT construction is often based on an iterative bottom-up region merging algorithm. Starting from individual pixels or any other initial partition, the region merging algorithm is an iterative process in which regions are iteratively merged. Each iteration requires three different tasks: 1) the pair of most similar neighboring regions is merged, 2) a new region containing the union of the merged regions is formed, and 3) the algorithm updates the distance between the newly created region with its neighboring regions. Working with hyperspectral data, the definition of a region-merging algorithm is not straightforward [48]. The first difficulty is the high intra-class spectral variability which can be found in a region from the same material. In the case of remote sensing images, this variability is introduced by several factors such as the noise resulting from atmospheric conditions, the sensor influence or the illumination effects. Because of this variability, special care has to be taken in modeling hyperspectral regions (it cannot be assumed that the spectra of pixels belonging to a region are strictly homogeneous) [44]. The second important issue is the definition of a spectral similarity measure to establish the merging order between regions. The main difficulty in defining a spectral similarity measure is that most of the spectral signatures cannot be discriminated broadly along all the wavebands [49]. The reason of this difficulty is the redundancy of the spectral information or equivalently the correlation between consecutive values of the spectral radiance curve. As a result, the definition of a region model and a similarity metric defining a good merging order for the construction of BPT is an open research problem.

On the other hand, it can be noticed that once the BPT representation has been computed, this tree is a generic and scalable image representation. This representation enables many application-dependent processing strategies to select tree nodes to form a specific partition in a robust fashion. Different

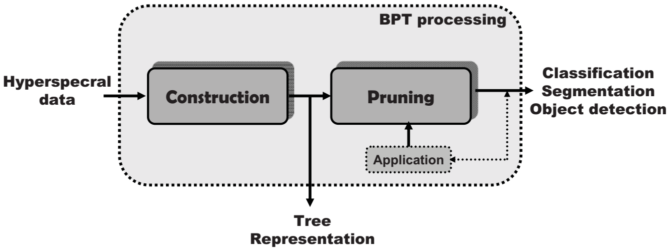
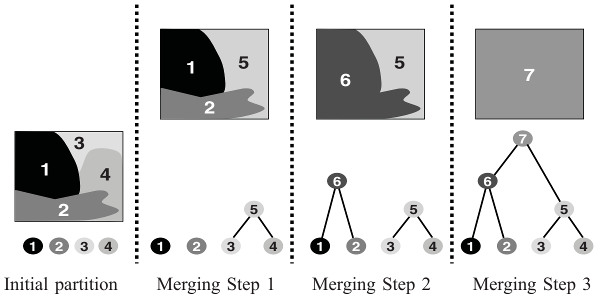


Fig. 2. Diagram of BPT philosophy.



Initial partition Merging Step 1 Merging Step 2 Merging Step 3

Fig. 3. Example of BPT construction using a region-merging algorithm.

processing techniques can be defined using the BPT according to the different applications. The processing of BPT, which is highly application-dependent, generally consists in defining a pruning strategy. This is true for filtering (with connected operators), classification [24], compression, and segmentation [25], [26] or object detection [27].

The goal of a tree pruning is to remove subtrees composed of nodes which are considered to be homogeneous with respect to some criterion of interest (homogeneity criterion, e.g., intensity or texture). Hence, the hyperspectral image processing framework based on BPT relies on two steps illustrated in Fig. 2. The first one corresponds to the construction of the BPT in the case of hyperspectral data, enabling the exploitation of the spectral/spatial correlation. The second corresponds to an application of a pruning strategy which is completely linked to a specific application.

This paper introduces the construction and the processing of BPT representation for the case of hyperspectral images. Firstly, the construction of a robust region-merging algorithm for hyperspectral data is studied. The work presented here investigates and analyzes various region models and similarity metrics defining different merging orders for the BPT construction. Besides BPT construction, an example of BPT processing is also presented here dealing with classification. The organization of this paper is as follows. Section II briefly introduces the BPT and focuses on its construction. The BPT pruning for classification is discussed in Section III. Experimental results are shown in Section IV. Finally, conclusions are drawn in Section V.

II. BPT Construction

Binary Partition Tree (BPT) is a hierarchical representation of a set of regions obtained from an initial partition. Note that the regions of this initial partition may correspond to individual pixels. If the initial partition involves n regions, a BPT generates a tree structure containing 2n − 1 nodes. The BPT should be created in such a way that the most interesting or useful regions of the images are represented by nodes. A possible solution, suitable for a large number of cases, is to create the tree by the execution of a region-merging algorithm [50]. In a bottom-up strategy starting from the leaves, the tree construction is then performed by keeping track of the merging steps. Following an iterative region-merging algorithm, the most similar adjacent regions are merged at each iteration. Fig. 3 shows an example of Binary Partition Tree construction.

In this last figure, tree leaves corresponds to the regions belonging to the initial partition. However, in our BPT construction, each leaf of the tree corresponds to an individual pixel of the original image. The creation of BPT relies on two important notions. The first one is the *region model* M_{R_i} which specifies how regions are represented and how to model the union of two regions. The second notion is the *merging criterion* $O(R_i, R_j)$, which defines the similarity of neighboring regions and hence determines the order in which regions are going to be merged. Therefore, the challenge related to the construction of BPT to represent an hyperspectral image are the definition of a *region model* (to model a set of spectra) and the definition of a *merging criterion* (to measure the similarity between two spectrum data sets).

Past hyperspectral works focusing on spectral classification and hierarchical segmentation have presented different strategies to model a set of spectra [19], [20], [29]. The most popular solution to describe a set of spectra is the first-order parametric model, that is the mean spectrum. As detailed in the following, the key of this model popularity is its simplicity which allows simple definitions of merging orders. However, this model can have an important drawback because it assumes the spatial homogeneity inside the region. In order to solve this problem, parametric models for hyperspectral data have also been studied for some approaches [17], [30]. In this case, the strategy is to model regions by a gaussian probability density function by estimating its mean and covariance matrix. This model presents two important drawbacks: 1) the estimation of the covariance matrix is not easy, in particular for small regions, and 2) this model, as in the case of first order model, is also unimodal.

In this paper, different region models and similarity metrics to construct a robust hyperspectral BPT are studied. The study can be roughly split in two important categories depending on the type of region models. Firstly, the classical first-order parametric model is studied. Then, besides the first-order parametric model, a non parametric statistical region model is also studied in the following sections [31]. This non parametric statistical region model is proposed in order to avoid making any assumption as homogeneity or gaussian probability distribution inside the regions.

A. Region Model

1) First-Order Parametric Model: Given a hyperspectral region R formed by N_{R_p} spectra containing N_z different radiance values, the first-order parametric model M_R is defined as a vector with N_z components which corresponds to the

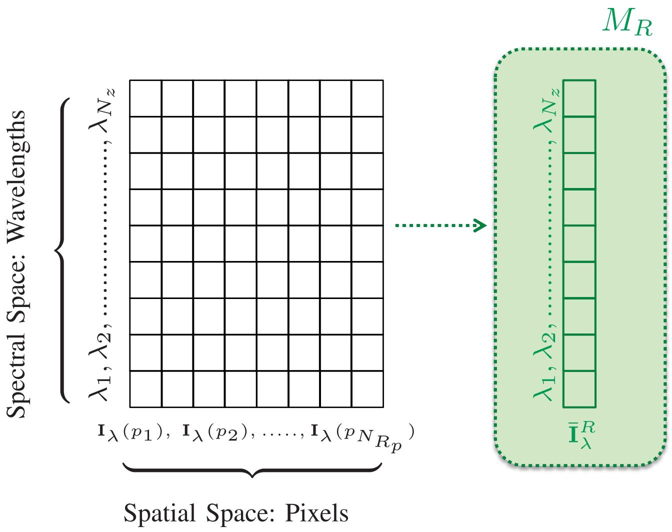

M_R
Spectral Space: Wavelengths
$\lambda_1, \lambda_2, \ldots\ldots, \lambda_{N_z}$
$\mathbf{I}_\lambda(p_1), \mathbf{I}_\lambda(p_2), \ldots\ldots, \mathbf{I}_\lambda(p_{N_{R_p}})$
$\lambda_1, \lambda_2, \ldots\ldots, \lambda_{N_z}$
$\bar{\mathbf{I}}^R_\lambda$
Spatial Space: Pixels

Fig. 4. First-order parametric model.

average of the values of all spectra $p \in R$ in each band λ_i.

$$M_R(\lambda_i) = \bar{I}^R_{\lambda_i} = \frac{1}{N_{R_p}} \sum_{j \leq N_{R_p}} I_{\lambda_i}(p_j) \quad i \in \left[1, \ldots, N_z\right]. \tag{1}$$

Fig. 4 shows how this region model is interpreted. The grid on the left represents the set of spectra of R. In this grid, the horizontal dimension corresponds to the labels assigned to the pixels of the spatial space whereas the vertical dimension corresponds to the spectral domain for each spectrum. Hence, each cell of the grid $I_{\lambda_i}(p_j)$ represents the radiance value in the wavelength λ_i of the pixel whose spatial coordinates are p_j. In this same figure, the green square on the right illustrates the model M_R corresponding to the vector $\bar{\mathbf{I}}^R_\lambda$ which contains in each position $\bar{I}^R_{\lambda_i}$ the mean radiance values for each wavelength on the region.

Note that the M_R model can be considered as a random variable in the λ dimension. The probability distribution of such variable $P_R(\lambda)$ can be easily estimated by applying the spectrum normalization of Eq. 2 in each λ_i

$$P_R(\lambda_i) = \frac{\bar{I}^R_{\lambda_i}}{\sum_{t=1}^{N_z} \bar{I}^R_{\lambda_t}} \tag{2}$$

where $\bar{I}^R_{\lambda_i}$ corresponds to the average of the values of all spectra $p \in R$ in each band λ_i.

Using the spectral distribution $P_R(\lambda)$, a classical spectral similarity measure taking into account the overall shape of the reflectance curves can be proposed as $O(R_i, R_j)$. The use of the Spectral Information Divergence [32] is analyzed here since it can characterize spectral similarity and variability more effectively than other measures [49].

2) Non Parametric Statistical Model: This region model is directly estimated from the pixels of the region where neither spectral nor texture homogeneity are assumed [31]. To formally tackle this idea, this M_R supposes that a region formed by a set of connected pixels is a realization of statistical variables which can be characterized by the corresponding discrete estimated probability distribution. In fact, considering region pixels as a set of independent samples, their common statistical distribution can be directly estimated. Therefore, the region model is the probability density function representing the pixels of the region. In other words, this region model corresponds to the normalized histogram of the pixel values belonging to each region.

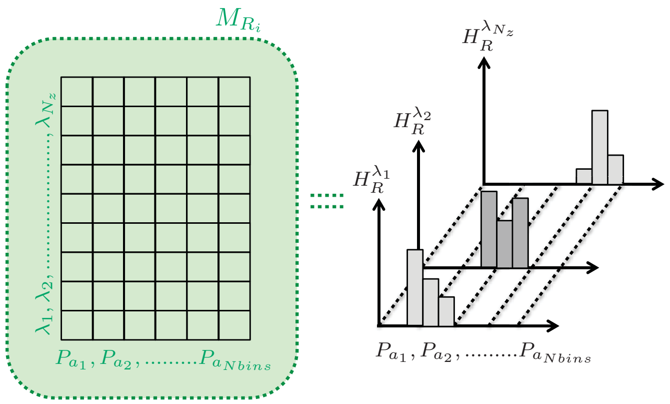

M_{R_i}
$\lambda_1, \lambda_2, \ldots\ldots, \lambda_{N_z}$
$P_{a_1}, P_{a_2}, \ldots\ldots P_{a_{Nbins}}$
$H_R^{\lambda_{N_z}}$
$H_R^{\lambda_2}$
$H_R^{\lambda_1}$
$P_{a_1}, P_{a_2}, \ldots\ldots P_{a_{Nbins}}$

Fig. 5. Nonparametric statistical model.

Consequently, the region model is then represented by a set of N_z non parametric probability density functions (pdfs), one for each band $H_R^{\lambda_j}$, with no assumptions about the nature of the regions nor the shape of the pdfs.

$$M_R = \{H_R^{\lambda_1}, H_R^{\lambda_2}, \ldots, H_R^{\lambda_{N_z}}\}. \tag{3}$$

Fig. 5 shows the non parametric statistical model interpretation. It is observed how M_R is a matrix where each cell represents the probability of the region pixels to have a radiance value a_s in a specific band λ_k. The region model is then formed by the set of the rows $H_R^{\lambda_k}$, each one corresponding to the empirical spatial distribution (histogram) of the region R in the band λ_k. As Fig. 5 shows each $H_R^{\lambda_k}$ is coded by N_{Bins} bins.

For regions made of individual pixels, the histogram of each band is a unit impulse as only one instance of pixel is available. However, the pdf of individual pixels can be more precisely estimated by exploiting the self-similarity present in the image [33]. The key assumption behind the pixel pdf estimation consists in considering that the image is locally a general stationary random process and that it is possible to find many similar patches in an image. Let be p a pixel of the image I and $P(p)$ the square neighborhood patch centered at p defined by the dimensions $Wx \times Wy$. The probability distribution H_R^λ of each individual pixel p given its neighborhood $P(p)$ can be estimated by looking for the similar patches centered at different p_y pixels in a search window. It is assumed that the probability distribution of p depends only on the values of the pixels in $P(p)$ and it is independent of the rest of the image (markovian model). This patch similarity is interpreted as a weight $w(p, p_y)$ which is considered as an additive contribution to the probability of the pixel p of having the value of $\mathbf{I}_\lambda(p_y)$. Computing all $w(p, p_y)$ associated to the pixel p for all the possible pixels p_y in a search window $\mathbf{\Omega}$, the function $w(p, p_y)$ can be used to estimate the probability density function for the individual pixel p.

The computation of the weight $w(p, p_y)$ associated to all the pixels p_y in Ω in the hyperspectral context is proposed in this work as

$$w(p, p_y) = \frac{1}{Z(p)} e^{-\sum_{t=1}^{N_z} \frac{d(P_{\lambda_t}(p), P_{\lambda_t}(p_y))}{h_{\lambda_t}^2}} \tag{4}$$

where $P(p_y)$ is one possible similar patch centered in p_y in the search window Ω. The $Z(p)$ is the normalizing factor to assure $\sum_{\forall p_y \in \Omega} w(p, p_y) = 1$. It is given by

$$Z(p) = \sum_{\forall p_y \in \Omega} e^{-\sum_{t=1}^{N_z} \frac{d(P_{\lambda_t}(p), P_{\lambda_t}(p_y))}{h_{\lambda_t}^2}}. \tag{5}$$

Concernig the similarity between the pixel values of a patch centered at p and a patch centered at p_y, it is computed by using the following expression:

$$d(P_{\lambda_t}(p), P_{\lambda_t}(p_y)) = \sum_{b_y=-W_y}^{W_y} \sum_{b_x=-W_x}^{W_x} \frac{((I_{\lambda_t}(p + b_x + b_y) - I_{\lambda_t}((p_y + b_x + b_y))^2}{(2 * d_p + 1)^2} \tag{6}$$

where $d_p = \sqrt{b_x^2 + b_y^2}$ is the local displacement on the patch regarding the central pixel.

The smoothing parameter h_{λ_t}, which stands for the typical distance between similar patches, controls for each λ_t the decay of the function w. This parameter h_{λ_t} depends on the standard deviation of the noise of the image band I_{λ_t}. The standard deviation for each hyperspectral band can be automatically estimated by calculating the pseudo-residuals of each pixel p as described in [34].

B. Merging Criterion

Different merging criteria are proposed according to the previous region models. On the one hand, the Spectral Information Divergence is proposed for the first-order parametric model. On the other hand, following the statistical analysis, three different similarity metrics between histograms are proposed as merging criterion. Battacharyya Coefficient, Diffusion Distance, and Association Measure via Multidimensional Scaling, respectively.

1) Spectral Information Divergence: The Spectral Information Divergence computes the probabilistic discrepancy between two corresponding spectral signatures modelled by $P_{R_i}(\lambda)$ and $P_{R_j}(\lambda)$. Then, this measure can be proposed to define the merging criterion defined by

$$O_{SID}(R_i, R_j) = \underset{R_i, R_j}{\operatorname{argmin}} \left\{ D(R_i, R_j) + D(R_j, R_i) \right\} \tag{7}$$

with $D(R_i, R_j)$ the Kullback Leibler divergence between two probability distributions

$$D(R_i, R_j) = \sum_{k=1}^{N_z} P_{R_i(\lambda_k)} \log \frac{P_{R_i(\lambda_k)}}{P_{R_j(\lambda_k)}}. \tag{8}$$

2) Battacharyya Coefficient: The bin-to-bin Bhattacharyya distance between two statistical discrete distributions measures the amount of overlap between them. Given two adjacent regions R_i and R_j, modeled by their non parametric statistical region models, the Battacharyya distance at band λ_k between the distributions $H_{R_i}^{\lambda_k}$ and $H_{R_j}^{\lambda_k}$ is defined by

$$BC(H_{R_i}^{\lambda_k}, H_{R_j}^{\lambda_k}) = -\log \left(\sum_{s=1}^{N_{Bins}} H_{R_i}^{\lambda_k}(a_s)^{\frac{1}{2}} H_{R_j}^{\lambda_k}(a_s)^{\frac{1}{2}} \right) \tag{9}$$

where N_{Bins} are the number of bins used to quantify the images intensities. Therefore, the merging criterion O_{BAT} can be defined by

$$O_{BAT} = \underset{R_i, R_j}{\operatorname{argmin}} \sum_{k=1}^{N_z} BC\left(H_{R_i}^{\lambda_k}, H_{R_j}^{\lambda_k}\right). \tag{10}$$

It can be observed that this merging criterion assumes that the histograms are already aligned. To address this weakness, a cross-bin measure between probability distribution is proposed in order to be less sensitive to quantization, noise effect and histogram misalignment. The second similarity measure is called diffusion distance [47].

3) Diffusion Distance: The diffusion distance D_K is a cross-bin distance defined to measure the similarity between two discrete probability distributions, which may overlap or not. The main idea of this distance is to measure the difference between two histograms at various resolution scales through a diffusion process. If the histograms are different, the difference between them will exist at several scales.

The diffusion process is computed by convolving the histogram difference $d_l(a_s)$ with a Gaussian filter $\phi_{\sigma_G(a_s)}$, where $a_s \in \mathbb{R}^m$ is a vector. Thus, each diffusion scale l is computed by a convolution and a downsampling step as

$$d_0(a_s) = H_{R_i}^{\lambda_k}(a_s) - H_{R_j}^{\lambda_k}(a_s) \tag{11}$$

$$d_l(a_s) = \left[d_{l-1}(a_s) * \phi_{\sigma_G(a_s)} \right] \downarrow_2 \quad l \in [1, \ldots, L]. \tag{12}$$

The notation $\downarrow_2$ denotes downsampling by a factor of two. L is the number of pyramid layers and σ_G is the constant standard deviation for the Gaussian filter ϕ. From the Gaussian pyramid constructed by Eq. 12, a distance D_K between the histograms can be computed summing up the L1 norms of the various levels

$$D_K\left(H_{R_i}^{\lambda_k}, H_{R_j}^{\lambda_k}\right) = \sum_{l=0}^{L} \sum_{s=1}^{N_B} \left| d_l(a_s) \right|. \tag{13}$$

Consequently, the proposed merging criterion using the diffusion distance defined through the equations is derived as

$$O_{DIF} = \underset{R_i, R_j}{\operatorname{argmin}} \sum_{k=1}^{N_z} D_K\left(H_{R_i}^{\lambda_k}, H_{R_j}^{\lambda_k}\right). \tag{14}$$

Before concluding on merging criteria using classical histogram distances, it should be remembered that hyperspectral bands are processed separately by the last two criteria: O_{BAT} and O_{DIF}. As a result, the correlation between bands is not taken into account in these merging criteria. In order to

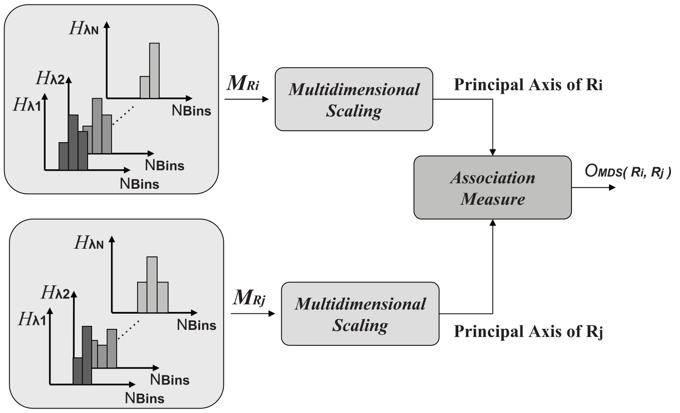


Fig. 6. Methodology for similarity measure via multidimensional scaling.

improve this limitation, a new merging criterion is defined in the following. This criterion tries to exploit the distances between wavebands to remove redundant information contained in each region model. This last studied similarity criterion consists in a new similarity measure based on distances between observations and canonical correlations [35].

4) Association Measure Via Multidimensional Scaling: The merging criterion is divided in two steps summarized in Fig. 6. The first step, corresponds to a local dimensionality reduction by analyzing the inter-waveband similarity relationships for each data set M_R. The purpose of this stage is to remove the redundant hyperspectral information via multidimensional metric scaling (MDS), [43], [45]. As a result, the principal components of the regions containing the most relevant information are obtained. Afterwards, a similarity measure correlating the principal axis of both data sets obtained via multidimensional scaling is performed. This similarity measure, relies on a statistical test based on the multivariate analysis of variance (MANOVA) [41], [46]. The goal is to test whether there is a dependence (or correlation) between the principal components of the regions or not.

The objective of Multidimensional scaling (MDS) [36] transformation is to provide a lower-dimensional data where the dissimilarities between the data points of the multidimensional domain correspond to the dissimilarities of the lower-dimensional domain. In our case, MDS attempts to reduce the dimension formed by the N_z probability distributions of each M_R. To perform it, the probability distribution similarities (or dissimilarities) of M_R can be represented by a $N_z \times N_z$ distance matrix $\Delta_R = (\delta_{kl})$, where $\delta_{kl} = \delta_{lk} \geq 0$ is the diffusion distance value computed by Eq. 13. Hence, being A the matrix with entries $A = -(\frac{1}{2})\delta_{kl}^2$ and the centering matrix $H = I_N - \frac{1}{N}11'$, the so-called inner product matrix B_R associated to Δ_R can be computed by $B_R = HAH$ for each M_R [36]. The inner product matrix B_R is an $N_z \times N_z$ symmetric matrix which can be spectrally decomposed as $B_R = U_R\Lambda_R^2 U_R'$. Assuming the eigenvalues in Λ_R are arranged in descending order, the matrices $U_R\Lambda_R$ and U_R contain the principal and the standard coordinates of region R, respectively. The aim of MDS is then achieved by taking the D_s first most representative principal or standard coordinates of each M_R.

Given two regions defined by M_{R_i} and M_{R_j}, our interest is to measure the similarity between their D_s first standard coordinates. Therefore, two distance matrices Δ_{R_i} and Δ_{R_j} to find $B_{R_i} = U_{R_i}\Lambda_{R_i}^2 U_{R_i}'$ and $B_{R_j} = U_{R_j}\Lambda_{R_j}^2 U_{R_j}'$ should be computed using the explained procedure.

The number D_s of dimensions is an important aspect in most multivariate analysis methods. In MDS, the number of dimensions is based on the percentage of variability accounted for by the first dimensions. Here, a criterion extending the one proposed in [35] is used to set the value of D_s. Firstly, a number of dimensions N_s suggested by the data should be fixed. Then, being u_i and v_i, $i = 1, \ldots, N_s$, the first N_s columns of U_{R_i} and U_{R_j}, a sequence C_k is defined as

$$C_k = \frac{\sum_{t=1}^{k}\sum_{p=1}^{k}\lambda_{tR_i}^2(u_t'v_p)^2\lambda_{tR_j}^2}{\sum_{t=1}^{N_s}\sum_{p=1}^{N_s}\lambda_{tR_i}^2(u_t'v_p)^2\lambda_{tR_j}^2} \qquad k \in [1, \ldots, N_s] \quad (15)$$

where $\lambda_{tR_i}^2$ $\lambda_{tR_j}^2$ are the eigenvalues of B_{R_i} and B_{R_j} which are proportional to the variances of the corresponding principal axes. Here N_s is the minimum dimension for which $\sum_{t=1}^{N_s}\lambda_{tR}^2/\sum_{t=1}^{N}\lambda_{tR}^2 \approx 1$ and $(u_t'v_p)^2$ is the correlation coefficient between the t-th and p-th coordinates. Thus the numerator in C_k is a weighted average of the relationships between principal axes. Clearly $0 \leq C_1 \leq, \cdots \leq C_{D_s} \leq, \cdots \leq C_{N_s} = 1$. The dimension D_s is then chosen such that C_{D_s} is high, for instance $C_{D_s} = 0.9$.

At this point, having two regions defined by their principal coordinates ($U_{R_i}\Lambda_{R_i}$ and $U_{R_j}\Lambda_{R_j}$), a statistical test to measure the similarity between the regions is defined by interpreting the D_s columns of $U_{R_i}\Lambda_{R_i}$ and $U_{R_j}\Lambda_{R_j}$ as a predictor X and a response variable Y of a multivariate linear regression model.

Given a predictor X and a response variable Y, their multivariate linear regression model is defined by

$$Y = X\beta + e \quad (16)$$

where β is the matrix of parameters containing the regression coefficients and e is a matrix of errors. The least-squares estimation of $\hat{\beta}$ is given by $\hat{\beta} = (X'X)^{-1}X'Y$ and the prediction matrix is $\hat{Y} = X\hat{\beta} = PY$ where $P = (X'X)^{-1}X$ is the hat matrix [37]. Clearly, if there is no relationship between X and Y, the matrix β is equal to 0. Considering this, the idea is to perform a test verifying the hypothesis $\beta = 0$ to measure if a significant relationship between X and Y exists. Here, the likelihood ratio test W (or Wilks' lamba) is proposed to measure if the hypothesis $\beta = 0$ is true or false through Eq. 17. This measure has been proposed following the study presented in [38].

Being in our case $Y = U_{R_j}\Lambda_{R_j}$ and $X = U_{R_i}\Lambda_{R_i}$, the predicted model corresponds to $\hat{Y} = PY = U_{R_i}U_{R_i}'U_{R_j}\Lambda_{R_j}$. Eq. 17 is then defined by using $E = \Lambda_{R_j}(I - U_{R_j}'U_{R_i}U_{R_i}'U_{R_j})$ Λ_{R_j} and $E + H = Y'Y = \Lambda_{R_j}U_{R_j}'U_{R_j}\Lambda_{R_j} = \Lambda_{R_j}^2$. These two last equations define the Wilks' lambda test $W(R_i, R_j)$ as

$$W(R_i, R_j) = \frac{det(E)}{det(E + H)} = det(I - U_{R_j}'U_{R_i}U_{R_i}'U_{R_j}). \quad (17)$$

The Wilks' criterion of Eq. 17 can also be defined by $W(R_i, R_j) = \lambda_w^1 \times \lambda_w^2 \cdots \times \lambda_w^{D_s}$ where λ_w^i are the eigenvalues corresponding to

$$det(E - \lambda_w(E + H)) = 0. \tag{18}$$

Being these values $0 \leq \lambda_w^i \leq 1$, the squared canonical correlation r_i^2 is defined by $1 - \lambda_w^i$. The Wilks' criterion can thus be expressed in terms of canonical correlations as

$$W(R_i, R_j) = \prod_{i=1}^{D_s}(1 - r_i^2). \tag{19}$$

This last equation satisfies that $0 \leq W(R_i, R_j) \leq 1$ and $W(R_i, R_j) = 1$ if R_i is equal to R_j. Thus, this leads to the definition of the proposed merging criterion

$$O_{MDS} = \underset{R_i, R_j}{\operatorname{argmin}}\, W(R_i, R_j). \tag{20}$$

To conclude this section, it should be remarked that the area of the regions is not included in any proposed merging order. Thus, these approaches may suffer from small and meaningless regions into the generated partition. To overcome this limitation, the fusion between small regions in the first merging levels has been set as a priority. The approach consists in forcing the merging of regions having an area smaller than a given percentage (typically 15%) of the average size of the regions created by the merging process [31].

III. Pruning Strategy for Hyperspectral Image Classification

The processing of a BPT representation can be understood as the extraction of a set of nonoverlapping regions coded in BPT nodes according to a specific criterion. This analysis of the tree can be performed by a pruning strategy aiming at removing redundant subtrees from the original tree. A subtree is redundant if all its nodes can be considered homogeneous with respect to some criterion of interest (e.g., homogeneity of intensity or of texture). This task can be performed by analyzing a pruning criterion along the tree branches to retrieve the nodes of largest area fulfilling the criterion. Three different results of BPT pruning are shown in Fig. 7. Note that using the pruned tree, a partition composed of N_R regions can be easily constructed by selecting the leaf nodes of the resulting pruned tree.

The tree analysis may follow a bottom-up or a top-down strategy. The pruning strategy discussed in this section corresponds to a bottom-up analysis of the BPT. The task consists in evaluating some regions (or nodes) criterion by performing an analysis running from the leaves to the tree root. In practice, this particular cost directly depends on the considered application. Classification, filtering, object detection and segmentation are different examples of applications. In this section, as an illustrative example, a classification application is discussed. Other examples of applications involving object detection and segmentation can be found in [25], [26], and [27]. In this work, the goal of this pruning is to remove subtrees composed of nodes belonging to the same class and to construct a classification map. To address it, the analysis of the tree consists of two important steps. The first one is the BPT population which computes and assigns specific region descriptors to each node of the tree structure. The second step is the pruning decision whose task is to evaluate a cost function ϕ_R associated to the region descriptors and eventually to decide where to prune the tree.

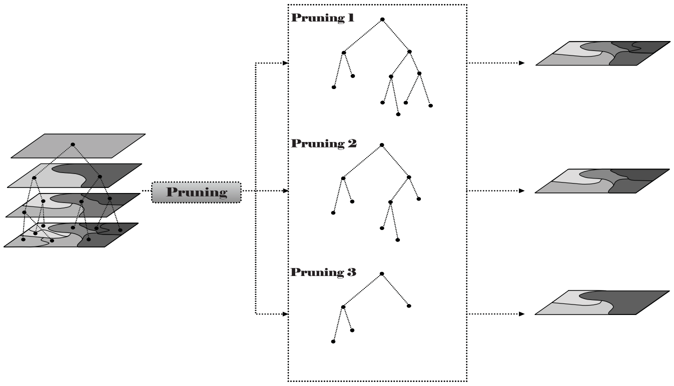


Fig. 7. Pruning examples.

A. BPT Population

Given the classification aim, the main information, or node descriptor, to be used to define the pruning involves the class probability distribution $\mathcal{P}_R$. This probability distribution is a vector containing the probabilities that the node belongs to each class C_i. The resulting class distribution is denoted by $\mathcal{P}_{R_i}(\mathcal{C}_j)$.

The task of node population can be easily achieved in a supervised way by using a multiclass classifier. Here, Support Vector Machine is used as an example of probabilistic classifiers which have proved to be well suited to classify hyperspectral data [5], [6], [1]. The standard Gaussian kernel is chosen in this work since it is one of the most used kernels in hyperspectral data.

Being SVMs supervised [28], the kernel parameters should be first computed by a training step. In our case, this step follows the classical crossvalidation strategy: the training set is divided into k parts, then the SVM is trained using $(k - 1)$ parts and the obtained parameters are tested on the remaining part. The SVM training step is done by using some leaf nodes which correspond to single spectra. The selection of these nodes directly depends on the available ground truth.

Once the kernel function is constructed, it is used to classify all the BPT nodes by assigning to each of them their $\mathcal{P}_R$. In order to classify the data, the kernel function usually uses a spectrum as an input parameter. However, note that in our case, each BPT node represents a region formed by a set of spectra not a single spectrum. For this reason, each BPT node is modeled by its mean spectrum to be able to apply the kernel function on the node.

The information provided by the class probability distribution is used to evaluate the node *Misclassification Rate* ($\mathcal{MR}$). *Misclassification Rate* can be understood as the error of assigning to a node a wrong class. The use of *Misclassification Rates* has been previously studied in binary decision tree prunings [39]. In these decision trees, a reliable classification result of a node R_i implies a likely minimum misclassification rate $\mathcal{MR}(R_i)$ which has been previously mathematically expressed by

$$\mathcal{MR}(R) = 1 - \max_i \mathcal{P}_R(C_i). \qquad (21)$$

The misclassification rate of Eq. 21 can have two important problems in our context. The first problem comes when a node is formed by merging a very large region with a small one. Assume the node R is formed by two sibling nodes R_L and R_R having an area relation such that $\mathcal{A}_{R_L} >> \mathcal{A}_{R_R}$. If R_L belongs to class $\mathcal{C}_i$ and R_R to class $\mathcal{C}_j$, the union of both regions will belong to $\mathcal{C}_i$ since the region contained in R_L is much larger than R_R. Thus, the reliability of the SVM classifier for the node R will not significantly change even if both regions belong to two different classes.

The second important problem of Eq. 21 is the presence of mixed pixels in the image forming mixed regions. The mixed pixels in hyperspectral context are spectra which are formed with some materials involving different ground truth classes. Consequently, pixels belonging to these regions do not have a high probability of belonging to any given class. As a result, an important misclassification rate can appear for this type of regions. In order to solve these problems, the misclassification rate of Eq. 21 for nonleaf nodes has been modified as follows:

$$\mathcal{MR}(R) = 1 - BC(\mathcal{P}_{R_R}, P_{R_L}) \qquad (22)$$

where $BC(\mathcal{P}_{R_R}, \mathcal{P}_{R_L})$ is the Battacharryya coefficient between the probability class distributions of the left and the right children of R. With N_c different ground truth classes, the Battacharrya coefficient in this classification context is described by

$$BC(\mathcal{P}_{R_R}, \mathcal{P}_{R_L}) = \sum_{i=1}^{N_c} \mathcal{P}_{R_R}(\mathcal{C}_i)\mathcal{P}_{R_L}(\mathcal{C}_i). \qquad (23)$$

Last expression cannot be used for leaf nodes as they have no children. Hence, two types of misclassification rates are used: 1) eq. 21 is the misclassification rate used for BPT leaves, and 2) eq. 22 is the one used for nonleaf nodes. The use of eq. 22 solves the weak area relation problem allowing to detect if two reliable but different regions are going to be merged in an unique node. However, as this last equation is sensitive to small regions, a node formed by a very small wrongly classified region (for instance 1 pixel) can give a high $\mathcal{MR}(R)$. Thus, this weakness should be solved setting that if a node has a very small area (for instance smaller than 3), its parent will have a very small $\mathcal{MR}(R)$. In other words, very small regions cannot cut BPT branches since they are not considered as reliable.

B. Pruning Decision

The pruning of a sub-tree $\mathcal{T}_s$ hanging from a node R consists in deciding if all its descendants, can be replaced by R. This is done by the function ϕ_R which compares the misclassification rate at node R with the misclassification rate corresponding to the set of leaf nodes of the sub-tree $\mathcal{T}_s$. Fig. 8 shows an example of the concepts presented in the evaluation of a nonleaf node R. In this example, the misclassification rate associated with the node R_i should be compared with the error associated to the 3 leaves $R_i^{leaves} = \{l_1, l_2, l_3\}$ contained in $\mathcal{T}_s$.

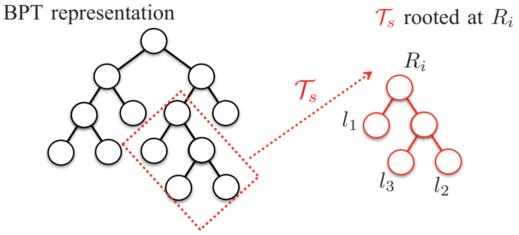


Fig. 8. Subtree definition.

Mathematically, the function defining the pruning function ϕ_R is given by

$$\phi_R(R_i) = \mathcal{MR}(R_i) - \overline{\mathcal{MR}(R_i^{leaves})} \qquad (24)$$

where $\overline{\mathcal{MR}(R_i^{leaves})}$ represents the average misclassification rates of the leaves of the subtree rooted at of R_i. The aim is to detect when ϕ_R is higher than an allowed threshold α. Considering a node R_i, if the cost function $\phi_R(R_i) < \alpha$, the subtree hanging from R_i can be pruned and replaced by R_i. Contrarily, if $\phi_R(R_i) > \alpha$, the node R_i cannot be a leaf in the pruned BPT. Note that the α value determines the size of the pruned BPT. When α is small, the penalty term is small, so the size of the pruned tree will be large. Contrarily, as α increases, the pruned BPT has fewer and fewer nodes.

IV. Experimental Results

In this section, a complete evaluation of the BPT-based representation is provided. Firstly, experiments have been performed to evaluate the different merging order criteria proposed in Section II. To this goal, some partitions obtained during the construction of the BPT following the merging sequence are compared between themselves and also with the RHSEG technique [20], which is the reference hierarchical representation and segmentation tool for hyperspectral data. Secondly, experiments are conducted to evaluate the pruning technique described in Section III. In this context, two different data sets are used. Firstly, an AVIRIS hyperspectral image is used to study how the construction of BPT affects the pruning results in our classification context. Secondly, the last experiment is devoted to compare the classification results obtained by pruning the BPT with the spectral-spatial classification approach [40].

A. Evaluation of the BPT Construction

This experiment evaluates the partitions that are obtained following the merging sequence involved in the BPT construction. Note that, appropriate pruning techniques can produce a much larger set of partitions, but the partitions involved in the tree creation allows us to make an evaluation of the region models and the associated merging criteria proposed in Section II. The initial partition is composed of regions formed by individual pixels and, therefore, involves N_p regions. In order to get a partition with N_R regions, a number of $N_p - N_R$

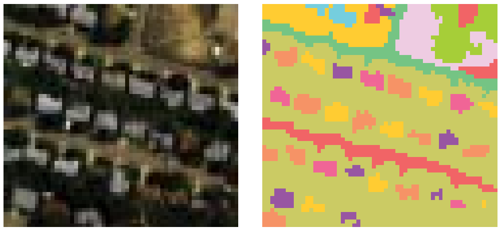


Fig. 9. Urban Hydice data set. (a) False-color composition. (b) Ground truth.

merging steps have to be performed. The quality of the partitions having N_R regions is then evaluated using two different partition distances. Both measures have been defined in the context of image segmentation and previously used in [31].

The first quality measure is the asymmetric partition distance d_{asym} which is ranging between 0 and 1. This distance measures the minimum number of pixels whose labels should be changed so that partition P_1 becomes finer than partition P_2, normalized by the total number of pixels of the image minus one. If P_1 is the ground truth and P_2 is the computed partition, $d_{asym}(P_1, P_2)$ measures the undersegmentation and $d_{asym}(P_2, P_1)$ the oversegmentation. In this work, an average asymmetric distance has been used: $d^T_{asym} = (d_{asym}(P_1, P_2) + d_{asym}(P_2, P_1))/2$.

The second used partition distance corresponds to the symmetric distance d_{sym} which is used as a global error measure to establish a compromise between under- and oversegmentation error since it is measured between partitions with equal number of regions. The distance is defined as the minimum number of pixels whose labels should be changed in P_1 to achieve a perfect matching with P_2 (P_1 and P_2 become identical), normalized by the total number of pixels of the image minus one. Here, both distances d_{sym} and d^T_{asym} are used to measure the quality of the BPT hierarchical levels obtained by studying two different data sets.

1) Urban HYDICE Data Set: The first experiments have been performed using a portion of a publicly available HYDICE hyperspectral image. After removing water absorption and noisy bands, the data contain 167 spectral bands in a range from 0.4 to 2.5 micrometers. The studied image has 60 × 60 pixels having a spatial resolution of a few meters. Fig. 9(a) shows a false color composition of three of them and Fig. 9(b) features a manually designed segmentation ground truth.

For this image, the BPT is computed by the procedure described in Section II. The number of bins to represent the histograms depends on the image range (here $N_B = 256$). For the multidimensional scaling approach, the number of used components found by the sequence C_k is $D_s = 3$. To visually illustrate these results, some partitions obtained following the merging sequence are shown in Fig. 10. This figure shows the partitions obtained by BPT constructed by different merging orders and the RHSEG algorithm [20]. In the case of RHSEG, the similarity criterion used is SAM [9]. The spectral clustering weight has not been used in this experiment. It should be noticed that the RHSEG algorithm also uses the mean region model as the BPT with the O_{SID} distance.

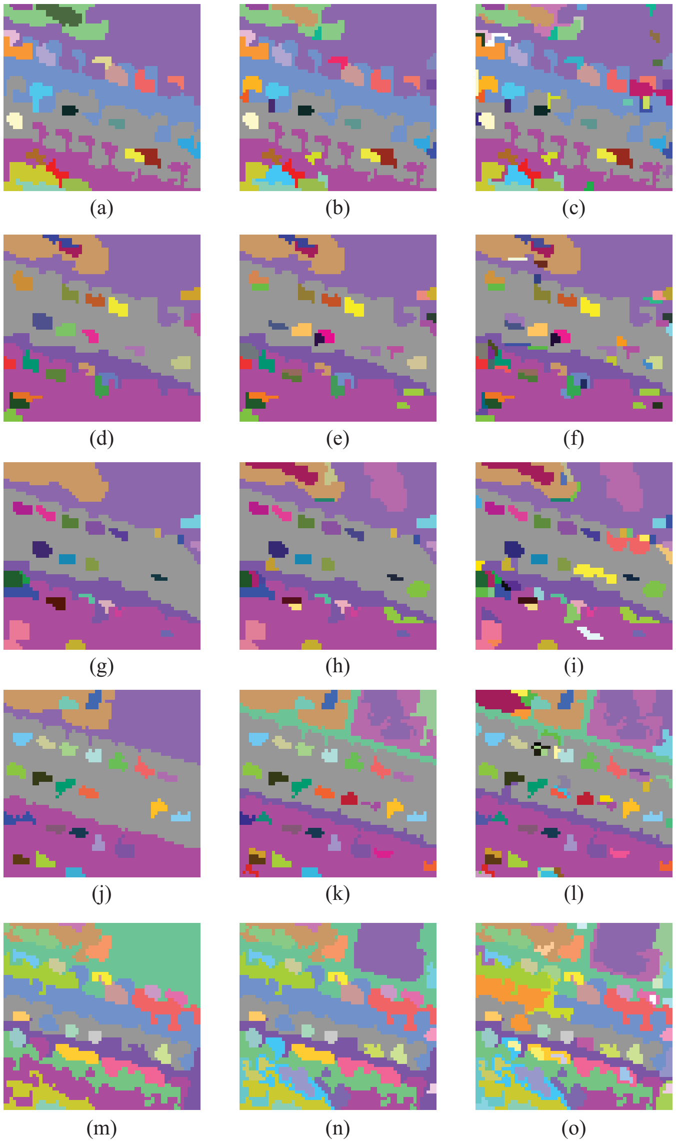
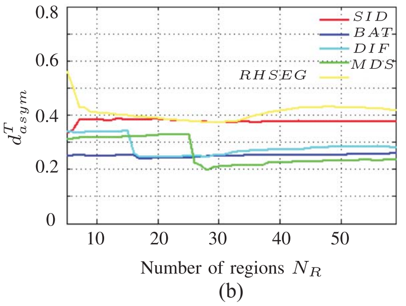


Fig. 10. Visual evaluation of the results over the HYDICE data set. (a) $O_{\mathrm{SID}}, N_R = 27$. (b) $O_{\mathrm{SID}}, N_R = 37$. (c) $O_{\mathrm{SID}}, N_R = 56$. (d) $O_{\mathrm{BAT}}, N_R = 27$. (e) $O_{\mathrm{BAT}}, N_R = 37$. (f) $O_{\mathrm{BAT}}, N_R = 56$. (g) $O_{\mathrm{DIF}}, N_R = 27$. (h) $O_{\mathrm{DIF}}, N_R = 37$. (i) $O_{\mathrm{DIF}}, N_R = 56$ (j) $O_{\mathrm{MDS}}, N_R = 27$. (k) $O_{\mathrm{MDS}}, N_R =37$. (l) $O_{\mathrm{MDS}}, N_R = 56$. (m) RHSEG, $N_R = 27$. (n) RHSEG, $N_R = 37$. (o) RHSEG, $N_R = 56$.

Merging Criterion	d_{sym}
O_{SID}	0.54
O_{BAT}	0.35
O_{DIF}	0.33
O_{MDS}	0.236
$RHSEG$ [31]	0.65

(a)



(b)

Fig. 11. Distance evaluation for Hydice image. (a) Symmetric. (b) Asymmetric.

The first test to evaluate the quality of the BPT construction is done by computing d_{sym} using the manually created ground truth image shown in Fig. 9(b) that contains 37 regions.

TABLE I
CLASS SPECIFIC ACCURACY USING 20% OF TRAINING SAMPLES

Class	Simple SVM	Pruned BPT O_{SID}	Pruned BPT O_{DIF}	Pruned BPT O_{MDS}
1	75.61	**82.93**	**82.93**	80.49
2	83.46	**92.75**	92.19	**92.75**
3	84.35	95.21	**96.81**	96.01
4	76.14	**96.59**	92.05	91.48
5	94.37	95.44	95.17	**95.71**
6	97.15	97.65	**98.40**	97.50
7	**92.31**	88.89	88.89	88.89
8	98.09	**99.73**	99.18	**99.73**
9	90	**100**	85.71	**100**
10	83.06	88.02	**89.94**	88.57
11	91.52	97.24	96.06	**99.73**
12	86.55	92.62	91.76	**94.36**
13	96.22	97.48	98.11	**98.76**
14	95.57	97.53	97.89	**97.94**
15	67.72	84.21	80.35	**97.89**
16	91.67	95.83	93.06	**97.22**
Overall	87.74	93.89	92.40	**94.69**

The d_{sym} between Fig. 9(b) containing 37 regions and the partitions obtained by doing N_p-37 merging steps over the initial partition is computed. This distance is also computed for the partition involving 37 regions obtained by the RHSEG to compare the BPT results with a state of the art technique. Table I shows the values of the symmetric distance between the ground truth of Fig. 9(b) and the partitions obtained by BPT constructed by different merging orders and the RHSEG. It should be noticed that all the results shown in Fig. 11(a) are obtained by partitions involving 37 regions.

Comparing the results of Fig. 11(a), it can be observed that region-merging algorithms using the non parametric statistical region model obtain better results. As this model is more accurate than the traditional mean, the BPTs constructed by using O_{MDS}, O_{DIF} and O_{BAT} achieve smaller d_{sym} values. A small improvement is introduced by O_{DIF} regarding O_{BAT} [31]. This is explained by the fact that the diffusion distance is more robust to histogram misalignment. Comparing all the obtained results, it can be observed that O_{MDS} achieves the best results. Besides relying on a non parametric statistical region model, this distance takes into account correlation between bands. O_{MDS} removes redundant information through multidimensional scaling which allows the introduction of the spectral information inside the merging criterion.

For this image, a second test evaluating the merging orders for BPT construction is carried out using d^T_{asym}. This measure is computed for various partitions having different number of regions N_R. The evolution of d^T_{asym} according to the number of regions is shown in Fig. 11(b).

As it can be seen, the effectiveness of the statistical region model and the good performances of O_{MDS} can be corroborated. For the case of O_{SID} some peaks can be observed in the d^T_{asym} curve. They correspond to the merging of regions without any meaning because of the poor mean region model. The quantitative evaluation can be corroborated by observing Fig. 10. Looking at the second column, the results described by d_{sym} can be corroborated. O_{MDS} with $N_R\ = 37$ corresponds to $d_{sym}\ = 0.236$ which is the best result.

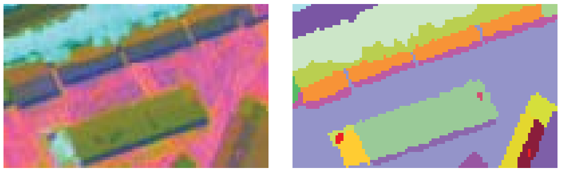

(a) (b)

Fig. 12. Pavia Center ROSIS data set. (a) False-color composition. (b) Ground truth.

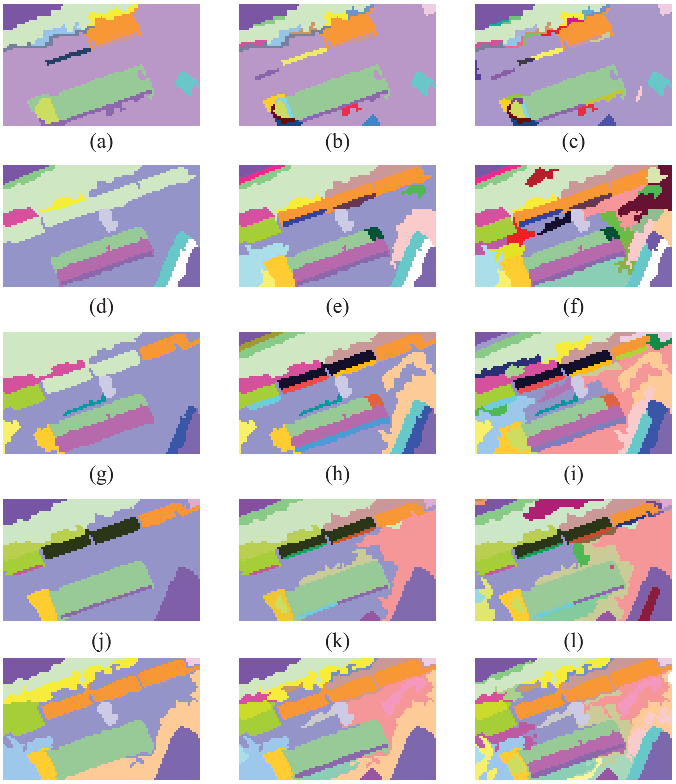

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (o)

Fig. 13. Visual evaluation of the results over the Pavia Center data set. (a) $O_{\mathrm{SID}}, N_R = 13$. (b) $O_{\mathrm{SID}}, N_R = 25$. (c) $O_{\mathrm{SID}}, N_R = 39$. (d) $O_{\mathrm{BAT}}, N_R = 13$. (e) $O_{\mathrm{BAT}}, N_R = 25$. (f) $O_{\mathrm{BAT}}, N_R = 39$. (g) $O_{\mathrm{DIF}}, N_R = 13$. (h) $O_{\mathrm{DIF}}, N_R = 25$. (i) $O_{\mathrm{DIF}}, N_R = 39$. (j) $O_{\mathrm{MDS}}, N_R = 13$. (k) $O_{\mathrm{MDS}}, N_R = 25$. (l) $O_{\mathrm{MDS}}, N_R = 39$. (m) RHSEG, $N_R\ = 13$. (n) RHSEG, $N_R\ = 25$. (o) RHSEG, $N_R\ = 39$.

2) ROSIS Pavia Data Set: A second data set is processed to confirm the previous results. In this case, a portion of the Pavia Center image from the hyperspectral ROSIS sensor is used having a spatial resolution equal to 1.3 m per pixel. These data contain 99 × 60 pixels and 102 spectral bands. Fig. 12(a) shows a false-color composition of three hyperspectral bands while Fig. 12(b) shows the ground truth that has been manually created. Similar experiments have been performed for this image. The number of bins used to represent the histograms is Nbins = 256. The dimension D_s of the multidimensional scaling reduction techniques is 2.

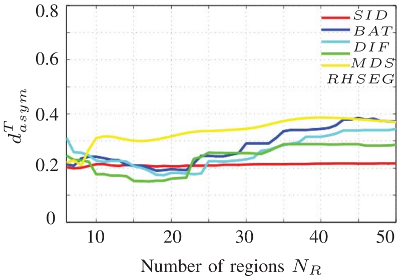
The same evaluation is also carried out for this second data set. The visual evaluation is shown in Fig. 13. Concerning the quantitative evaluation, Table II in Fig. 14(a) shows the symmetric distance values d_{sym} between the ground truth partition and the partitions generated by the proposed methods, both with the same number of regions (equal to 25).

TABLE II
CLASS SPECIFIC ACCURACY FOR PAVIA UNIVERSITY DATA SET

Class	Simple SVM	Spectral-Spatial Approach [24]	Pruned BPT
1	85.93	83.6	**88.84**
2	76.66	**77.9**	71.69
3	70.46	82.9	**91.95**
4	**97.55**	96.7	95.14
5	**99.55**	98.7	98.81
6	91.99	95.2	**97.08**
7	92.48	94.0	**99.02**
8	92.31	95.0	**98.13**
9	99.26	**97.4**	95.99
Overall	88.58	91.26	**92.96**

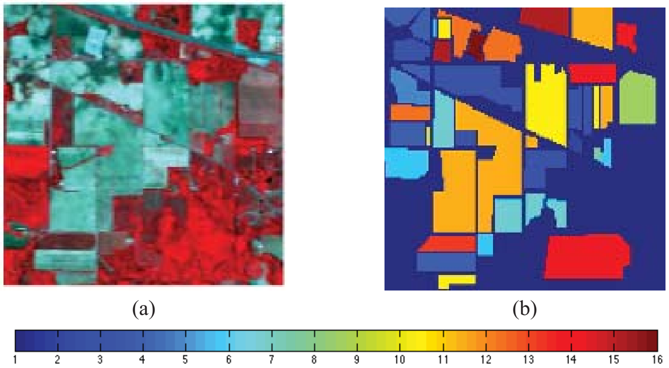
Merging Criterion	d_{sym}
O_{SID}	0.336
O_{BAT}	0.286
O_{DIF}	0.274
O_{MDS}	0.227
$RHSEG$ [31]	0.48

(a)



(b)

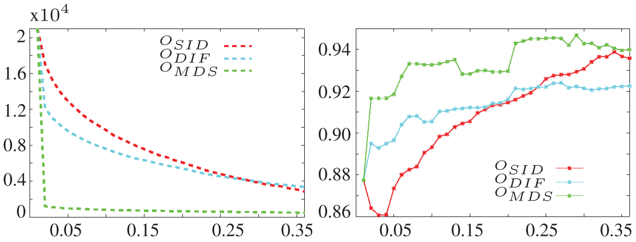
Fig. 14. Distance evaluation for Pavia Center ROSIS image. (a) Symmetric. (b) Asymmetric.



(a)

(b)

Fig. 15. (a) False-color Indian pines composition. (b) Available ground truth image.



(a)

(b)

Fig. 16. Indian pines pruning evaluation. (a) Number of regions versus α. (b) Overall accuracy versus α.

Fig. 14(b) plots the evolution of the average asymmetric distance according to the number of regions. Fig. 14(b) confirms the good performances of the non parametric statistical region model. Also, O_{MDS} obtains the best results for this second data set. The efficiency of O_{MDS} against the other merging criteria can also be verified looking at Fig. 13 and 14.

As can be seen, the results obtained with criteria using the mean region model are similar to those obtained with the non parametric region model but with a larger number of regions. However, it can be seen that when regions become more complex, the simple model becomes less accurate. This explains why the average asymmetric distance value obtained with the nonparametric statistical model starts decreasing when the number of regions gets close to the number of regions of the ground truth. The curve also shows that in some iterations d^T_{asym} is smaller by using the O_{DIF} than O_{MDS}. This is because, at this level of the tree construction, the pixels forming the background have been merged earlier with O_{DIF} than with O_{MDS}.

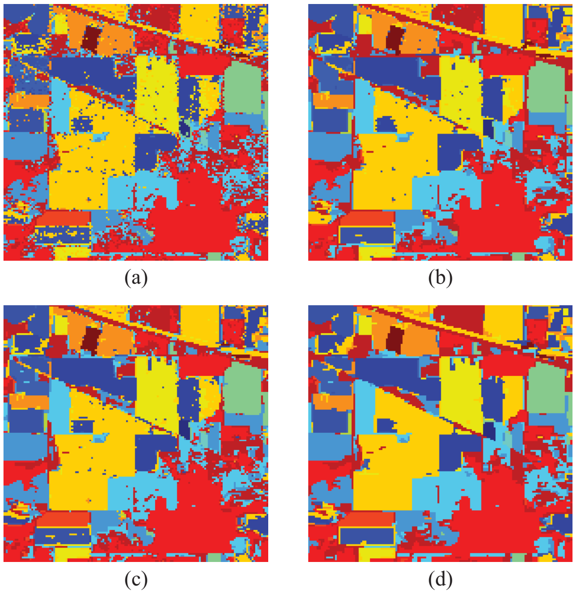
B. Evaluation of the Classification Pruning

1) Aviris Indian Pines: In the first pruning experiment, Indian Pines AVIRIS hyperspectral data containing 200 spectral bands having a spatial dimension of 145 × 145 pixels is used. Fig. 15(a) shows a false-color composition of this data set. In this image, it can be seen how a simple RGB composition of the hyperspectral data does not allow to discriminate between the different materials. The whole image is formed by 16 different classes having an available ground truth as illustrated on Fig. 15(b). For this image, three different BPT are constructed using the following merging criteria O_{SID}, O_{DIF} and O_{MDS}, respectively. In the case of non parametric statistical region model, the histogram quantification is set to $N_{bins} = 150$. Concerning the merging criterion O_{MDS}, the estimated D_s value defining the number of principal components is equal to 3.

Once the three different BPT have been created, the populating BPT strategy described in Sec. III-A is performed. The SVM classifier is trained by selecting randomly 20% of samples for each class from the reference data described in Fig. 15(b). Using the constructed SVM model and the BPT representation, the $\mathcal{P}_R$ probability distributions are assigned to all BPT nodes in order to compute their misclassification rates.

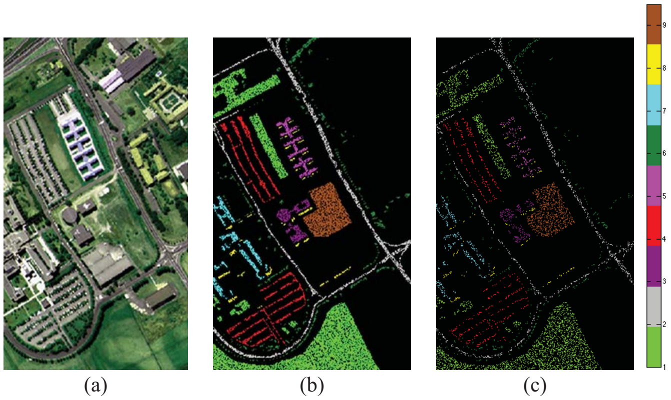
In this example, different α threshold values are used to compare the different classification maps obtained by the three BPTs. Two different evaluations are carried out for different α values ranging from 0 to approximately 0.4. It has been considered that α higher than 0.4 means a high misclassification error. The first evaluation corresponds to the number of BPT leaves obtained after the pruning. This measure gives information about the BPT construction. For a given class accuracy, if a pruning strategy removes more BPT nodes from a tree, this means that the BPT has been better constructed. The second evaluation corresponds to the overall class accuracy obtained by the classification maps achieved by the BPT pruning. Both experiments are shown in Fig. 16.

Fig. 16(b) shows how the highest accuracies are obtained with $\alpha_C \approx 0.30$ where the results obtained by O_{MDS} outperforms the other results for all the α_C values. However, it should be noticed that in some cases, O_{SID} can lead to similar classification accuracies than O_{MDS}. Contrarily, the merging


(a) (b)

(c) (d)

Fig. 17. Obtained classification map using 20% of training samples. (a) Pixel-wise classification. (b) Pruned BPT, O_{SID}, $\alpha = 0.34$. (c) Pruned BPT, O_{SID}, $\alpha = 0.27$. (d) Pruned BPT, O_{MDS}, $\alpha = 0.29$.



(a) (b) (c)

Fig. 18. Pavia University data set. (a) False-color composition. (b) Test data set. (c) Training data set.

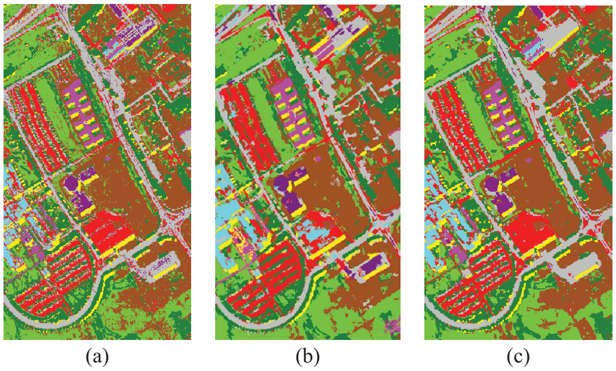

(a) (b) (c)

Fig. 19. Comparison between obtained classification maps. (a) Pixel-wise classification. (b) Spectra-spatial approach. (c) Pruned BPT.

criterion O_{DIF} achieves the worst results. These results can be explained by the fact that the criterion processes separately the different bands and this turns out to be a serious drawback for classification.

Comparing the curves of O_{SID} and O_{MDS} in Fig. 16, it could be said that both criteria have similar performances in a given α_C interval. However, comparing the results on the curves of Fig. 16(a), it can be remarked how O_{MDS} removes more BPT nodes in its pruning. Note that the number of regions corresponding to the pruned BPT leaves, is much lower for O_{MDS} than for the other two merging criteria.

Following this evaluation, the classification maps corresponding to the highest overall accuracy of Eq. 16(b) are shown on Fig. 17. The obtained results are compared with the classical SVM pixel-wise classification of Fig. 17(a). The same training samples are used for all the classification results.

Looking at the BPT pruning results, it can be observed that the classification maps are formed by quite homogeneous regions. In particular, the BPT nodes selection according to the proposed pruning criterion provides a less noisy classification. This can be noticed in the case of Fig. 17(d) corresponding to the O_{MDS} merging criterion. The obtained results also corroborate the BPT performances since extracted nodes reflect semantic real-world objects of the image. It should be remarked that Indian Pines has a high spectral variability due to its low spatial resolution.

The interest of using the BPT structure to obtain the classification map can be summarized as follows: the classification noise observed with pixel-wise approaches is strongly reduced, while no edge noise is introduced (a classical post-processing or a Markovian regularization would suffer from this drawback). The regions do actually map real borders as determined while constructing the tree. In addition, the final partition can contain small (but meaningful) regions as well as large regions. All these regions are selected in different levels of the hierarchy.

According to Fig. 17, Table II illustrates the corresponding class-specific and the global classification accuracies. The best class accuracies are highlighted in bold. Observing these results, the proposed BPT pruning classification improves the classification accuracies for almost all the classes compared to pixel-wise classification. Studying the different merging criteria, O_{MDS} leads to the best results.

2) ROSIS Pavia University: The second data set used to evaluate the classification pruning corresponds to Pavia University from ROSIS sensor. The image is formed by 103 channels and has 610 × 340 pixels. Fig. 18(a) shows a false-color composition for this second data set. For this data set, Fig. 18(b) and Fig. 18(c) illustrate the used test and training data sets, respectively.

This second experiment tries to verify that the classification accuracies obtained by pruning the BPT can be comparable to one of the state of the art recent spectral-spatial classification approach [40]. This classification approach combines two kernel functions to include both the spatial and the spectral classification in the SVM classification process. The spatial information is extracted by a morphological area filtering of size 30. In this experiment, the results obtained by this classical method are compared with the results obtained by pruning a BPT which is constructed by using O_{MDS} as merging criterion. The region model has been defined by using $N_{bins} = 256$ and the number of principal components is equal to 2. Fig. 19 shows the classification maps obtained by the pixel-wise classification on Fig. 19(a), the spectral-spatial approach [40] on Fig. 19(b) and the results obtained after applying BPT pruning on Fig. 19(c). Comparing with the pixel-wise approach, it can be observed that, using the BPT, a better classification map is also obtained for this second data set. BPT pruning improves the classification accuracy preserving most of the edges and shapes. In order to compare the results obtained by Fig. 19(b) and (c), Table II shows the global accuracies. The global accuracy of the proposed approach presented in Fig. 19(c) obtains the best results.

V. Conclusion

In this work, Binary Partition Trees have been proposed as a new representation for hyperspectral images. Obtained through a recursive region-merging algorithm, they can be interpreted as a new region-based and hierarchical representation of the hyperspectral data. The main advantage of BPT is that it can be considered as a generic representation. Hence, it can be constructed once and used for many applications such as segmentation, classification, filtering, object detection, etc. Many tree processing techniques can be formulated as pruning strategies. Concerning the BPT construction, two concepts have been highlighted to define the recursive merging algorithm. The first concept is the use of non parametric statistical region models which efficiently deal with the problems of spectral variability and textures for clustering hyperspectral data. The second one is the use of a new similarity measure called MultiDimensional Scaling (MDS) depending on canonical correlations relating principal coordinates. Note that, in this approach, as in many hyperspectral image processing algorithms, there is a dimension reduction step represented by the number of principal components. However, by contrast to classical approaches, the dimension reduction is not defined and applied globally on the entire image but locally between each pair of regions. It has been demonstrated that BPT enables the extraction of a hierarchically structured set of regions representing well the image. As a first example of BPT processing, we have proposed and illustrated a pruning strategy to classify hyperspectral data. Experimental results obtained from different data sets have shown that the proposed method improves the classification accuracies of a classical SVM and a spectral-spatial approach. Obtained classification maps contain a reduced amount of noise. preserving most of the edges and shapes. Future work will be conducted for the pruning strategy. New global techniques are currently being studied to improve the accuracy and the robustness of the results. We will also develop pruning strategies for different types of applications including object detection and segmentation.

Acknowledgment

The authors would like to thank Prof. Cuadras for his mathematical interest and support of this work. They would also like to thank Dr. Fauvel for his support in performing the spectral-spatial classification comparison.

References

[1] A. Plaza, J. A. Benediktsson, J. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, J. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, no. 1, pp. S110–S122, 2009.

[2] C.-I. Chang *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. Dordrecht, Norwell, MA: Kluwer Acad. Publ., 2003.

[3] D. A. Landgrebe, *Signal Theory Methods Multispectral Remote Sens.* New York: Wiley, 2003.

[4] G. Camps-Valls and L. Bruzzone "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[5] L. Bruzzone, M. Chi, and M. Marconcini "A novel transductive SVM for semisupervised classification remote sensing images," *IEEE Trans. Geosci. Remote Sen.*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006.

[6] M. Chi and L. Bruzzone "Semisupervised classification of hyperspectral images by SVMs optimized in the primal," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1870–1880, Jun. 2007.

[7] 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.

[8] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.

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

[10] Y. Tarabalka, J. C. Tilton, J. A. Benediktsson, and J. Chanussot, "A marker-based approach for the automated selection of a single segmentation from a hierarchical set of image segmentations," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 5, no. 1, pp. 262–272, Jan. 2012.

[11] J. Angulo and S. Velasco-Forero, "Semi-supervised hyperspectral image segmentation using regionalized stochastic watershed," *Proc. SPIE*, vol. 7695, pp. 1–12, May 2010.

[12] A. Farag, R. Mohamed, and A. El-Baz "Unified framework for map estimation in remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1617–1634, Jul. 2005.

[13] J. Li, J. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.

[14] G. Martin and Plaza, A, "Spatial-spectral preprocessing prior to endmember identification and unmixing of remotely sensed hyperspectral data," *IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens.*, vol. 5, no. 2, pp. 380–395, Apr. 2012.

[15] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-France, and J. Calpe-Maravilla "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[16] M. Fauvel, J. Chanussot, and J. A. Benediktsson "Adaptive pixel neighborhood deÞnition for the classiÞcation of hyperspectral images with support vector machines and composite kernel," in *Proc. IEEE ICIP Conf.*, Oct. 2008, pp. 1–5.

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

[18] A. Darwish, K. Leukert, and W. Reinhardt "Image segmentation for the purpose of object-based classification," in *Proc. IGARSS Conf.* , vol. 3. 2003, pp. 2039–2041.

[19] N. Gorretta, J. M. Roger, G. Rabatel, V. Bellon-Maurel, C. Fiorio, and C. Lelong, "Hypersectral image segmentation: The butterfly approach," in *Proc. IEEE Workshop Hyperspectral Image Signal Process.*, Aug. 2009, pp. 1–4.

[20] J. A. Gualtieri and J. Tilton, "Hierarchical segmentation of hyperspectral data," in *Proc. AVIRIS Earth Sci. Appl. Workshop*, 2002, pp. 5–8.

[21] P. Salembier and F. Marques, "Region-based representations of image and video: Segmentation tools for multimedia services," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1147–1169, Dec. 1999.

[22] 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, Apr. 2000.

[23] F. Van der Meer, "The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery," *Int. J. Appl. Earth Observat. Geoinformat.*, vol. 8, no. 1, pp. 3–17, 2006.

[24] S. Valero, P. Salembier, and J. Chanussot, "New hyperspectral data representation using binary partition tree," in *Proc. IEEE IGARSS Conf.*, Jul. 2010, pp. 80–83.

[25] S. Valero and P. Salembier, and J. Chanussot, "Comparison of merging orders and pruning strategies for binary partition tree in hyperspectral data," in *Proc. IEEE ICIP Conf.*, Sep. 2010, pp. 2565–2568.

[26] S. Valero, P. Salembier, and J. Chanussot, "Hyperspectral image segmentation using binary partition trees," in *Proc. IEEE 11th ICIP*, Jun. 2011, pp. 1273–1276.

[27] S. Valero, P. Salembier, J. Chanussot, and C. M. Cuadras, "Improved binary partition tree construction for hyperspectral images: Application to object detection," in *Proc. IEEE 11th IGARSS*, Mar. 2011, pp. 2515–2518.

[28] V. N. Vapnik, *Statistical Learn. Theory*. New York: Wiley, 1998.

[29] S. Lee and M. Crawford, "Unsupervised multistage image classification using hierarchical clustering with a Bayesian similarity measure," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 312–320, Mar. 2005.

[30] L. Gomez-Chova, J. Calpe, G. Camps-Valls, J. D Martin, E. Soria, J. Vila, L. Alonso-Chorda, and J. Moreno "Semi-supervised classification method for hyperspectral remote sensing images," in *Proc. IGARSS Conf.*, Sep. 2003, pp. 1776–1778.

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

[32] C.-I. Chang "Spectral information divergence for hyperspectral image analysis," in *Proc. IEEE IGARSS*, Oct. 1999, pp. 509–511.

[33] M. Dimiccoli and P. Salembier "Hierarchical region-based representation for segmentation and filtering with depth in single images," in *Proc. ICIP Conf.*, Nov. 2009, pp. 3533–3536.

[34] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot "An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 425–441, Apr. 2008.

[35] C. M. Cuadras, A. Arenas, and J. Fortiana "Some computational aspects of a distance-based model for prediction," *Communi. Stat., Simul. Comput.*, vol. 25, no. 3, pp. 593–609, 1996.

[36] T. F. Cox and M. A. Cox, *Multidimensional Scaling*. Ed., K. Fernandez and A. Morineau, London, U.K.: Chapman Hal, 1994.

[37] M. H. Kutner, C. J. Nachtsheim, and J. Neter, *Application Linear Regression Model*. New York: McGraw-Hill, 2004.

[38] C. M. Cuadras, S. Valero, D. Cuadras, P. Salembier and J. Chanussot "Distance-based measures of association with applications in relating hyperspectral images," *Commun. Stat., Theory Meth.*, vol. 41, nos. 13–14, pp. 2342–2355, 2012.

[39] L. Breiman, J. Friedman, R. Olshen, and C. Stone "Classification and regression trees," in *Proc. Wadsworth Int. Group Conf.*, 2004, pp. 1–8.

[40] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A spatial-spectral kernel based approach for the classification of remote sensing images," *Pattern Recognit.*, vol. 45, no. 1, pp. 281–392, 2011.

[41] T. W. Anderson, *An Introduction to Multivariate Analysis*. New York: Wiley, 2003.

[42] J. Benediktsson, L. Bruzzone, J. Chanussot, M. D. Mura, P. Salembier, and S. Valero "Hierarchical analysis of remote sensing data: Morphological attribute profiles and binary partition trees," in *Proc. ISMM, Int. Symp. Math. Morphol.*, 2011, pp. 306–319.

[43] C. M. Cuadras "Multidimensional and dependencies in classification and ordination," Eds. K. Fernandez and A. Morineau, 2009.

[44] M. Dundar and D. Landgrebe "A model-based mixture-supervised classification approach in hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 12, pp. 2692–2699, Dec. 2002.

[45] Y. Escoufier "Le traitement des variables vectorielles," *Int. Biometr. Soc.*, vol. 29, no. 4, pp. 751–76, 1973.

[46] J. C. Gower, "Some distance properties of latent roots and vector methods used in multivariate analysis," *Biometrika*, vol. 53, nos. 3–4, pp. 325–338, 1966.

[47] H. Ling and K. Okada, "Diffusion distance for histogram comparison," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 246–253.

[48] G. Noyel, J. Angulo, and D. Jeulin, "Morphological segmentation of hyperspectral images," *Image Anal. Stereol.*, vol. 26, no. 3, pp. 101–109, 2007.

[49] K. M. Rajpoot and N. M. Rajpoot, "Wavelet based segmentation of hyperspectral colon tissue imagery," in *Proc. IEEE 7th Int. Multitop. Conf. INMIC*, 2003, pp. 142–149.

[50] I. Silverman, S. R. Rotman, and C. E. Caefer, "Segmentation of hyperspectral images from the histograms of principal components," in *Proc. Imag. Spectromet. Conf.*, 2002, pp. 24–29.

Silvia Valero (S'08–M'12) received the M.S. degree in electrical engineering from the Universitat Politcnica de Catalunya (UPC), Barcelona, Spain, in 2007, the M.S. degree in computer science from the Grenoble Institute of Technology (Grenoble-INP), France, in 2008, and the conjoint Ph.D. degree from the Grenoble-INP and at UPC in 2011. Her Ph.D. work is focuses on developing advanced image processing techniques for hyperspectral remote sensing images.

She joined the CESBIO Laboratory, Toulouse, France, in 2012, as Assistant Professor. She is currently working in Land Cover Mapping using multi-temporal images. Her current research interests include image processing, pattern recognition, information retrieval and tree processing techniques.

Dr. Valero was the recipient of the IEEE GRSS Symposium Best Paper Award in 2011.

Philippe Salembier (M'96–SM'09–F'11) received the engineering degree from the Ecole Polytechnique, Paris, France, in 1983, and the electrical engineering degree from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1985, and the Ph.D. from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1991.

From 1985 to 1989, he was with the Laboratoires d'Electronique Philips, Limeil-Brevannes, France, in the fields of digital communications and signal processing for HDTV. He was a Postdoctoral Fellow at the Harvard Robotics Laboratory, Cambridge, MA, in 1991. Then, he joined the Technical University of Catalonia (UPC), Barcelona, Spain, where he is currently a Professor lecturing in the area of digital signal and image processing. His research interests include image and video sequence processing, compression and indexing, mathematical morphology, level sets and nonlinear filtering, as well as remote sensing image processing and signal processing tools for genomics.

Dr. Salembier has served as an Associate Editor of various journals, including the *Journal of Visual Communication and Image Representation*, *Signal Processing (Elsevier), Signal Processing: Image Communication (Elsevier)*, the *Eurasip Journal on Image and Video Processing*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE SIGNAL PROCESSING LETTERS. He was a member of the Image and Multidimensional Signal Processing Technical Committee of the IEEE Signal Processing Society between 2000–2006 and was Technical Chair (with Prof. E. Delp) of the IEEE International Conference on Image Processing, ICIP'2003, organized in Barcelona.

Jocelyn Chanussot (M'04–SM'04–F'12) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from Savoie University, Annecy, France, in 1998.

He has been with Grenoble INP since 1999, where he is currently a Professor of signal and image processing. He is conducting his research at the Grenoble Images Speech Signals and Automatics Laboratory (GIPSA-Lab). His current research interests include image analysis, multicomponent image processing, nonlinear filtering, and data fusion in remote sensing.

Dr. Chanussot is the founding President of the IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010 which received the 2010 IEEE GRS-S Chapter Excellence Award. He was the co-recipient of the NORSIG 2006 Best Student Paper Award, the IEEE GRSS 2011 Symposium Best Paper Award, and the IEEE GRSS 2012 Transactions Prize Paper Award. Since 2011, he has been the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.