Clustering-Based Hyperspectral Band Selection Using Information Measures

Adolfo Martínez-Usó, Filiberto Pla, José Martínez Sotoca, and Pedro García-Sevilla

***Abstract*—Hyperspectral imaging involves large amounts of information. This paper presents a technique for dimensionality reduction to deal with hyperspectral images. The proposed method is based on a hierarchical clustering structure to group bands to minimize the intracluster variance and maximize the intercluster variance. This aim is pursued using information measures, such as distances based on mutual information or Kullback–Leibler divergence, in order to reduce data redundancy and nonuseful information among image bands. Experimental results include a comparison among some relevant and recent methods for hyperspectral band selection using no labeled information, showing their performance with regard to pixel image classification tasks. The technique that is presented has a stable behavior for different image data sets and a noticeable accuracy, mainly when selecting small sets of bands.**

***Index Terms*—Dimensionality reduction, feature clustering, feature selection, information theory.**

I. Introduction

THE benefits of hyperspectral imaging in several disciplines are becoming relevant in many emerging applications. Multi- or hyperspectral sensors acquire data from a range of wavelengths in the spectrum, and apart from the traditional remote sensing application, they are being introduced in important and demanding application fields, such as medical imaging, product quality inspection, and fine arts.

On the other hand, hyperspectral data usually entail dealing with large amounts of information, with little, or even no, labeled information, which makes difficult the application of supervised techniques, for instance, for pixel classification or image segmentation. Another common drawback present in hyperspectral images when performing classification or regression tasks is that hyperspectral information is commonly represented in a large number of bands, which are usually highly correlated; thus, the information provided can contain important data redundancies. Selecting the relevant range of wavelengths in the spectrum while keeping the accuracy for some given tasks is desirable to save computational efforts and data storage, and can also simplify the image acquisition step.

Manuscript received February 22, 2007; revised May 11, 2007. This work was supported by the Spanish Ministry of Science and Education under Projects ESP2005-07724-C05-05 and CSD2007-00018.

The authors are with the Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, 12071 Castellón de la Plana, Spain (e-mail: auso@uji.es; pla@uji.es; sotoca@uji.es; pgarcia@uji.es).

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

Digital Object Identifier 10.1109/TGRS.2007.904951

This reduction in the hyperspectral representation could be done using feature extraction [1]–[4] or feature selection techniques [5], [6]. In feature extraction, we would obtain a new and reduced data set representing the transformed initial information, whereas in feature selection, we would have a subset of relevant data from the original information. Other techniques try to exploit spatial information for dimensionality reduction purposes [7].

In hyperspectral imaging, feature or band selection is preferable to feature extraction for dimensionality reduction because of two main reasons [8]. On the one hand, feature extraction would need the whole (or most) of the original data representation to extract the new features, forcing to always obtain and deal with the whole initial representation of the data. In addition, since the data are transformed, some crucial and critical information may have been compromised and distorted, for instance, when dealing with physical measures that are represented in the hyperspectral image domain, while band selection has the advantage of preserving the relevant original information from the data.

The availability of little or no labeled information has also been a matter of attention, with increasing recent work on band reduction techniques for different hyperspectral imaging tasks [8], [9] being the most common one-pixel classification.

In summary, a very desirable preprocessing step in hyperspectral imaging and, particularly, on pixel classification tasks is to perform a band selection process to reduce the redundant information in the image representation without losing classification accuracy in a significant way and using no supervised information. To this end, we propose a new technique that does the following.

- Exploits band correlation through a clustering-based algorithm. A similar strategy has also been used in distributional clustering for text categorization [10] or data compression [11], due to the high dimensionality that these approaches have to deal with.
- Can use different measures to discriminate among the bands. In this paper, two different measures are proposed, resulting in two variants of the same algorithm. Both criteria are based on information theory measures [12].
- Obtains subsets of relevant bands to try to get the best classification performance, mainly when selecting small sets of bands, where the band selection methods have to really show their capabilities to extract the relevant information in the data.
- It is not a ranking or incremental method. That is, the best m bands are not the best $m-1$ bands plus another relevant band.

In order to show the performance of the proposed method, we also present a comparison against several unsupervised methods for band selection in hyperspectral imaging. These methods have been chosen for their relevance in band selection using no labeled information and with the intention of covering as many tendencies as possible in this field. The comparison is done by testing the proposed and reference methods with different hyperspectral images and several types of classifiers, using the pixel classification accuracy as a criterion to check the relevance of the bands selected.

II. Clustering-Based Band Selection

In hyperspectral imaging for remote sensing, it is very common to have very little or no labeled information. Therefore, a band selection technique that uses no supervised information can be really useful. On the other hand, techniques that do not use supervised information can use the whole data set available, while supervised data sets usually provide labeled information of only one part of the available data.

In text categorization [10], distributional clustering techniques have been used to deal with high dimensionality, compressing the initial data representation by means of word/feature clusters. These techniques treat each word as a single feature, and these clustering methods can be more effective than standard feature selection, particularly, at lower number of features [13]. These techniques use supervised information for selecting the most representative words. However, we are interested in a dimensionality reduction method that can manage with no labeled information.

Similar clustering strategies have been applied in order to select relevant filter responses from a bank of spectral channels [14] or for data compression, such as *vector quantization* [11]. Obviously, representing data by fewer instances or a reduced feature representation will lead to a loss of information. However, keeping only the most discriminative features or data points, data analysis can become simpler, losing no significant performance.

Therefore, the band selection technique proposed here is based on a clustering process that is performed in a similarity space that is defined among bands. Against other techniques that rank bands by means of a similarity measure, a process that joins similar bands together is proposed, constructing a family of derived clusters that preserves a low variance among the bands that belong to the same cluster and a high variance among different clusters, in an analogous way, as clustering is used in vector quantization for data compression. The final selected bands will be the best representative instances from each cluster. Moreover, in contrast to other authors that use a divisive clustering approach [10], we advocate for an agglomerative clustering strategy, in order to also reflect the hierarchical nature of the spectrum structure [1].

In addition to these essential requirements, one of the main objectives of this paper is a significant reduction of the redundant information, keeping a high accuracy in classification tasks. To this end, from information theory [15], we can find information measures that can quantify how much a given random variable can predict another one. We will particularly focus on this property. Therefore, we propose the use of two different measures to exploit this point: 1) the mutual information and 2) the Kullback–Leibler divergence.

Not only is mutual information widely used as a criterion for measuring the degree of independence between random variables but it also measures how much a certain variable can explain the information content about another variable, being a generalized correlation measure. Thus, a dissimilarity measure between two bands (random variables) can be defined based on this measure as a relevance criterion. On the other hand, the Kullback–Leibler divergence has been employed as a measure of discrepancy between any two probability distributions, and it can be interpreted as the cost of substituting a given probability distribution with another one. This criterion was already applied to compare hyperspectral image bands [9].

A. Dissimilarity Measures

1) Mutual-Information-Based Criterion: The first dissimilarity measure that is proposed tries to identify the subset of the selected bands that are as independent as possible among them. It is known that independence between bands [1] is one of the key issues to obtain relevant subsets of bands for classification purposes. As we will show in the experimental results, identifying subsets of bands that are as much independent as possible among them indeed produces very satisfactory classification rates with regard to other band selection approaches.

The use of information measures, such as mutual information, in order to quantify the degree of independence, provides a methodology to find generalized correlations among image bands. Thus, this technique exploits this concept for band selection in order to reduce data redundancy and nonuseful information.

Let us introduce some information theory concepts and properties [12], [16]. The Shannon entropy of a random variable X with probability density function $p(x)$ for all possible events $x \in \Omega$ is defined as

$$H(X) = -\int_{\Omega} p(x)\ \log p(x)\ dx. \tag{1}$$

In the case of a discrete random variable X, entropy $H(X)$ is expressed as

$$H(X) = -\sum_{x \in \Omega} p(x)\ \log p(x) \tag{2}$$

where $p(x)$ represents the mass probability of an event $x \in \Omega$ from a finite set of possible values. Entropy is often taken as the related amount of *information* of a random variable.

On the other hand, *mutual information* I is a measure of independence between random variables. I can be interpreted as a generalized correlation measure, which includes the linear and nonlinear dependence between variables. In other words, mutual information quantifies the statistical dependence of random variables or how much a variable can predict another one.

Due to the complexity in calculating the joint distribution in high-dimensional spaces [12], estimation of $I(\hat{\mathbf{S}}, \mathbf{S})$, where $\hat{\mathbf{S}}$ is a subset of random variables out of the original set $\mathbf{S}$

such that $\hat{\mathbf{S}} \subset \mathbf{S}$, becomes complex and highly computational expensive. This is a critical issue from a practical point of view. In this sense, the technique proposed here tries to overcome this drawback by only using comparisons between pairs of random variables, through defining a similarity measure based on the mutual information between two random variables.

Let us consider a set of L random variables that represent their corresponding bands $X_1, \ldots, X_L$ from a hyperspectral image. $I(X_i, X_j)$ is defined as

$$I(X_i, X_j) = \sum_{x_i \in \Omega} \sum_{x_j \in \Omega} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)}. \quad (3)$$

I is always a nonnegative quantity for two random variables, being zero when the variables are statistically independent. The higher the I, the higher the dependence between the variables. Furthermore, the following property about two random variables always holds:

$$0 \leq I(X_i, X_j) \leq \min\{H(X_i), H(X_j)\}. \quad (4)$$

Mutual information I can be expressed in terms of entropy measures according to the following expression:

$$I(X_i, X_j) = H(X_i) + H(X_j) - H(X_i, X_j) \quad (5)$$

where $H(X_i, X_j)$ is the joint entropy, which is defined from the joint probability distribution $p(x_i, x_j)$.

So far, I has been introduced as an absolute measure of common information shared between two random sources. However, as we can infer from (5), I by itself would not be suitable as a similarity measure. The reason is that it can be low because either the X_i, X_j variables present a weak relation (such as it should be desirable) or the entropies of these variables are small (in such a case, the variables contribute with little information). Thus, it is convenient to define a proper measure, so that it works independently from the marginal entropies and also measures the statistical dependence as a similarity measure.

Thus, the following measure of similarity between two random variables will be used:

$$NI(X_i, X_j) = \frac{2 \cdot I(X_i, X_j)}{H(X_i) + H(X_j)} \quad (6)$$

which is a normalized measure of I. Furthermore, this normalized mutual information is used as a dissimilarity or distance measure as follows [14]:

$$D_{NI}(X_i, X_j) = \left(1 - \sqrt{NI(X_i, X_j)}\right)^2. \quad (7)$$

Fig. 1 represents the dissimilarity matrix D_{NI} as a gray-level *image*, where darker values represent high correlated bands for the *HyMap* image with 128 bands (see Section III-C for a database details). Using this distance in a clustering process will lead to K selected bands from the final K clusters, having low correlation (mutual information) and therefore a significant degree of independence.

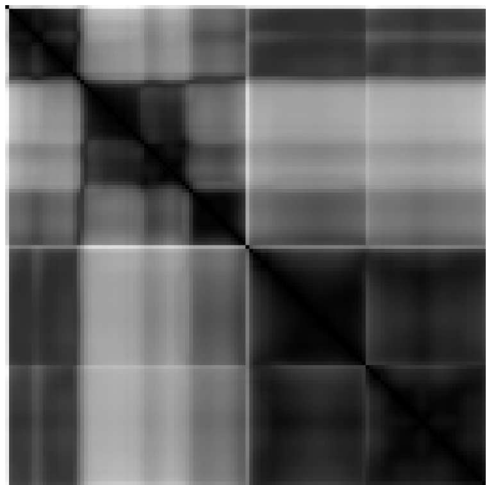

Fig. 1. Dissimilarity matrix for a hyperspectral image with 128 bands.

2) Divergence-Based Criterion: Another information measure to be considered is the Kullback–Leibler divergence, which can be interpreted as a kind of dissimilarity distance between two probability distributions, although it is not a real distance measure because it is not symmetric. Thus, a symmetric version of the Kullback–Leibler divergence is often used [12], [16].

Let us call X_i and X_j two random variables that are defined in Ω space, representing the ith and jth bands of a hyperspectral image. Let us assume that $p_i(x)$ and $p_j(x)$ are the probability distributions of these random variables. Thus, the symmetric Kullback–Leibler divergence can be expressed in the discrete domain as follows:

$$D_{KL}(X_i, X_j) = \sum_{x \in \Omega} p_i(x) \log \frac{p_i(x)}{p_j(x)} + \sum_{x \in \Omega} p_j(x) \log \frac{p_j(x)}{p_i(x)} \quad (8)$$

The Kullback–Leibler divergence is always nonnegative, being zero when $p_i(x)$ and $p_j(x)$ are the same probability distribution. This divergence measure can be used as a criterion to know how far two distributions are, and it can be interpreted as the cost of using one of the distributions instead of the other one. In the hyperspectral band selection framework, it can be used as a measure of dissimilarity between two image bands, which are represented by their corresponding probability distributions.

This divergence measure is the second criterion that is proposed to be used as a distance for the clustering process, and it has been frequently used in order to compare different probability distributions, also in hyperspectral imaging to measure the overlapped information that is contained in a pair of image bands, as a band-decorrelation algorithm [9].

B. Variance-Reduction Clustering Strategy

Using the introduced criteria either based on the mutual information or the Kullback–Leibler divergence as a dissimilarity measure between two image bands, a hierarchical clustering process is then proposed, in order to form clusters of bands as similar as possible among them within each cluster. The clustering is part of an information compression process, and

at the end of the clustering process, a representative band for each cluster is selected, which will substitute all bands in the cluster, at the lowest possible cost in terms of information loss. The selected representatives will constitute the subset of bands that were selected as a compressed image band representation for the whole original set of image bands.

1) Hierarchical Clustering: Hierarchical structures are a very intuitive way to summarize certain types of data sets. One interesting characteristic of hierarchical methods is the fact that different linkage strategies create different tree structures. The algorithm proposed here uses an agglomerative strategy. Thus, the number of groups is reduced one by one.

In particular, a hierarchical clustering algorithm based on Ward's linkage method [17] is used. Ward's linkage has the property of producing minimum variance partitions. Thus, this method is also called minimum variance clustering, because it pursues to form each possible group in a manner that minimizes the loss that is associated with each grouping (internal cohesion). Several studies point out that this method outperforms other hierarchical clustering methods [18], but, in our case, the process also helps us to form groups with low variance in their level of similarity.

Briefly summarizing the linkage strategy, let us suppose that clusters C_r and C_s are merged. The general expression for the distance between the new cluster (C_r, C_s) and any other cluster (C_k) is defined as

$$D\left[(C_k),(C_r,C_s)\right] = \alpha \cdot D(C_k,C_r) + \beta \cdot D(C_k,C_s) + \gamma \cdot D(C_r,C_s) + \delta \cdot |D(C_k,C_r) - D(C_k,C_s)| \quad (9)$$

where α, β, γ, and δ are the merging coefficients. Ward's intercluster distance results from the following coefficients:

$$\alpha = \frac{n_r + n_k}{n_r + n_s + n_k} \quad \beta = \frac{n_s + n_k}{n_r + n_s + n_k}$$
$$\gamma = \frac{-n_k}{n_r + n_s + n_k} \quad \delta = \emptyset$$

where n_i is the number of instances in group i.

The algorithm starts with the disjoint partition where each cluster is formed as a single pattern (hyperspectral band). At this step, dissimilarity matrix $D_{L\times L}$ is initialized by means of the dissimilarity measures that were described in either Section II-A1 or Section II-A2. After that, the algorithm looks for the two most similar clusters that will have the minimum distance value in matrix $D_{L\times L}$. Then, these two clusters are merged into one, and matrix $D_{L\times L}$ is updated using expression (9). The rows/columns corresponding to the merged clusters are deleted, and a row/column for the new cluster is added.

This process is repeated until K number of desired clusters are obtained. The resulting mutually exclusive clusters represent groups of highly correlated bands, and bands from two different clusters will have low correlation.

2) Selecting Cluster Representatives: Let us consider now a resulting cluster C with R bands. The weight of each band $X_i \in C$ is defined as

$$W_i = \frac{1}{R} \sum_{j\in C, j\neq i} \frac{1}{\epsilon + D(X_i,X_j)^2} \quad (10)$$

where ϵ is a very small positive value to avoid singular values, and function $D(X_i, X_j)$ returns the distance value between bands i, j. The representative band from each group is selected as the band with the highest W_i in the cluster.

A low value of W_i means that band i has an average large distance from the other bands in the cluster, i.e., in this case, band i will have an average low correlation with regard to the other bands in the cluster. In a reverse way, a high value of W_i means that band i has, on average, a high correlation with regard to the other bands in the cluster.

Therefore, when selecting cluster representative bands by using dissimilarity measure D_{NI}, choosing the band in the cluster with the highest average correlation (mutual information) with regard to the other bands in the cluster is equivalent to choosing the band that better predicts the information content of the other bands in the cluster; this is because the more mutual information that two random variables share, the more one of the variables can predict about the other one, which, in this sense, creates a high degree of dependence among them.

On the other hand, when selecting cluster representative bands by using distance D_{KL}, choosing the band in the cluster with the highest average divergence with regard to the other bands in the cluster is equivalent to select the band that would produce the lowest cost, in the average sense, when substituting every band in the cluster by its representative.

As a result of the algorithm, there will be K bands selected, representing K different clusters. The bands within the same cluster will have a high correlation. The selected bands will also cover the dissimilarity space, being a compressed representation that tries to explain most of the information that is contained in the original representation.

C. Implementation and Computational Issues

One of the key implementation issues is the estimation of probability function $p(x)$ for each band X. Probability density function $p(x)$ for all events $x \in \Omega$, where Ω is the set of possible values that a random variable X can take, will be estimated for each image band as $p(x) = h(x)/(MN)$, with $h(x)$ being the gray-level histogram and MN being a normalizing factor, which is the number of pixels in the image. In an analogous way, to estimate the joint probability distribution $p(x_i, x_j)$ between two bands X_i and X_j, the corresponding joint histogram $h(x_i, x_j)$ is first required, and the probability distribution is then computed as $p(x_i, x_j) = h(x_i, x_j)/(MN)$.

The whole algorithm can be divided into two main parts: 1) the operations done before the clustering (*preclustering*) and 2) the operations properly involved in the hierarchical clustering process (*clustering*). Note that, in the *preclustering* part, we shall distinguish two different processes depending on the measure used when we calculate the distances between any

TABLE I
PROCESSING TIMES FOR EACH OF THE ANALYZED METHODS USING THE HYMAP IMAGE WITH 128 BANDS (SECTION II-C)

	WaLuMI	WaLuDi	BCM-BDM
CPU time	1m48s	49s	141m58s
	BCC-BDC	MVPCA	ID
CPU time	149m47s	3m10s	10s

pair of bands: 1) based on mutual information D_{NI} or 2) based on the divergence criterion D_{KL}.

- *Preclustering*: When the process begins, each band in the image is considered as a separated cluster. Then, a distance matrix of size $L \times L$ is initialized with the corresponding distances between pairs of bands, obtaining a symmetrical matrix.
  - — When using the distance based on mutual information D_{NI}, the histogram for each single band and the cojoint histogram for each pair of bands must be computed. Thus, assuming that $MN > G$, where G is the number of gray levels in the bands, the temporal cost of this part is $O(L^2MN)$.
  - — When using the divergence criterion D_{KL}, the cojoint histograms are not required. Now, the temporal cost of this part is $O(LMN + L^2G)$.

  Although, the D_{KL} criterion requires less computational effort for the matrix initialization, both methods are computationally affordable. From the point of view of the spatial cost, only the distance matrix, the histograms, and the image bands are required in the process. Furthermore, only one pair of bands is required in main memory at a time when the cojoint histograms must be computed.
- *Clustering*: This part is related to the operations that the Ward's linkage method involves. Once the distance matrix has been initialized, its minimum value is found in order to choose two clusters to be joined. Then, a new row and a new column are added for the new cluster that was created, and their entries in the distance matrix are computed according to (9). Afterward, the rows and columns for the old clusters in the distance matrix are removed. The process repeats until the desired number of clusters is reached. There are no additional requirements of memory in this step. The temporal cost of this part is $O(L^3)$, which will be significantly lower than the preclustering part for usual images. Only if we had to deal with small images and a large number of bands would the temporal cost of this part comparable with the cost of the preclustering part.

As an illustrative example of these computational costs, using an Intel Pentium IV 3.00-GHz CPU and a HyMap image with 128 bands of 700 $\times$ 670 pixels (see Section III-C for a database description), our current implementation required about 47 s for the *preclustering* part (using D_{KL}), whereas the *clustering* part required just 2 s. Table I gives a simple quantitative analysis of the computational cost of each method for HyMap image. Note that no optimizations were considered, but a direct implementation of the algorithms that were described were run. Implementations were developed in C++. For the linearly constrained minimum variance (LCMV)–constrained band selection (CBS) methods, it is important to point out that the practical optimization that was described in [8] has been assumed, i.e., since band correlation minimization (BCM), band dependence minimization (BDM) and band correlation constraint (BCC), band dependence constraint (BDC) share the resulting bands that were selected, the best time of each pair has been taken into account, otherwise BDM and BDC methods are much more time-consuming.

In summary, the method presented here is computationally affordable, even for hyperspectral images with a large number of input bands since it is based on probability estimations from histogram pixel values of, at most, pairs of bands, avoiding unfeasible high-dimensional probability estimations.

III. EXPERIMENT AND RESULT

The experimental results will consist of comparing our method with relevant techniques from recent literature using different classifiers and databases. Methods, databases, and classifiers are described in the succeeding sections.

A. Comparison With Other Techniques

In order to assess the performance of the proposed method regarding state-of-art techniques, a comparison study has been made with different methods that perform band selection using no labeled information and that are a reference in the field [8], [9], [19]. In addition, the comparison also includes other dimensionality reduction methods, such as maximum-variance principal component analysis (MVPCA) and information divergence (ID) methods [9]. The succeeding sections summarize briefly the methods that were used in the comparison.

1) CBS Method: CBS [8] is an approach that is different from the variance-based methods or information theoretic criteria-based methods. This technique constrains linearly a band while minimizing the correlation or dependence of this particular band with respect to the other bands in a hyperspectral image. CBS methods propose four different solutions to an optimization problem, with two based on correlation and two based on dependence. At the same time, these four solutions arise from two different approaches: 1) constrained energy minimization (CEM) and 2) LCMV. Since the experimental results show that both approaches perform similarly, LCMV is usually used, because the computational complexity is reduced substantially [20], [21]. Moreover, the sizes of the images that were used in this comparison make the use of the CEM-CBS implementation unfeasible.

In [8], the authors propose four criteria for band selection (see Table II for analytical details).

1) bcm_L (BCM) represents the minimal correlation of the lth band with the entire hyperspectral image in the least square sense by constraining band image b_L. Thus, the larger the bcm_L, the higher the correlation of the band, therefore becoming the better choice in the hyperspectral image.
2) bdm_L (BDM) represents how much dependence the lth band has on the other bands. The larger the bcm_L, the more significant the band.

TABLE II
SOLUTIONS BASED ON CORRELATION OR DEPENDENCE FILTERS

Correlation
$bcm_l = v_l^T \Sigma v_l$
$bcc_l = \sum_{k=1,k\neq l}^{L} 1_N^T (b_k^T v_l)$
Dependence
$bdm_l = \tilde{v}_l^T \tilde{\Sigma}_l \tilde{v}_l$
$bdc_l = \sum_{k=1,k\neq l}^{L} 1_N^T (b_k^T \tilde{v}_l)$

3) bcc_L (BCC) measures the degree of the band constraint, taking into account its correlation on all other bands from the hyperspectral image.
4) bdc_L (BDC) measures the degree of the band constraint, taking into account its dependence on all other bands from the hyperspectral image.

2) MVPCA Method: MVPCA is a joint band-prioritization and band-decorrelation approach to band selection, which was introduced in [9] for hyperspectral image classification and also used in some comparative works for band selection [8]. This band prioritization is based on an eigenanalysis, decomposing a matrix into an eigenform matrix from which a loading factor matrix could be constructed and used to prioritize bands. The loading factors determine the priority of each band and rank all bands in accordance with their associated priorities. Thus, bands so sorted are bands sorted from high to low variance.

3) ID Method: The divergence-based criterion that was described in Section II-A2 has been previously used in other works as a discriminative criterion between probability distributions for spectral band selection. In [9], it was used in the experimental comparison as a dissimilarity measure. In [8], it was also included in the comparative results, using it to assess how different a probability distribution that is associated to a band of the hyperspectral image is from a Gaussian probability distribution. Thus, the *ID method* measures how far from a Gaussian behavior a probability distribution is, sorting the bands according to the decreasing order of ID [8], i.e., from non-Gaussian bands to Gaussian ones.

B. Notational Comments

Hereafter and, particularly, in graphs/tables, the following notation will be used: *WaLuMI* (*Ward's Linkage strategy Using Mutual Information*) will denote the proposed method that uses the distance based on mutual information, as described in Section II-A1. Analogously, *WaLuDi* (*Ward's Linkage strategy Using Divergence*) will denote the method that uses the Kullback–Leibler divergence measure that was introduced in Section II-A2.

About the LCMV–CBS approach that was presented in the comparison of this work, the *BCM/BDM* and *BCC/BDC* alternatives have been joined together due to the similar results that they produced (results were also joined in this way in [8]). These methods are also referred as LCMV–CBS in graphs/tables.

C. Databases Description

To test the proposed method and the described approaches in the comparison, four different databases of hyperspectral images were used in the experimental results. Examples of these images are shown in Fig. 2.

1) The *92AV3C* source of data corresponds to a spectral image (145 $\times$ 145 pixels, 220 bands, and 17 classes composed of different crop types, vegetation, man-made structures, and an unknown class) that is acquired with the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data set and collected in June 1992 over the Indian Pine Test site in Northwestern Indiana (http:/dynamo.ecn.purdue.edu/~biehl/MultiSpec). As described in [22] and [23], several bands should be discarded from this database due to the effect of atmospheric absorption. Thus, 185 out of the 220 bands were used, discarding the lower signal-to-noise ratio (SNR) bands.
2) The *DAISEX'99* project provides useful aerial images about the study of the variability in the reflectance of different natural surfaces. This source of data, which is referred to as *HyMap* in figures/tables, corresponds to a spectral image (700 $\times$ 670 pixels and seven classes that are composed of crops and an unknown class) acquired with the 128-band HyMap spectrometer during the DAISEX'99 campaign (http:/io.uv.es/projects/daisex/). In this case, 126 bands were used, discarding also the lower SNR bands.
3) The third database is an example of the application of hyperspectral imaging to other applications that are different from remote sensing. It contains hyperspectral images of orange fruits obtained by an imaging spectrograph (RetigaEx, Opto-knowledged Systems Inc., Canada). It has two groups of hyperspectral images. The first one covers the spectral range extended from 400 to 720 nm in the visible (VIS, 676 $\times$ 516 pixels), obtaining a set of 33 spectral bands for each image. The second group covers the spectral range from 650 to 1050 nm in the near-infrared (NIR, 676 $\times$ 516 pixels), obtaining a set of 41 spectral bands for each image. In both cases, the camera has a spectral resolution of 10 nm. Regarding the database content, each hyperspectral image has classes ranging from three to nine, depending on the orange view/type. Classes are composed basically of background, healthy skin, and unhealthy skin, but images can include a stem class and several types of unhealthy skin (defects as rot, trip, overripe, or scratch). Concretely, the hyperspectral image that is used in our experiments has four classes, i.e., background, healthy skin, unhealthy skin (rot), and stem. No band was discarded in this case.
4) Satellite PROBA has a positional spectroradiometric system (CHRIS) that measures the spectral radiance, i.e., the amount of light that passes through or is emitted from a particular area, and falls within five given angles in a specified direction. System CHRIS–PROBA is able to operate in several acquisition modes. The images that are used in this paper come from the mode that operates on an area of 15 $\times$ 15 km, with a spatial resolution of 34 m, obtaining a set of 62 spectral bands that range from 400 to 1050 nm (641 $\times$ 617 pixels and nine classes that are composed of crops and an unknown class). The camera

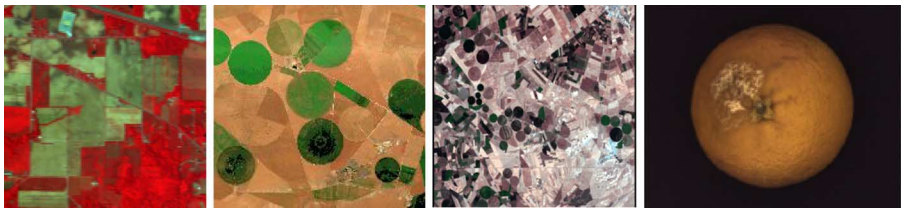

Fig. 2. Database examples. First for AVIRIS (92AV3C), second for HyMap spectrometer, third for a CHRIS–PROBA system, and fourth for the orange image from the VIS collection. The images are presented as red–green–blue compositions.

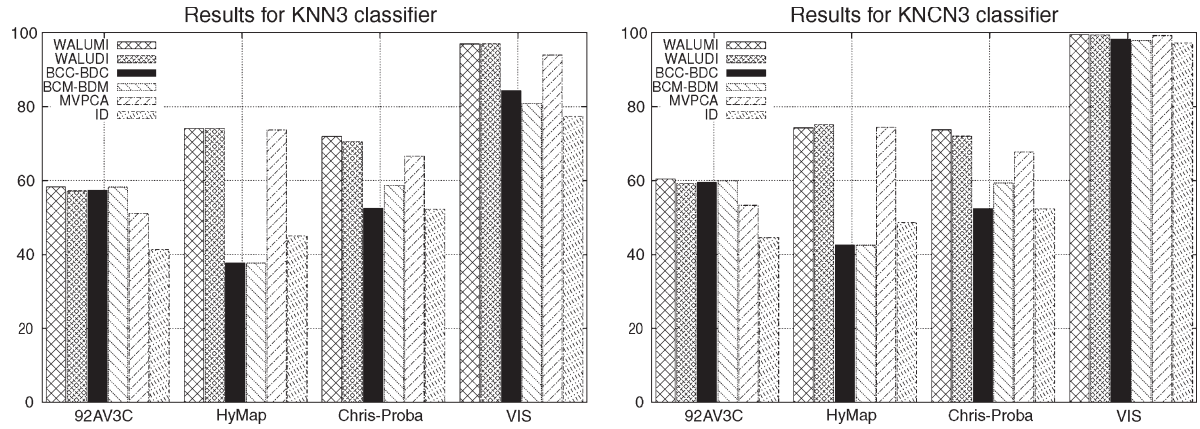

Fig. 3. KNN3 and KNCN3 classifier results for each image. Results cover the average classification rate up to $K = 15$ selected bands.

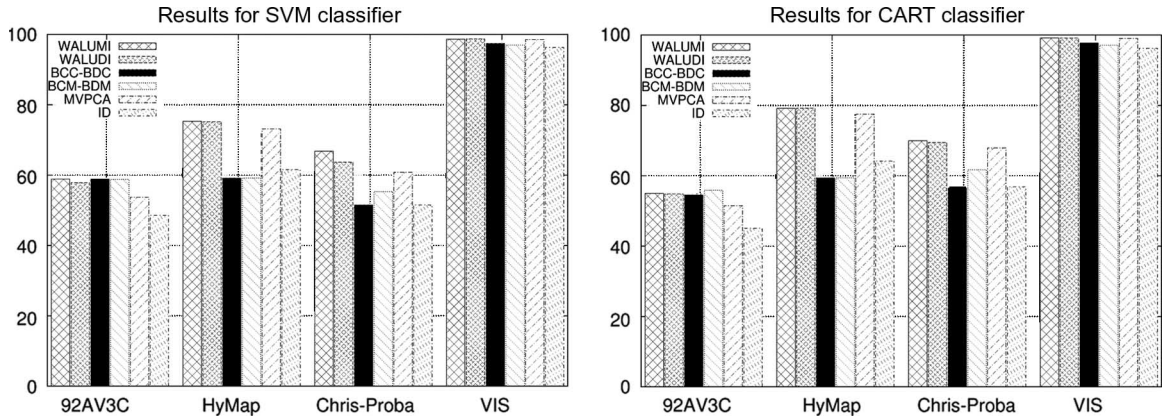

Fig. 4. SVM (polynomial) and CART classifier results for each image. Results cover until $K = 15$.

has a spectral resolution of 10 nm. Concretely, images cover the area that is known as Barrax (Albacete, Spain). In this case, 52 bands were used, discarding the lower SNR bands.

D. Classifier Description

As a validation criterion, in order to evaluate the performance of each band selection method, a supervised pixel classification process using a labeled image with the different classes has been used. To this end, four classifiers have been used to compare the significance of the subsets of selected image bands that were obtained when using different classification schemes.

1) *K-Nearest Neighborhood* (*KNN3*) [24]: Among the various methods of supervised statistical pattern recognition, the k-nearest neighbor rule achieves high performance when a sufficiently large number of samples is available, without *a priori* assumptions about the distributions from which the training examples are drawn. A new sample is classified by calculating the distance to the k-nearest training cases. The class of those training points then determines the classification of the new sample by a majority-voting scheme.
2) *K-Nearest Centroid Neighborhood* (*KNCN3*) [25]: Let p be a sample whose k neighbors are found in a training set. These neighbors have to be found such that (a) the first nearest centroid neighbor of p corresponds to its nearest neighbor (for example, q_1) and (b) the ith nearest centroid neighbor [for example, q_i $(i \geq 2)$] is such that the centroid of this and all previously selected nearest centroid neighbors $q_1, \ldots, q_i$ is the closest to p. This produces a neighborhood in which both closeness and

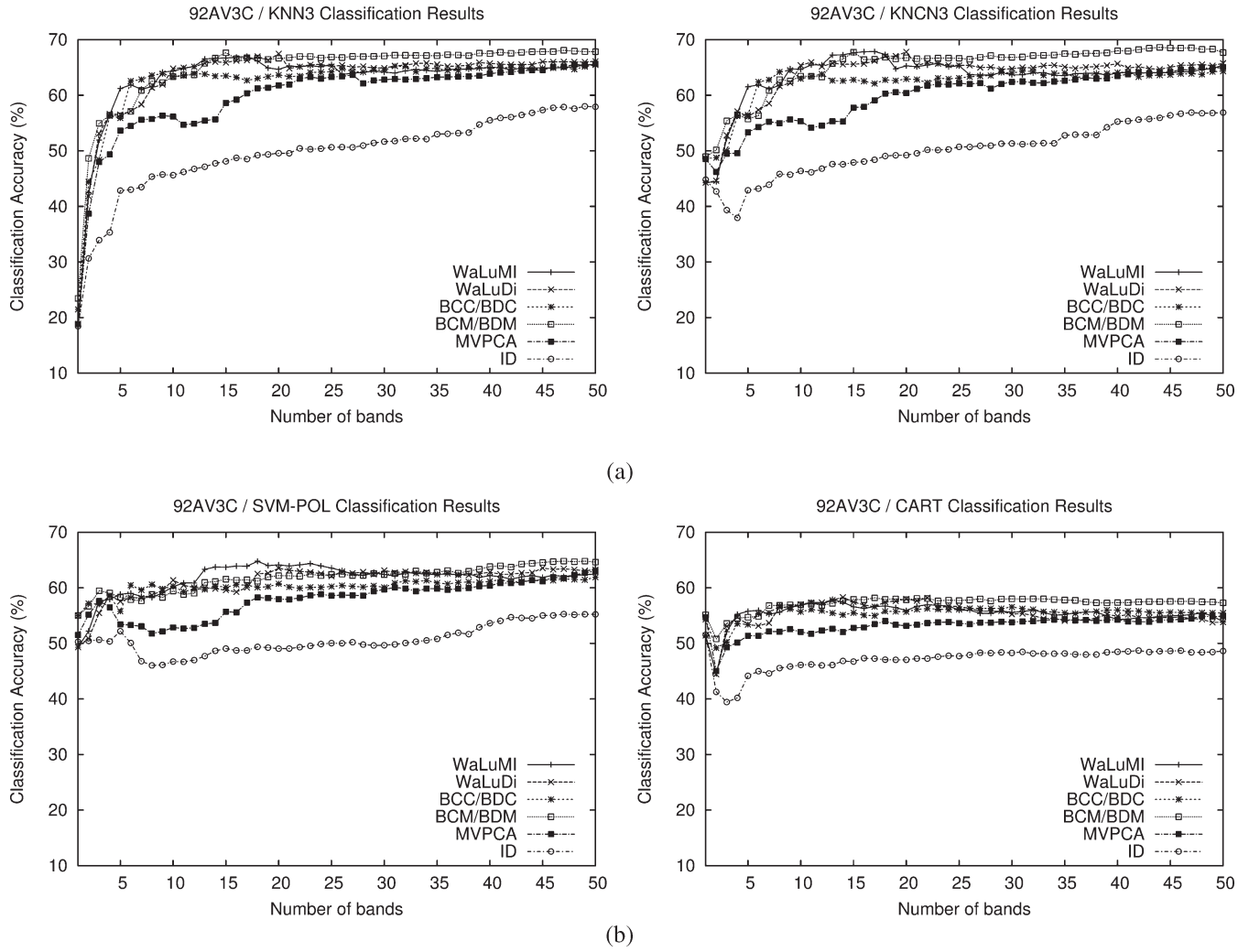

Fig. 5. (Top row) AV3C DB–KNN3 and KNCN3, and (bottom row) SVM (polynomial) and CART classifiers.

geometrical distribution of neighbors are taken into account because of the centroid criterion. On the other hand, the proximity of the nearest centroid neighbors to the sample is guaranteed because of the incremental nature of the way in which those are obtained from the first nearest neighbor.

3) *Support Vector Machine* (*SVM*) [26], [27]: SVMs are a set of related supervised learning methods that are used for classification and regression. In this case, a family of generalized *linear/polynomial* classifiers has been used. A special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin of the decision boundary. The effectiveness of SVM classifiers has been widely tested, showing a good performance against other classifiers [28].
4) *Classification And Regression Trees* (*CART*) [29]: The CART methodology is a binary-decision-tree-based technique. For the experimental results, the Gini criterion was used for splitting nodes, and the resulting tree was not pruned.

In order to increase the statistical significance of the experimental results, classification rates were provided by averaging the classification accuracy that was obtained by the classifiers over five random partitions in each image data set. The samples in each partition were randomly assigned to the training and test sets with equal sizes as follows: HyMap $=$ $37\,520$ pixels, 92AV3C $= 2102$ pixels, VIS $= 34\,882$ pixels, and CHRIS$-$PROBA $= 1788$ pixels. The proposed setup satisfies that the sum of the elements from the different partitions constitutes the entire original set, and the *a priori* probabilities for each class in the data sets are preserved, as well as the statistical independence between the training and test sets of every partition. It is important to point out that, in the case of the SVM classifier, the test set size has been preserved, but the training set size had to be reduced due to the huge computational cost, maintaining the *a priori* probabilities for each class. Several tests have been carried out in order to evaluate how the number of pixels that were used to train the SVM classifier affects both classification rate and computational cost. Thus, for SVM, 400 pixels were used in the training set, because we found that, for this value, the classification rate has already reached its ceiling, and at the same time, the computational cost is still affordable.

Finally, it is important to point out that, due to space limitations, only quantitative results are presented for *KNN3*, *KNCN3*,

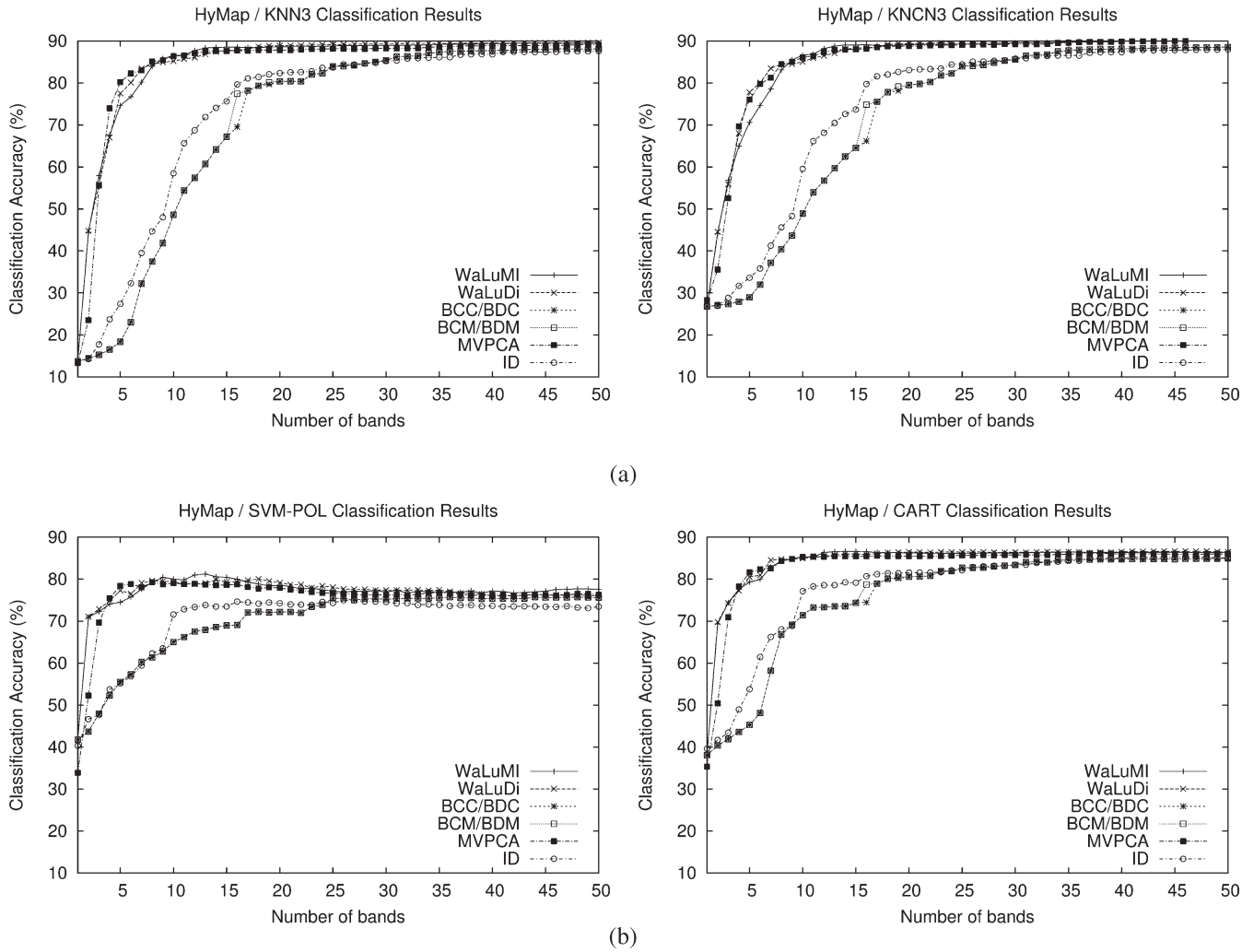

Fig. 6. (Top row) HyMap DB–KNN3 and KNCN3, and (bottom row) SVM (polynomial) and CART classifiers.

SVM with a polynomial kernel, and *CART* classifiers. However, *nearest neighbor* and *SVM* with a linear kernel were also tested with very similar classification results. The *Naive Bayes* classifier [30] was also used in the experiments, but its results were not included because of the significant lower performance that it provided with regard to the other classifiers.

E. Discussion

Using the previously described experimental setup, the methods that were described for the comparison and the proposed approach were applied in order to obtain a ranking of relevance of the selected spectral bands with respect to the classification performance that was achieved. The results that were obtained can be summarized here.

- Classifiers *KNN3* and *KNCN3* use $k = 3$ neighbors. $k = 5$ and $k = 7$ alternatives have also been tested, obtaining similar results.
- The graphs that are presented in Figs. 3 and 4 summarize the experimental results that were carried out. Each graph shows the results that were obtained for each classifier. The x-axis represents the different band selection approaches that are grouped for each image database where they were applied, whereas the y-axis shows the average classification rate that was obtained by each method for the corresponding image. The average classification is computed over the first 15 subsets of features, which has been considered to be the approximate transitory period to reach a stable performance (flat zone of the classification). The transitory zone is considered to be the most important phase of the learning curve shown in Figs. 5–8, where the band selection methods show their potential to really select relevant bands. From these results, we have certain observations.
  - —The *VIS* database is the one with the lowest difficulty, whereas *92AV3C/HyMap/CHRIS–PROBA* have a similar complexity.
  - —The WaLuMI and WaLuDi techniques present consistent behavior, achieving good results in all databases and for all classifiers, getting either the best or the nearly best results.
- Figs. 5–8 show graphs with the classification rates related to the subset of K bands that were selected by each method. Note that around $K = 15$, practically all the methods for all the classifiers have reached the maximum classification performance, which is named here as the

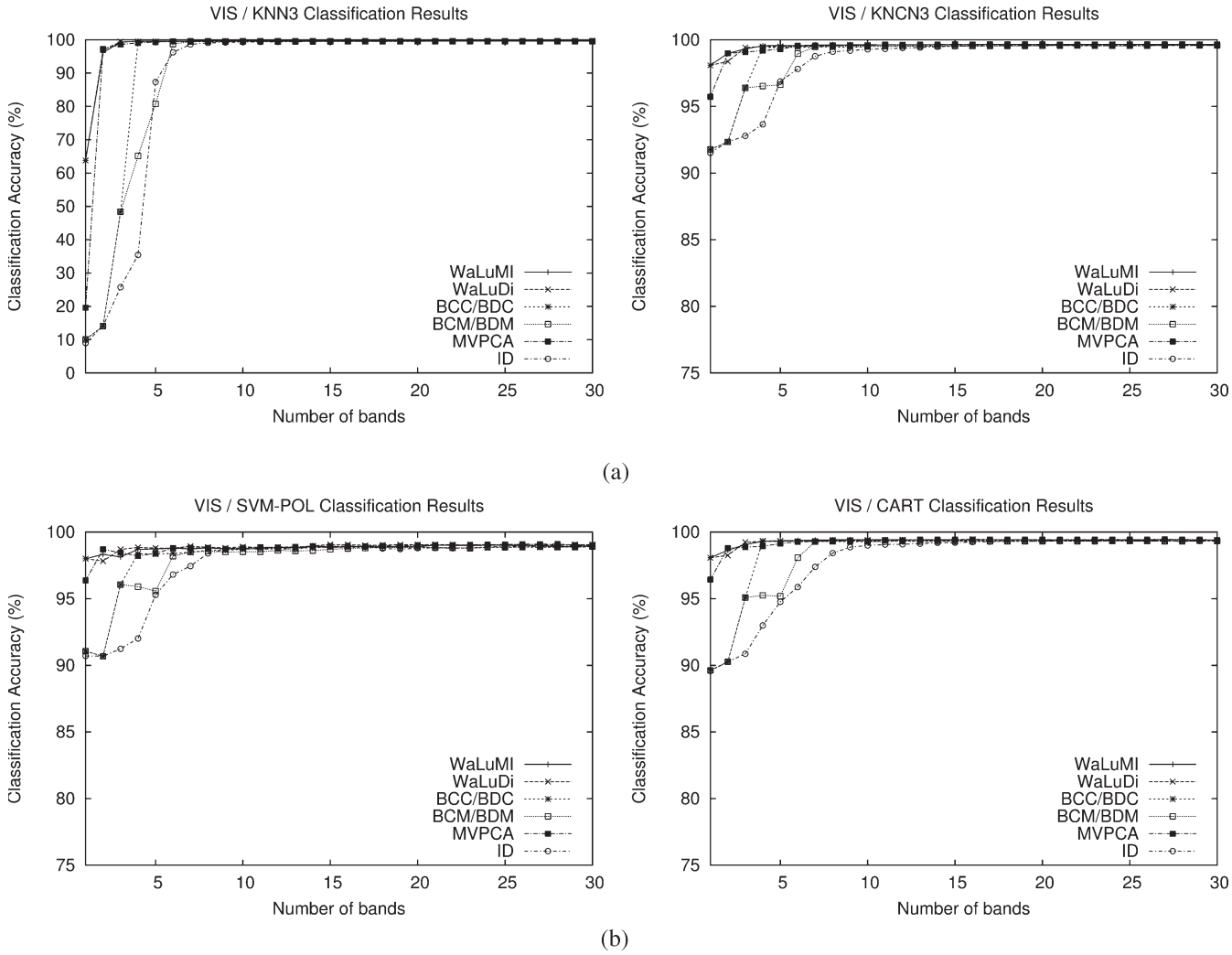

Fig. 7. (Top row) VIS DB–KNN3 and KNCN3, and (bottom row) SVM (polynomial) and CART classifiers.

flat zone of the learning curve. The initial zone, up to approximately $K = 15$, is considered the transitory zone of the learning curve, where the selection of a certain band is more critical.

- Tables III–VI show the same results as the graphs but in an alternative way. These tables summarize the classification rates numerically. Taking into account all possible values of K, results in rows *Up to* $K = 5$, *Up to* $K = 10$, and *Up to* $K = 15$ present the average classification rate from 1 to 5, 10, and 15 bands, respectively. These three points of reference are used to test the behavior of each method from the first selected band to the 15th selected band, which is considered to be the transitory zone of the learning curve, before reaching the flat zone of the curve.

From this comparison, several interesting points arise.

1) The WaLuMI and WaLuDi methods generally obtained equal or better performance with respect to the rest of the methods in all databases. Therefore, regarding the band selection problem, where there exists high correlation among different features (image bands), the principle of looking for noncorrelated bands from the different regions of the spectrum by reducing the mutual information or a divergence measure between two bands has proved to be effective measures to obtain subsets of bands, from the point of view of pixel classification tasks.
2) From the experimental results, the Kullback–Leibler divergence that was used in the WaLuDi method and the mutual-information-based distance that was used in WaLuMI show quite a similar behavior. Kullback–Leibler divergence measures the cost of replacing one distribution by another one, while mutual information measures how much a random variable with a certain probability distribution can predict about another random variable. Although both concepts quantify different properties when comparing two probability distributions, because of the high correlation that image bands have in this context, the relative differences that were measured in both cases behave similarly, and they are not significant enough to obtain a noticeable change in the result after the clustering process. Thus, in the technique proposed here, the clustering process plays an important role.
3) It is worth remarking how important the methodology that was used to achieve the final set of selected bands is. Note that only the WaLuMI and WaLuDi methods involve a measure among the bands into a global strategy

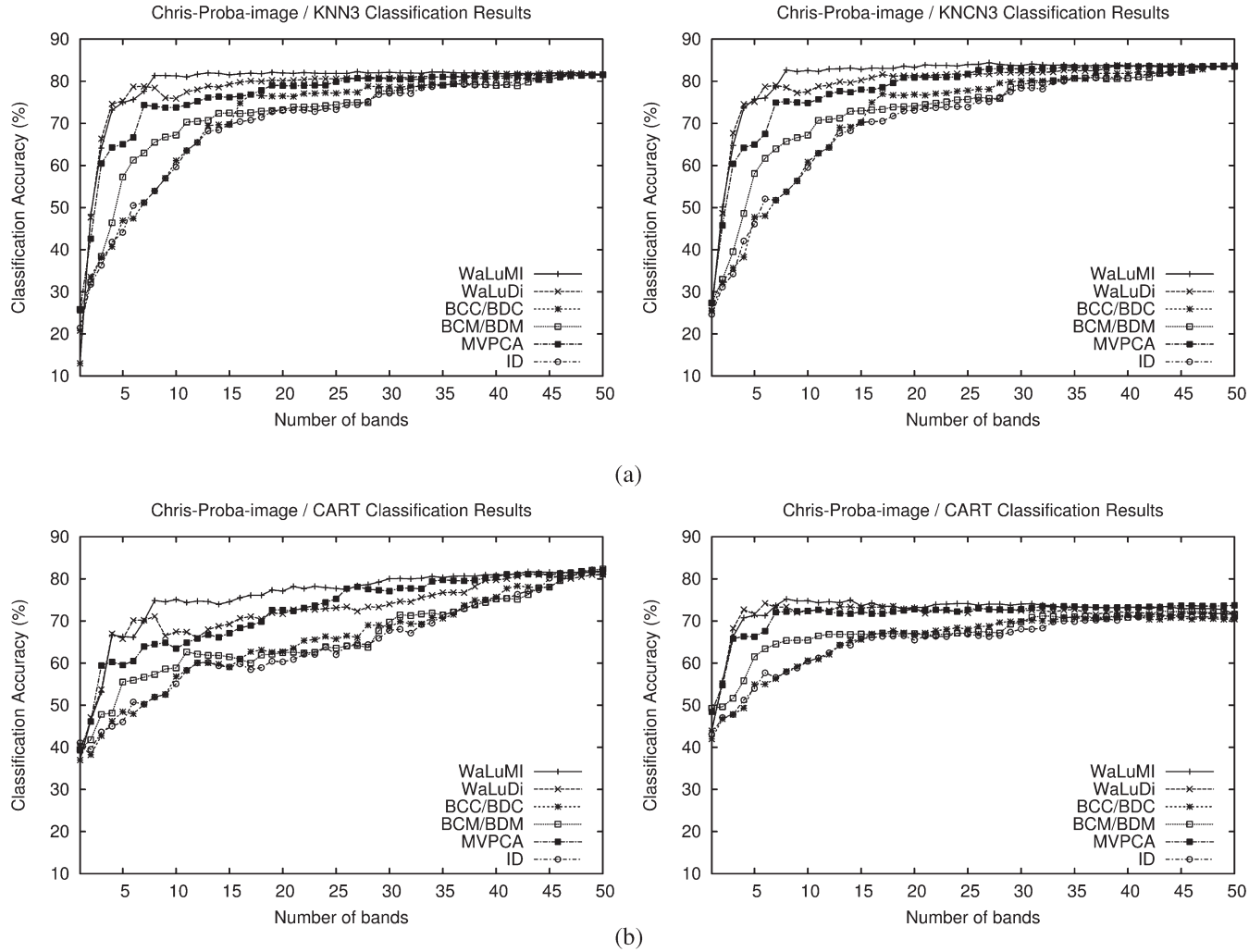

Fig. 8. (Top row) CHRIS–PROBA DB–KNN3 and KNCN3, and (bottom row) SVM (polynomial) and CART classifiers.

TABLE III
CLASSIFICATION ACCURACY FOR 92AV3C DB

KNN3	WaLuMI	WaLuDi	BCC/BDC	BCM/BDM	MVPCA	ID
Up to K=5	46.1968	45.3920	45.2952	**47.9512**	41.7004	32.2388
Up to K=10	54.3994	53.0554	54.2662	**54.5890**	48.6832	38.4402
Up to K=15	**58.2623**	57.2229	57.3911	58.2217	51.0749	41.3549
KNCN3						
Up to K=5	51.8656	50.9924	52.0228	**53.3480**	49.4236	41.5404
Up to K=10	**57.3562**	56.0056	57.6986	57.2758	52.2584	43.2708
Up to K=15	**60.4424**	59.2087	59.4807	59.9663	53.3100	44.5837
SVM						
Up to K=5	55.1720	54.3700	56.6920	**57.6940**	54.8440	50.7840
Up to K=10	57.0870	56.7870	**58.4230**	58.0460	53.7360	48.9640
Up to K=15	58.8907	57.8533	**58.8973**	58.7947	53.7140	48.5813
CART						
Up to K=5	51.8100	51.4000	52.1260	**53.7500**	50.1020	43.3180
Up to K=10	53.9140	53.3930	54.0120	**55.1140**	51.0740	44.3660
Up to K=15	55.0220	54.8080	54.5447	**55.8620**	51.4833	45.0373

of clustering. Thus, not only are mutual information and divergence measures adequate correlation or dependence measures, but the optimization process applied using a clustering strategy also acquires a special relevance. In fact, the robustness that was proven by these methods in all databases shows that the effectiveness of the final selected bands is probably due to the clustering process that was proposed since the selected bands are the final cluster representatives. The better these clusters, the more adequate the final K selected bands become.

4) LCMV–CBS methods and, particularly, the BCM/BDM method provided a slightly better performance in the 92AV3C image, although WaLuMI, WaLuDi, and even MVPCA achieve similar classification results in this

TABLE IV
CLASSIFICATION ACCURACY FOR HYMAP DB

KNN3	WaLuMI	WaLuDi	BCC/BDC	BCM/BDM	MVPCA	ID
Up to K=5	51.6708	**51.7908**	15.6685	15.6700	49.3344	19.3644
Up to K=10	67.1592	**67.7140**	26.1508	26.1516	66.8928	31.9862
Up to K=15	**74.0872**	74.0684	37.6936	37.6941	73.6891	45.0575
KNCN3						
Up to K=5	52.6732	**54.6012**	27.6176	27.6176	52.4198	29.5700
Up to K=10	67.1812	**69.0601**	34.0264	34.0264	67.8815	37.8408
Up to K=15	74.2511	**75.1045**	42.5148	42.5148	74.4265	48.6313
SVM						
Up to K=5	66.7060	**67.5720**	48.2520	48.2520	61.9540	48.7120
Up to K=10	72.6570	**73.1960**	54.8000	54.8000	70.3790	55.7240
Up to K=15	**75.3127**	75.1900	59.1540	59.1540	73.1600	61.6153
CART						
Up to K=5	67.7580	**68.0000**	41.8580	41.8580	63.3040	45.5040
Up to K=10	75.6130	**75.9980**	52.2950	52.2950	73.5900	56.9270
Up to K=15	79.1713	**79.2220**	59.3993	59.3993	77.5340	64.2093

TABLE V
CLASSIFICATION ACCURACY FOR VIS DB

KNN3	WaLuMI	WaLuDi	BCC/BDC	BCM/BDM	MVPCA	ID
Up to K=5	91.6732	**91.8176**	54.2684	43.6916	82.7620	34.3248
Up to K=10	95.6346	**95.6842**	76.8526	71.5120	91.1670	66.4028
Up to K=15	96.9599	**96.9904**	84.4017	80.8429	93.9879	77.4091
KNCN3						
Up to K=5	**99.1028**	98.9748	95.8700	94.7300	98.4540	93.4348
Up to K=10	**99.3544**	99.2672	97.6610	97.0616	99.0122	96.1330
Up to K=15	**99.4447**	99.3839	98.2815	97.8815	99.2211	97.2307
SVM						
Up to K=5	98.3760	**98.4300**	94.9020	93.8540	98.0260	91.9820
Up to K=10	98.5660	**98.6410**	96.7370	96.1550	98.4080	95.0050
Up to K=15	98.6560	**98.7333**	97.4193	96.9727	98.5680	96.2727
CART						
Up to K=5	**98.8820**	98.8320	94.7240	93.0800	98.4300	91.6900
Up to K=10	**99.1430**	99.0970	97.0080	96.0660	98.8750	94.7980
Up to K=15	**99.2293**	99.2033	97.7813	97.1433	99.0527	96.2427

TABLE VI
CLASSIFICATION ACCURACY FOR CHRIS–PROBA DB

KNN3	WaLuMI	WaLuDi	BCC/BDC	BCM/BDM	MVPCA	ID
Up to K=5	54.7268	**55.3940**	36.0116	40.0312	51.6092	35.0832
Up to K=10	**67.0872**	66.4850	45.0690	52.3860	62.0608	44.7650
Up to K=15	**71.9287**	70.4696	52.5717	58.6817	66.5905	52.1984
KNCN3						
Up to K=5	57.9368	**58.2844**	35.9004	41.3032	52.5360	35.6304
Up to K=10	**69.1882**	68.2460	45.0242	53.1710	63.0118	45.1588
Up to K=15	**73.7387**	72.0052	52.3892	59.3647	67.7236	52.3288
SVM						
Up to K=5	53,8780	**54.0900**	43.2200	46.5440	52.9240	43.0440
Up to K=10	**63.0050**	61.5760	47.5560	52.0010	58.1790	47.5770
Up to K=15	**66.8167**	63.6720	51.5173	55.3293	60.8320	51.5167
CART						
Up to K=5	61.6100	**62.3240**	48.1520	53.5760	60.3120	48.6320
Up to K=10	**67.7750**	67.6750	52.9840	59.1950	65.8140	53.4580
Up to K=15	**69.9700**	69.4467	56.5320	61.7220	67.9073	56.8360

image. However, the LCMV–CBS methods lacked consistency when HyMap or VIS databases were used where they achieved poorer results.

5) MVPCA has generally achieved a good performance, although it never achieves the best classification rate. One of its main characteristics is that it also shows quite a consistent behavior in all databases.
6) The ID method was the weaker one since its classification rates are quite poor compared with the other ones. This measure of non-Gaussianity seems to be unsuitable in performing band selection for classification tasks.
7) Although the best CPU time comes from the ID method, this is not relevant due to the poor results that were given by this method. It is important to distinguish the subset of methods that are computationally affordable, i.e., WaLuDi, WaLuMI, and MVPCA, from the LCMV–CBS methods, which are much more expensive in this sense.

IV. CONCLUSION

A technique for band selection in multi/hyperspectral images has been presented, which was aimed at removing redundant

information while keeping significant information for further classification tasks. The proposed method uses a clustering process strategy to group bands that minimizes the intracluster variance and maximizes the intercluster variance, in terms of some dissimilarity measures based on band information content, such as mutual information and Kullback–Leibler divergence.

The results that were obtained, from the point of view of pixel classification in hyperspectral images, provide experimental evidence about the importance that the proposed clustering strategy in a band dissimilarity space plays in the problem of classification.

The band selection method presented here is fully unsupervised, i.e., it uses no labeling information. It is computationally affordable, since it is based on probability estimations from histogram pixel values, as a maximum, from pairs of bands, avoiding unfeasible high-dimensional probability estimations.

The results in further pixel classification tasks demonstrate that the proposed technique has a more consistent and steadier behavior for different image databases and classification schemes, with respect to the other methods that were used in the experimental comparison. In particular, the method presented here exhibits a good behavior in the transitory zones of the learning curve (for small subsets of bands), where selecting the appropriate and significant bands is more critical, being a desirable property for a good band/feature selector in classification tasks.

ACKNOWLEDGMENT

The authors would like to thank Prof. C.-I Chang and, particularly, S. Wang, both from RSSIPL, for their help in the implementation of the LCMV–CBS and CEM–CBS methods.

REFERENCES

[1] S. Kumar, J. Ghosh, and M. M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1368–1379, Jul. 2001.

[2] S. B. Serpico and G. Moser, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 2, pp. 484–495, Feb. 2007.

[3] L. O. Jimenez-Rodriguez, E. Arzuaga-Cruz, and M. Velez-Reyes, "Unsupervised linear feature-extraction methods and their effects in the classification of high-dimensional data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 2, pp. 469–483, Feb. 2007.

[4] J. Wang and C.-I Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, Jun. 2006.

[5] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, Jul. 2001.

[6] L. Bruzzonne, F. Roli, and S. B. Serpico, "An extension to multiclass cases of the Jeffreys-Matusita distance," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 6, pp. 1318–1321, Nov. 1995.

[7] A. Plaza, P. Martinez, J. Plaza, and R. Perez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 466–479, Mar. 2005.

[8] C.-I Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006.

[9] C.-I Chang, Q. Du, T. L. Sun, and M. L. G. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2631–2641, Nov. 1999.

[10] I. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1265–1287, 2003.

[11] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ: Wiley, 1991.

[13] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *Proc. 23rd Eur. Colloq. Inf. Retrieval Res.*, 2001, pp. 191–200.

[14] R. Dosil, X. R. Fdez-Vidal, and X. M. Pardo, "Dissimilarity measures for visual pattern partitioning," in *Proc. IbPRIA*, 2005, vol. 3523, pp. 287–294.

[15] J. Aczel and Z. Daroczy, *On Measures of Information and Their Characterization*. New York: Academic, 1975.

[16] A. Webb, *Statistical Pattern Recognition*, 2nd ed. Hoboken, NJ: Wiley, 2002.

[17] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 236–244, Mar. 1963.

[18] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[19] A. Martinez-Uso, F. Pla, J. M. Sotoca, and P. Garcia-Sevilla, "Clustering-based multispectral band selection using mutual information," in *Proc. ICPR*, 2006, pp. 760–763.

[20] C.-I Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. New York: Plenum, 2003.

[21] C.-I Chang, "Target signature-constrained mixed pixel classification for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 5, pp. 1065–1081, May 2002.

[22] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.

[23] L. O. Jimenez and A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2653–2667, Nov. 1999.

[24] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[25] J. S. Sánchez, F. Pla, and F. J. Ferri, "Improving the k-NCN classification rule through heuristic modifications," *Pattern Recognit. Lett.*, vol. 19, no. 13, pp. 1165–1170, Nov. 1998.

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

[27] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, 2001. [Online]. Available: www.csie.ntu.edu.tw/~cjlin/libsvm

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

[29] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. New York: Chapman & Hall, 1984. New edition.

[30] T. Mitchell, *Machine Learning*. New York: McGraw-Hill, Oct. 1997. McGraw-Hill Education (ISE Editions).

Adolfo Martínez-Usó received the B.Sc. degree in computer science from Universitat Jaume I, Castellón de la Plana, Spain. He is currently working toward the Ph.D. degree in computer science engineering at the Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I.

His current research interests are hyperspectral image analysis, particularly, band selection and image segmentation.

Filiberto Pla received the B.Sc. and Ph.D. degrees in physics from the University of Valencia, Valencia, Spain, in 1989 and 1993, respectively.

He is currently a Full Professor in the Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Castellón de la Plana, Spain. He has been a Visiting Scientist at several universities and research centers in the U.K., France, Italy, Portugal, and Switzerland. He has authored more than 100 scientific papers in the fields of computer vision and pattern recognition. He has also been a coeditor of two books and acted as reviewer for several international journals in the field of computer vision and pattern recognition. His current research interests are color and spectral image analysis, visual motion analysis, active vision, and pattern recognition techniques applied to image processing.

Prof. Pla is a member of the International Association for Pattern Recognition.

José Martínez Sotoca received the B.Sc. degree in physics from the Universidad Nacional de Educación a Distancia, Madrid, Spain, in 1996 and the M.Sc. and Ph.D. degrees in physics from the University of Valencia, Valencia, Spain, in 1999 and 2001, respectively. His Ph.D. work was on surface reconstructions with structured light.

He is currently an Assistant Lecturer in the Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Castellón de la Plana, Spain. He has collaborated in different projects, most of which are in the medical application of computer science. He has published more than 45 scientific papers in national and international conference proceedings, books, and journals. His research interests include pattern recognition and biomedical applications, including image pattern recognition, hyperspectral data, structured light, and feature extraction and selection.

Pedro García-Sevilla received the B.Sc. degree from the University of Castilla-La Mancha, Albacete, Spain, in 1990, the M.Sc. degree from the University of Malaga, Malaga, Spain, in 1992, and the Ph.D. degree from Universitat Jaume I, Castellón de la Plana, Spain, in 1999, all in computer science.

He is currently with the Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, as a Lecturer, where he has been with the Computer Vision Group since it was created. His research interests are image processing and pattern recognition, particularly, texture analysis, hyperspectral imaging, industrial inspection, and medical imaging.