Band Selection in Multispectral Images by Minimization of Dependent Information

José Martínez Sotoca, Filiberto Pla, and José Salvador Sánchez, *Member, IEEE*

***Abstract*—In this paper, a band selection technique for hyperspectral image data is proposed. Supervised feature extraction techniques allow a reduction of the dimensionality to extract relevant features through a labeled training set. This implies an analysis of the existing class distributions, which usually means, in the case of hyperspectral imaging, a large number of samples, making the labeling process difficult. A possible alternative could be the use of information measures, which are the basis of the proposed method. The present approach basically behaves as an unsupervised feature selection criterion, to obtain the relevant spectral bands from a set of sample images. The relations of information content between spectral bands are analyzed, leading to the proposed technique based on the minimization of the dependent information between spectral bands, while trying to maximize the conditional entropies of the selected bands.**

***Index Terms*—Band selection, feature selection, information theory, multispectral images.**

I. Introduction

IN CERTAIN application fields where visual information processing is involved, the use of spectral information is of most importance to perform certain tasks, e.g., in remote sensing, medical imaging, fine arts, product quality assessment, etc. The trend for these systems is the use of spectral and spatial information, i.e., hyperspectral image representations, to estimate and analyze the presence of chemical compounds, pathologies, or other information, providing a qualitative and quantitative evaluation of those features.

Multispectral images are a kind of multimodality, where spectral imaging is combined with digital image processing [4], [17], [22]. While the images produced by usual digital cameras contain the intensity, or some color representations (e.g., RGB), multispectral images provide spectral information for each pixel in a wavelength range with a given spectral resolution. In our experimental work, each spectral image band is captured at each selected wavelength with a narrow bandpass filter, allowing a multiband representation for each pixel in a spectral range.

When having available multispectral data, a common question to be solved is how to select the right spectral bands to characterize the problem. When dealing with multispectral images, the amount of information to be treated can be very large. Moreover, the spectral information is highly correlated along a given spectral range; therefore, instead of having an exhaustive representation of the whole spectrum, selecting some key bands can considerably reduce the amount of data without practically losing relevant information [20].

Considering multispectral images contain information represented by means of a set of bidimensional signals, the band selection problem in multispectral images could be addressed from the point of view of information theory. Prior work using information theory concepts has been done in pattern recognition for simulated data and benchmark database techniques [5], [19], [27] and in medical applications [28].

The main objective of band selection in multispectral imaging is getting rid of redundant information and reducing the amount of data to be processed. Therefore, from the point of view of pattern recognition, we would be interested in feature selection [6], [15], [16] rather than in feature extraction [12], [18]. For instance, obtaining a new set of reduced image representations from a linear combination of the whole set of original image bands is not desirable, since we would need the total amount of information to obtain the new features. On the other hand, selecting a subset of relevant bands from the original set allows the process of image acquisition to be reduced to a certain number of bands instead of dealing with the whole amount of data, making simpler the image acquisition and analysis.

When analyzing spectral information with supervised methods, it is necessary to fix beforehand the number of classes present in the images, and the adequate number of training samples for a given number of features or spectral bands [11], which are usually large. The supervised feature selection algorithms establish a relevance criterion according to class separability measures for different subsets of features.

In the framework of multispectral imaging, another possible answer to the problem of feature selection would be using an unsupervised approach. One way to solve it consists of grouping the data in the feature space by using clustering techniques [2], [7], [13], [21]. Another approach is to minimize the classification error by selecting bands that provide the highest image contrast [10]. In this work, our approach consists of analyzing the amount of information in a subset of features (bands), measuring the degree of independence between image bands as a relevance criterion. Thus, by means of an unsupervised process, the connections of information between each subset of image bands are analyzed, leading to a criterion that allows the search of spectral bands aiming at removing the redundant information and, at the same time, trying to maximize the amount of information in the selected bands.

The organization of the rest of this paper is as follows. Section II provides a discussion about higher order information

Manuscript received March 8, 2004; revised August 8, 2004. This work was supported in part by Grants IST-2001-37306 (IST Project European Union), DPI2001-2956-C02-02 (Spanish CICYT), TIC2003-08496 (Spanish CICYT), and P1 1B2004-08 (Fundacio Caixa-Castello).

The authors are with the Department of Lenguajes y Sistemas Informáticos, Universitat Jaume I, 12071 Castellón, Spain (e-mail: sotoca@lsi.uji.es; sanchez@lsi.uji.es).

Digital Object Identifier 10.1109/TSMCC.2006.876055

measures and their implications in the band selection problem. Section III describes the proposed criterion to measure the dependent information among bands. Section IV describes the problematic about multidimensional probabilities of the different events and their computational costs. Section V reports the empirical results, including the classification performance with respect to the number of bands obtained, comparing them with some supervised approaches. Eventually, some concluding remarks are depicted in Section VI.

II. Information in Multispectral Images

Let us consider an ensemble of image bands $A_1, \ldots, A_n$, where A_i is a random variable representing the image band i. Thus, the amount of information contained in a multispectral image can be expressed as the joint entropy $H(A_1, \ldots, A_n)$, that is

$$H(A_1, \ldots, A_n) = \sum_{a_1, \ldots, a_n} p(a_1, \ldots, a_n) \log_2 \frac{1}{p(a_1, \ldots, a_n)} \tag{1}$$

where $p(a_1, \ldots, a_n)$ represents a joint probability distribution. The term $\log_2 1/p(a_1, \ldots, a_n)$ means that the amount of information gained from an event with probability $p(a_1, \ldots, a_n)$ is inversely related to the probability that takes place in this event. The rarer is an event, the more meaning is assigned to the occurrence of this event. Therefore, the information per event is weighted by the probability of its occurrence. The resulting entropy term is the average amount of information gained from a set of possible events.

In the case of multispectral imaging or, in general, multimodal images, the joint probability distribution $p(a_1, \ldots, a_n)$ can be estimated as [25]

$$p(a_1, \ldots, a_n) = \frac{h(a_1, \ldots, a_n)}{MN} \tag{2}$$

where $h(a_1, \ldots, a_n)$ is the joint gray-level histogram of bands $A_1, \ldots, A_n$, and the normalizing factor, MN (M columns and N rows) is the image size, assuming all image bands with equal size.

Mutual information $H(A_1 : \ldots : A_n)$ is a basic concept in information theory [1]. It measures a certain type of dependence between random variables. A general expression of the mutual information [24] can be obtained from

$$H(A_1 : \ldots : A_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < \cdots < i_k} H(A_{i_1}, \ldots, A_{i_k}) \tag{3}$$

where the sum $\sum H(A_{i_1}, \ldots, A_{i_n})$ runs over all possible combinations $\{i_1, \ldots, i_k\} \in \{1, \ldots, n\}$. The generalized mutual information may be interpreted as the common information shared among n random variables.

Some interesting properties of the mutual information and related measures are the following.

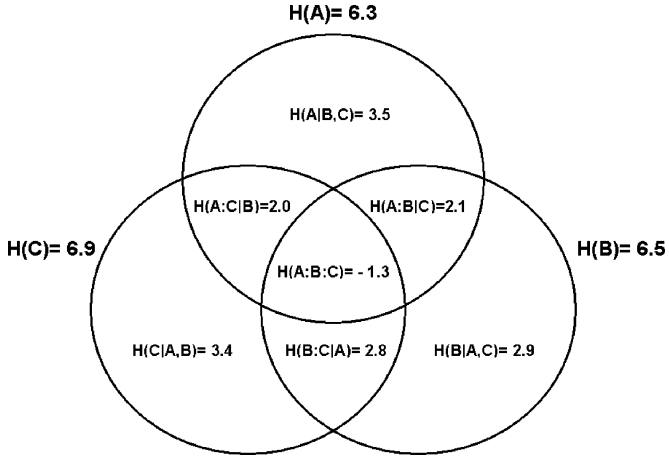


Fig. 1. Entropy Venn diagrams (each circle denotes the entropy of an image band. Joint entropy is the union of circles) for image bands with wavelengths: 450, 540, and 580 nm.

- $H(A_1 : \ldots : A_n)$ is symmetric under any permutation of $A_1, \ldots, A_n$.
- The entropies of an ensemble of image bands satisfy the following inequality:

$$H(A_1 : \ldots : A_n) \leq \min\{H(A_1), \ldots, H(A_n)\}. \tag{4}$$

On the other hand, the conditional entropy $H(A_{k+1} \mid A_1, \ldots, A_k)$ represents the amount of independent information in image band A_{k+1} having measured the rest image bands $A_1, \ldots, A_k$. It can be expressed in terms of joint entropies as

$$H(A_{k+1} \mid A_1, \ldots, A_k) = H(A_1, \ldots, A_{k+1}) - H(A_1, \ldots, A_k). \tag{5}$$

From the previous expression, it can be deduced that $0 \leq H(A_{k+1} \mid A_1, \ldots, A_k) \leq H(A_{k+1})$, which means that the larger $H(A_{k+1} \mid A_1, \ldots, A_k)$, the nearer to $H(A_{k+1})$, i.e., the more independent information provides the random variable A_{k+1}, being the limit $H(A_{k+1})$. Thus, when the condition is equal to the entropy of a random variable, this variable is fully independent.

An incremental way to compute mutual information, when adding a new random variable A_n to a given set of random variables $A_1, \ldots, A_{n-1}$, can be expressed as follows:

$$H(A_1 : \ldots : A_n) = H(A_1 : \ldots : A_{n-1}) + H(A_n) + \sum_{k=1}^{n-1} (-1)^k \sum_{(i_1 < \ldots < i_k) \neq n} H(A_{i_1}, \ldots, A_{i_k}, A_n). \tag{6}$$

Relationships among entropies can be conveniently represented by entropy Venn diagrams [24]. For instance, in Fig. 1, the entropy diagram with variables A, B, C has the following measures: $H(A \mid B, C)$, $H(B \mid A, C)$, $H(C \mid A, B)$ are the conditional entropies; $H(A : B \mid C)$, $H(B : C \mid A)$, $H(A : C \mid B)$ are the conditional informations; and $H(A : B : C)$ is the mutual information. All these measures can be expressed as a function of the seven entropies $H(A)$, $H(B)$, $H(C)$, $H(A, B)$, $H(A, C)$, $H(B, C)$, and $H(A, B, C)$. Therefore, estimating the

different single and joint entropies of a random variable set allow us to estimate any of the conditional or mutual information measures.

It is well known that all the above-mentioned information measures are positive, except for $H(A:B:C)$, which can be negative (see Fig. 1). Indeed, the monotonicity of Shannon entropies implies that conditional entropies, such as $H(A \mid B, C)$, are positive definite. Analogously, the conditional informations, such as $H(B:C \mid A)$, are nonnegative.

About the mutual information, it can be characterized by positive or negative values of $H(A:B:C)$. Therefore, if $H(A:B:C) > 0$, the mutual information $H(A:B)$ should contain part of the mutual information $H(A:C)$. On the other hand, if $H(A:B:C) < 0$, then the relation of pairs of variables is simultaneously unsatisfied. This situation is called *correlation frustration* [24]. The existence of negative values of higher orders of mutual information proves the existence of higher order correlations in the conditional informations.

One of the objectives of the present work is to develop a computationally affordable method to deal with the information measures needed. The mutual information is not a desirable measure we would need for our aims, mainly because it does not represent all possible types of dependent information. On the other hand, from the computational point of view, a simpler information measure would be more adequate.

Thus, a measure that can be worked out in an incremental way would be desirable, e. g., the conditional entropy. This measure allows to look at the problem from the opposite point of view, i.e., instead of looking for all types of dependent information among the random variables, the objective would be estimating the independent information a random variable provides, given the rest of variables. Focusing on measures of independence rather than correlation measures, leads us to use conditional entropies.

In order to establish a computational approach, the conditional entropy is less complex than the mutual information to be estimated from the experimental data [see (5) and (6)], in this case, from multispectral images. Thus, the proposed criterion to measure the relevance of a set of features will be based on conditional entropies.

III. Estimating Discriminant Information

According to [18], a multispectral band selector algorithm should have the following desirable properties.

- *Class dependence:* Different subsets of classes can be better discriminated by different feature sets.
- *Ordering constraint:* The fact that bands are ordered and adjacent bands are correlated should be exploited.
- *Discriminating transforms:* Any transformation of the initial feature set should try to maximize discrimination among classes, and thus, use class label information.

The first and third properties are related with supervised information, i.e., they are desirable properties for a band selection algorithm that tries to exploit the class label information. In the approach presented here, the aim is not using supervised information, to avoid the labeling process. Therefore, class information is not available when looking for a relevant subset of image bands.

The concepts of discriminant measures, when having the data structured in classes, have been substituted by information theory concepts such as amount of information and correlation measures. From this point of view, we will consider that a set of bands is more discriminant with respect to other subset of bands when the following occur.

- They contain more information.
- They are less correlated.

Therefore, from the desirable properties described above, we will exploit the fact that consecutive bands are correlated using information measures, and we will try to maximize the amount of information instead of looking for class discrimination. This way to tackle the problem will allow to deal with unlabeled data. Moreover, we will show in the experimental results that, approaching the problem in such a way also provides satisfactory results with respect to supervised feature selection criteria when looking for best classification rates.

An open question is how to define the dependent information among a given subset of image bands. One way could be measuring the mutual information among the image bands. In other problems involving image data, as in image registration, the goal is maximizing the mutual information to calculate the correspondence between multimodal images [25]. Band selection in multispectral images supposes the opposite problem.

Such as described before, in the band selection problem, we look for a subset of bands where the following occurs,

- The subset contains as much information as possible with respect to the whole multispectral image.
- The information each band represents has to be as less correlated as possible with respect to the other ones, not to have redundant information present in different bands.

The amount of joint information provided by a set of image bands $A_1, \ldots, A_n$ is represented by its joint entropy $H(A_1, \ldots, A_n)$. On the other hand, to make this information as discriminative as possible, the random variables that represent this joint entropy should be as independent as possible. Any type of correlation among the random variables would imply a decrease in the total amount of joint information.

The concept of *total correlation* [29] measures the total amount of dependence among random variables and can be defined as follows [23], [26]:

$$\phi_{\mathrm{DI}} = \left(\sum_{i=1}^{n} H(A_i) \right) - H(A_1, \ldots, A_n)$$

$$= \sum_{a_1, \ldots, a_n} p(a_1, \ldots, a_n) \log_2 \frac{p(a_1, \ldots, a_n)}{\prod_{i=1}^{n} p(a_i)}. \quad (7)$$

For the case of three image bands, this measure of dependence corresponds to the shaded area in Fig. 2 (darker area is counted twice). This expression measures all relations of redundant information that may exist in the ensemble of random variables, and it is nonnegative. One problem of this measure is that it does not allow a measure of the region representing the dependent information, since part of the dependent information

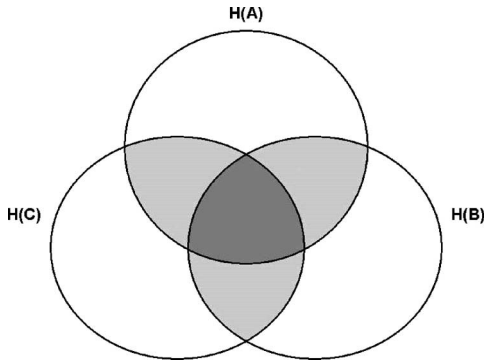


Fig. 2. Shared information among three image bands. The dark gray region corresponds to the dependent information obtained by (7).

is considered several times. Thus, this estimation can lead to biased values of some of the properties we are interested in.

A possible alternative to measure the dark region of Fig. 2, representing the dependent information, is considering the total information contained in a set of image bands, and subtracting the independent information. Therefore, if we discard the independent information, the rest of the information contained in a set of random variables can be considered as the union of all possible dependencies and connections of information among the image bands. The conditional entropies provide a measure of the independent information of an image band, given the rest of the image bands [nonshaded areas in Fig. 3(a)]. This fact will be the basis of our approach.

From the rationale described in the previous paragraph, a measure of the dependent information of a set of random variables could be defined by the following expression:

$$\Theta_{\mathrm{DI}} = H(A_1, \ldots, A_n) - \sum_{i_1=1}^{n} H(A_{i_1} \mid A_{i_2}, \ldots, A_{i_n}) \quad (8)$$

where $A_{i_2}, \ldots, A_{i_n}$ are the complementary variables of A_{i_1}; $H(A_1, \ldots, A_n)$ is the joint entropy, which represents the total amount of joint information of set of variables $A_1, \ldots, A_n$; and $H(A_{i_1} \mid A_{i_2}, \ldots, A_{i_n})$ is the conditional entropy of A_{i_1}, given $A_{i_2}, \ldots, A_{i_n}$. In the case of three [see Fig. 3(a)] image bands, the shaded area would correspond to such a measure.

The selection of a subset of bands that minimizes the above criterion function would provide a subset of bands with minimum interdependence, trying to keep as much information as possible. This criterion will be called hereafter the *minimization of the dependent information* (MDI).

Given a certain selection criterion, looking for an optimal subset of image bands (features) is a combinatorial problem with exponential cost with respect to the number of features [16]; thus, an exhaustive search for large number of features becomes unaffordable. There exist several approaches in the literature to look for subsets of features that minimize a criterion function. A simple strategy is the use of a sequential forward scheme as follows.

- *Input*: $A = \{A_1, \ldots, A_n\}$, the initial set of image bands; N is the number of bands to be selected.
- *Output*: B is the subset of bands selected, such as $|B| = N$.

Step 1) *Initialization*: Select the two image bands (A_i, A_j) with the lowest mutual information, being $H(A_i : A_j) = \min H(A_k : A_l)$ for all possible pairs of image bands (A_k, A_l). Initialize the selected pair of bands $B = \{A_i, A_j\}$.

Step 2) *Evaluation*: For the remaining bands, calculate for each band A_l the MDI criterion function (8), considering the subset of bands $(B \cup A_l)$.

Step 3) *Inclusion*: Choose the image band A_k that obtains the highest reduction of the MDI measure. Add this band to the selected subset $B = B \cup A_k$.

Step 4) Go back to Step 2) while $|B| < N$.

When increasing the number of image bands during the process, the joint entropy of the subset of selected image bands keeps growing, increasing the total amount of joint information, this increment being faster during the first selected image bands [see Fig. 3(b)]. On the other hand, the sum of their conditional entropies decreases with respect to the joint entropy, increasing the difference between the values of the two curves.

For a given number of selected bands, the difference between the two curves [Fig. 3(b)] measures the dependent information of the subset of spectral bands. For each image band included in the selected subset [Step 3) of the previous algorithm], the procedure described tries to keep this difference as minimum as possible when including a new image band.

Some important properties of the proposed MDI criterion can be expressed by the following lemmas.

Lemma 1: The dependent information defined by the MDI criterion is definite positive, since it always holds that

$$H(A_1, \ldots, A_n) \geq \sum_{i_1=1}^{n} H(A_{i_1} \mid A_{i_2}, \ldots, A_{i_n}). \quad (9)$$

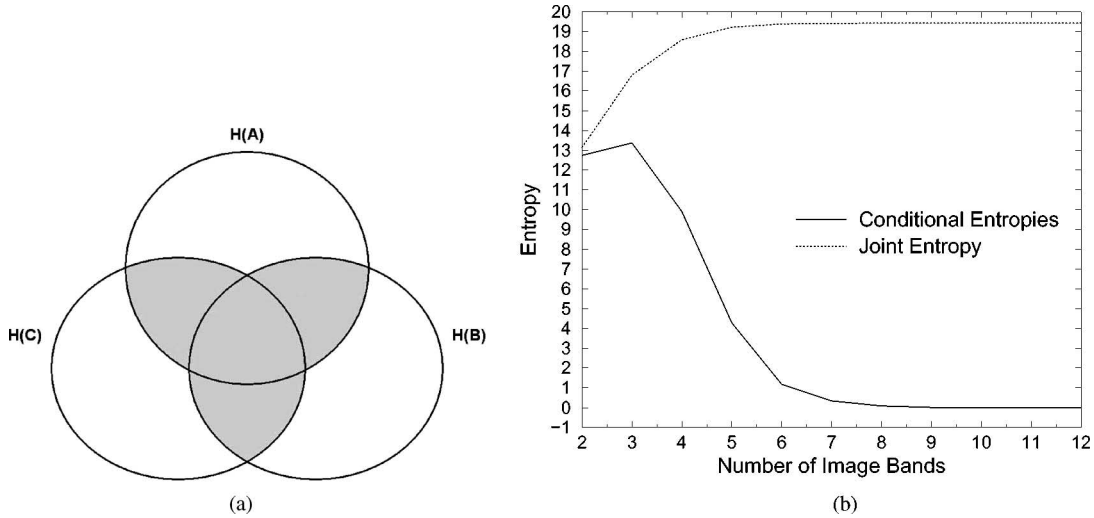
Lemma 2: If the dependent information is null, i.e., the MDI function is equal to zero, then the random variables are fully independent among them. This conclusion can be obtained in a straightforward way, since in such a case, the total amount of joint information (joint entropy) coincides with the conditional information, i.e., there is no redundant information among the random variables.

IV. Computational Complexity and Implementation Issues

From the computational point of view, calculating any of these information measures depends on the cost of the joint entropy computation, since any of these measures can be estimated from the joint entropy of different bands (see Section II).

Apart from the computational time, the main drawback in the calculation of joint entropy is in memory requirements. If we estimate the joint entropy of N bands from the cojoint histogram, we would need an array of N dimensions to allocate enough bins for all possible values of gray level co-occurrences in N bands, i.e., the spatial cost in this case would be exponential with respect to N.

In general, joint histograms for multispectral images are sparse, and the higher the dimensionality, the more sparse



(a) (b)

Fig. 3. (a) Shared information among three image bands. Shaded areas correspond to the dependent information obtained by (8). (b) Behavior of the sum of conditional entropies versus the joint entropy.

are the histograms. This is due to the fact that, given an image size, the number of possible different co-occurrences is bounded by the number of pixels in the image, and it does not depend on the number of bands. We can take advantage of this fact to reduce the spatial complexity of the joint entropy calculation.

To avoid this spatial complexity, the cojoint histogram is represented by a dynamic list of co-occurrences present in the image. An index is assigned to each co-occurrence, in this case a string, and the list is ordered according to this index. When calculating the the cojoint histogram, for a given co-occurrence value, it is searched in the list. If it is in the list, the corresponding bin is incremented; otherwise, a new element is inserted in the list with its corresponding index.

Therefore, the spatial complexity is significantly reduced, since the memory requirements only stands for the effective number of possible co-occurrences present in the image, which is bounded by the number of pixels in the image, avoiding the problem of sparse arrays.

With respect to time complexity, the cost of calculating the cojoint histogram depends on the cost of building and updating the list, which is basically the cost of searching and either inserting or increasing the value of an already existing element in the list. A binary tree data structure has been used in this case to represent the list, and the search procedure is performed using a binary search strategy. Therefore, the cost of estimating the joint entropy is of the order $O(N_{\mathrm{p}}\log N_{\mathrm{p}})$, being N_{p} the number of elements in the list. In the case of a single image, the upper bound of N_{p} is the size of the image in pixels.

Eventually, to calculate the criterion proposed in (8), the number of joint entropies to be calculated in each subset is $O(k+1)$, where k is the size of the subset of bands to estimate. Thus, if we are looking for a subset of k bands, the overall computational time to estimate the MDI criterion is of the order $O((k+1)N_{\mathrm{p}}\log(N_{\mathrm{p}}))$. The spatial cost is $O(N_{\mathrm{p}})$, the cost of building the cojoint histogram representation, which is linear with respect to the number of pixels.

V. Empirical Results

The collection of multispectral images used in the experiments was obtained by an imaging spectrograph (Retiga-Ex, Opto-knowledged Systems Inc., ON, Canada). The spectral range extended from 400 to 720 nm in the visible (VIS), with a spectral resolution of 10 nm, obtaining a set of 33 spectral bands for each image.

The image database consisted of 19 multispectral images corresponding to orange fruits, of 280×280 pixel size, with different types of defects and skin variations on their surfaces. To analyze the performance of the approach presented here, two types of data sets were built.

The first data set consisted of multispectral images with unlabeled data. The information for each pixel of the 33 image bands was stored as an unlabeled sample in a feature vector of 33 dimensions. Different data sets were built with one, four, seven, and nine fruit images. The proposed algorithm based on the MDI criterion was applied directly in each database as an unsupervised method to select a subset of bands, with the aim of selecting the subset of bands with higher discrimination power to detect the different types of skin areas and defects on the fruits' surface. This method assigns higher weights when more relevant is a feature.

The second data set consisted of a set of 135 540 pixels from the 19 multispetral images, manually labeled as representative examples of the different regions present in the image database, considering five different classes. One of the classes represents the healthy orange skin, and the other four classes correspond to different typologies of defects: scratch, trip, insect bite, and overripe (see Fig. 4). This data set was divided into two, a partition containing 94 875 instances for training and 40 665 instances for test. This data set was used to compare the proposed MDI band selection method with other supervised approaches.

To assess the performance of the method, a nearest neighbor (NN) classifier was used to classify pixels into the different classes established for defects and fruit surface types. The

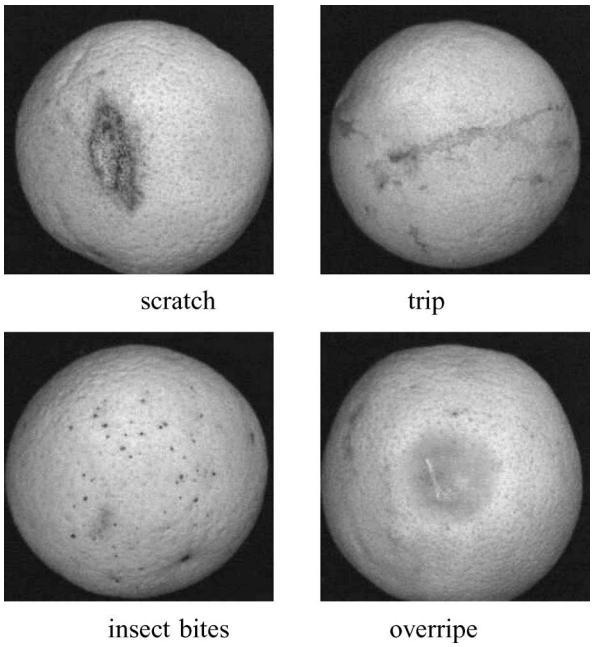


Fig. 4. Pathologies on the orange skin that appear in the multispectral image data set used.

performance of the NN classifier was considered as the validation criterion to compare the significance of the subsets of selected image bands obtained by the proposed approach and all the other supervised methods considered in the experiments carried out. To increase the statistical significance of the results, the average values over five random partitions were estimated. The samples in each partition were randomly assigned to the training and test sets as previously described.

A. Comparison With Supervised Feature Selection Criteria

To analyze the accuracy of the ranking of bands obtained by the presented approach, four supervised filter feature selection methods were also tested. Thus, the band selection process was considered as a supervised feature selection approach, in this case, using the labeled data set for the feature selection process.

The main motivation about comparing the proposed method with supervised approaches is that the labeled data set contains information about the distribution of classes existing in the hyperspectral data, and they allow the search for relevant feature subsets, with the aim of looking for a better class separability. Comparing the performance of those approaches, we can measure the capability of obtaining subsets of relevant features (image bands) by the introduced MDI approach without a prior knowledge of the class distributions in the multispectral image data.

The first method used in this comparative study is the well-known *ReliefF* algorithm [14], based on pattern distances. This algorithm initializes every feature weight to zero and then iterates m times looking for a set of feature weights that optimizes a criterion function.

The procedure begins by randomly selecting a sample x from the data set. For the selected sample, it determines its nearest neighbor prototype of the same class p^{hit} (nearest hit) and the nearest neighbor prototype of a different class p^{miss} (nearest miss). The algorithm updates each feature weight f_i according to the following criterion:

$$f_i^m = f_i^{m-1} - \frac{\text{diff}(x_i, p_i^{\text{hit}})}{m} + \sum_{c \neq \text{class}(x)} \frac{p(c)\text{diff}(x_i, p_i^{\text{miss}})}{m} \tag{10}$$

where $p(c)$ is the prior probability of class c and $diff(,)$ is the distance between the sample and the prototype for feature i. This algorithm was chosen because of its widespread use and good performance in general feature selection problems.

The second technique used is a multiple discriminant analysis for *n*-dimensional spaces and c classes. Let us suppose that the *total mean vector* $\boldsymbol{m}$, the *class mean vector* $\boldsymbol{m}_h$, and the number of elements of each class n_h have been calculated. We define the *between-class* covariance matrix $\mathbf{S}_B$ and the *within-class* covariance matrix $\mathbf{S}_W$ through the following expressions:

$$\mathbf{S}_B = \sum_{h=1}^{c} n_h(\boldsymbol{m}_h - \boldsymbol{m})(\boldsymbol{m}_h - \boldsymbol{m})^t \tag{11}$$

$$\mathbf{S}_W = \sum_{h=1}^{c} \sum_{x} (x - \boldsymbol{m}_h)(x - \boldsymbol{m}_h)^t. \tag{12}$$

A transformation matrix $\mathbf{w}$ that maximizes the ratio between the *between-class* covariance matrix and the *within-class* covariance matrix can be calculated by applying a statistical Fisher criterion [8]

$$J(W) = \frac{w^t S_B \mathbf{w}}{w^t S_W \mathbf{w}}. \tag{13}$$

In this work, we have applied the *Wilk's Lambda* test λ_W [9] as a measure of significance in multiple discriminant analysis for each subset of features. This measure takes values between 0 and 1. The nearer to 0 is λ_W, the higher discriminant power has the set of variables considered. λ_W can be defined through the following expression:

$$\lambda_W = \frac{|\mathbf{S}_W|}{|\mathbf{S}_W + \mathbf{S}_B|}. \tag{14}$$

Fisher-based discriminant has been choosen because it is a technique that has been used in some works [18], [30] for band selection in hyperspectral images.

The third technique is related to divergence measures between classes. One of the best-known distance measures utilized for feature selection in multiclass problems is the average Jeffries–Matusita (JM) distance [6]

$$\text{JM} = \sum_{h=1}^{c} \sum_{k>h}^{c} P_h P_k \text{JM}_{hk} \tag{15}$$

where

$$\mathrm{JM}_{hk} = \sqrt{2(1 - e^{-b_{hk}})}$$

$$b_{hk} = \frac{1}{8}(m_h - m_k)^t \left(\frac{\mathbf{S}_W^h + \mathbf{S}_W^k}{2} \right)^{-1} (m_h - m_k) + \frac{1}{2} \ln \left(\frac{\left| \frac{S_W^h + S_W^k}{2} \right|}{\sqrt{|\mathbf{S}_W^h|\,|\mathbf{S}_W^k|}} \right)$$

where P_i is the prior probability of class i; b_{hk} is the Bhattacharyya distance between classes h and k; and $\mathbf{S}_W^i$ and $\boldsymbol{m}_i$ are the covariance matrix and the mean vector of class i, respectively. In terms of class separability, the higher is the JM distance between two classes, the more separability between them.

The last technique considered in this comparison study is based on the mutual information feature selection (MIFS) algorithm, which uses as input the feature distributions and the classes distributions [3]. In this approach, the output class is treated as a random variable. At each step, the best feature A_i selected by the MIFS algorithm satisfies the following expression:

$$A_i = \max_{A_i} \left(H(C : A_i) - \beta \sum_{j=1}^{i-1} H(A_i : A_j) \right) \tag{16}$$

where C is the output class label; A_i is the ith feature considered, given a subset of $A_1, \ldots, A_{i-1}$ previous selected features; and β is a tunable parameter. If $\beta = 0$, the mutual information between input features is not taken into account and the algorithm selects the features considering the mutual information between input features and output classes. This criterion greedly selects the set of features with higher mutual information within the output classes, while trying to minimize the mutual information among the features selected. It has been chosen for comparison with the MDI-proposed method due to the fact it also uses information measures to select features, although in a supervised manner.

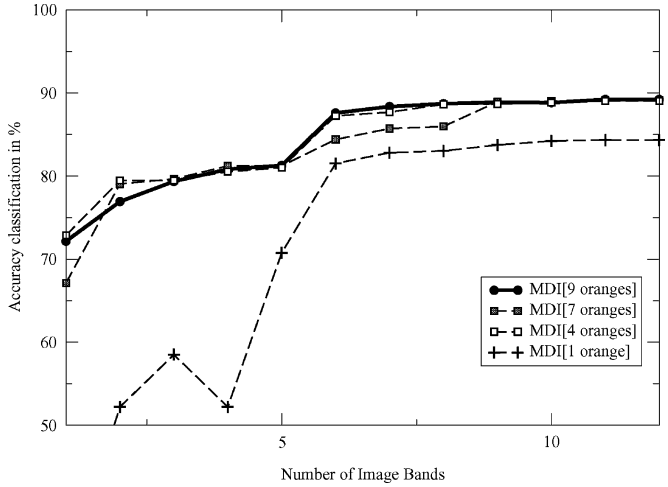


Fig. 5. Classification rate of the NN rule using the selected bands obtained by MDI[one orange], MDI[four oranges], MDI[seven oranges], and MDI[nine oranges].

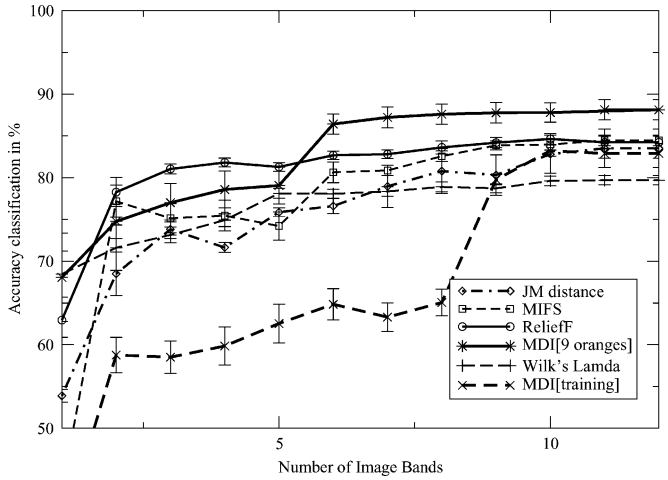


Fig. 6. Classification rate of the NN rule using the selected bands obtained by MDI[nine oranges], MDI[training set], JM distance, Wilks's Lambda, MIFS $\beta = 1{:}0$ and *ReliefF*.

B. Performance Evaluation

All experimental results shown in this section about classification rates correspond to the average classification accuracy obtained by the NN classifier over the five random partitions, such as described at the beginning of Section V.

To see the influence of the amount of data used to select the image bands in the MDI criterion, the performance of the proposed method has been tested with a variable number of multispectral images. Therefore, the plot of Fig. 5 shows the performance when using one, four, seven, and nine images as input data. A clear improvement of the classification accuracy can be observed from four to nine images.

The poor performace using only one or few images is due to the fact that, in a single image, there are not enough pixels that represent to all possible pixel classes considered, i.e., there may not appear enough pixels representing the different defects and orange skin types considered. When more image samples are used, if they contain enough information about all possible pixel values, the bands selected perform better. We can note a significant improvement from one to four images (Fig. 5). On the other hand, once the image data set contains a statistically significant amount of data representing all possible classes, there is no notable improvement in the bands selected. From the information theory point of view, this is because, from a given point, adding more data to a set of images, does not significantly increase the total amount of cojoint information.

Fig. 6 shows the performance of the proposed MDI method compared with the feature selection algorithms introduced in Section V-A. The plot represents the classification rate with respect to the subset of N bands selected. To show the statistical significance of these figures, the error bars represent the standard deviation of the classification rate obatined over the five random partitions mentioned at the end of Section V.

Note (see Fig. 6) that the proposed MDI method obtained from the data of nine images (MDI[9 oranges]) clearly outperforms the accuracy of the NN classifier with respect to the rest of methods as previously mentioned. This behavior is more notable

TABLE I
COMPUTATIONAL COST WHEN SELECTING 12 FEATURES

Criteria	Time (in minutes)
ReliefF	374
JM distance	29
Wilk's λ	13
MIFS	512
MDI[training]	139

from a certain number of bands, particularly from four image bands. The performance of the MDI method applied over the training set (MDI[training]) is poorer at the first selected bands, and keeps growing up to obtain a similar performance with respect to other supervised methods.

With respect to computational time, Table I shows the running times of the algorithms. In the case of the supervised feature selection criteria, the better performance are achieved by *ReliefF* and JM distance. In the case of *ReliefF*, this approach obtains a ranking of relevance for each single feature and the computational cost grows exponentially with respect to the number of samples in the data set. Moreover, it is not clear to what extent this technique can detect redundant or highly interacting attributes when increasing the number of features.

The JM distance provides a high classification accuracy and low computational cost, as well as Wilk's λ for a low number of classes, this being the particular case. In the case of MIFS, it does not seem the use of the mutual information of the image bands and the use of class labels information lead to an improvement of the selected features, with respect to the performance of the NN rule. The MDI criterion is not the most efficient method from computational point of view. This is mainly due to the cost of computing the joint probability distributions, as it was analyzed in Section IV. On the other hand, other methods like *ReliefF* and MIFS present a higher computational cost with no notable improvement in their classification performance (see Fig. 6).

Therefore, for the band selection problem, where there exists a high correlation among different features (image bands), the principle of looking for noncorrelated and, at the same time, as much information as possible has proven to be effective to obtain subsets of selected image bands that also provide satisfactory results from the classification accuracy point of view. Thus, the use of criteria based on information theory over unlabeled data as a feature selector is a valid approach to extract discriminant subsets of image bands from a multispectral image representation, with an affordable computational complexity, without providing labeled information.

Another interesting result in the image data set used is that, looking at Fig. 6, from a subset of six/seven image bands, the improvement of the classification accuracy is not significant. For instance, the MDI[9 oranges] criterion with only a subset of 11 image bands reaches a classification performance of 88.1%, while using the whole 33 image bands of the VIS range, the classification performance is 85.2%. This is due to the fact that the dependent amount of information present in the subset of the 11 chosen bands is already very close to the maximum joint entropy content that can be reached with any set of bands; thus, any new image band contributes very little to the conditional entropy terms in the MDI criterion.

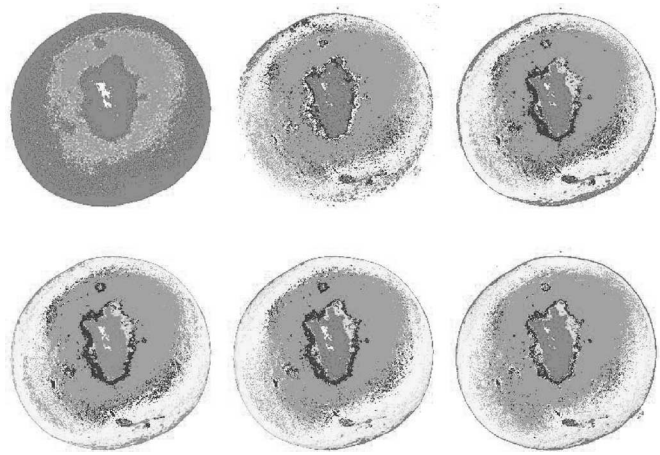


Fig. 7. Examples of labeling results for different subsets of bands corresponding to image in Fig. 8. (Top, left to right) Results with 1, 4, and 8 image bands selected. (Bottom, left to right) Results with 12, 20, and 33 image bands selected.

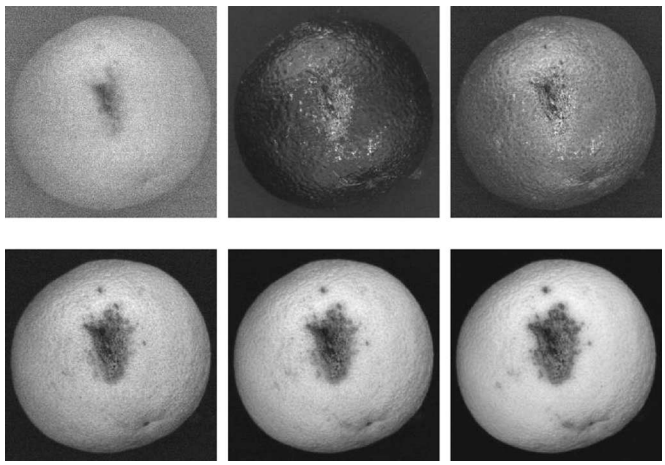


Fig. 8. Examples of image bands from a multispectral image in the visible spectrum. (Top, left to right) Violet, blue, and green. (Bottom, left to right). Yellow, orange, and red.

To graphically see this effect, in Fig. 7, the classification of the pixels in a new (not used in the training set) hyperspectral orange image is shown, assigning a class label for each pixel using the NN rule over the training set. Fig. 7 shows the behavior of the pixel labeling for different subsets of bands chosen from the ranking obtained by MDI. Note how the appearance of the resulting labeling does not have appreciable changes from the eight features, approximately.

Finally, Table II shows the ranking of image bands selected by the different techniques used in the experiments. In the case of the multispectral image data base used, the spectral bands in the violet region show a poorer signal-to-noise ratio; therefore, their inclusion in the feature set usually leads to a worse performance in the classification rate. The spectral bands around the yellow region are bands with larger value of entropy. In the orange–red

TABLE II
RANKING OF IMAGE BANDS SELECTED BY THE DIFFERENT TESTED METHODS IN NANOMETERS

	ReliefF	MIFS	Wilk's λ	JM distance	MDI [tra.]	MDI [9 oran.]
F1	700 nm	510 nm	650 nm	720 nm	670 nm	670 nm
F2	640 nm	680 nm	640 nm	710 nm	470 nm	480 nm
F3	680 nm	500 nm	660 nm	700 nm	480 nm	470 nm
F4	650 nm	670 nm	630 nm	530 nm	490 nm	490 nm
F5	690 nm	530 nm	670 nm	690 nm	500 nm	500 nm
F6	710 nm	710 nm	620 nm	520 nm	460 nm	710 nm
F7	660 nm	450 nm	610 nm	440 nm	450 nm	720 nm
F8	720 nm	720 nm	680 nm	640 nm	630 nm	700 nm
F9	590 nm	490 nm	600 nm	540 nm	520 nm	690 nm
F10	630 nm	430 nm	690 nm	680 nm	440 nm	680 nm
F11	600 nm	700 nm	590 nm	650 nm	510 nm	650 nm
F12	670 nm	570 nm	700 nm	510 nm	400 nm	660 nm

region of the spectrum, the bands are significantly correlated, and the contributions of information of any of them are similar.

Note how the bands selected by *ReliefF* are mainly in the red region, this approach being unable to detect redundant information. In the case of MDI[9 oranges], the image bands selected are mainly located in the blue and red regions alternatively. These regions are well separated in the spectrum and their image bands have a notable difference in their information content, increasing the amount of discriminant information of the ensembles.

VI. CONCLUDING REMARKS

An approach to select image bands in multispectral images based on information theory concepts has been introduced. From the information theory point of view, different properties can be estimated for multispectral images to know about their relationships in terms of shared and total amount of joint information.

The proposed approach tries to look at the problem estimating the independent information in a set of image bands, rather than looking at any type of correlation or dependent information measure. The extraction of selected subsets of spectral image bands can be obtained by means of a criterion based on the proposed MDI, which consists of a relation between the joint entropy and the union of the conditional entropies of the considered set of image bands.

This criterion looks for sets of spectral bands with minimum interdependency and high amount of cojoint information. This objective is achieved by calculating the subset of bands that have high conditional entropies, as a measure of the independent information each image band provides to the ensemble, being the principle of the MDI criterion.

Although this criterion has not been established in terms of class separability for supervised training sets, it has been shown in the experimental results that the image bands selected by the proposed approach behave as if they were obtained by an unsupervised feature selection algorithm, providing very satisfactory results with respect to classification accuracy when using the selected bands, even outperforming the other supervised methods used in the comparison in most situations. This constitutes an important advantage of this technique. Therefore, we could consider the proposed method uses data in an unsupervised way, avoiding tedious labeling of prototypes, which allows to easily deal with very large data sets, due to the fact that labeling is not necessary.

REFERENCES

[1] J. Aczél and Z. Daróczy, *On Measures of Information and their Characterization*. New York: Academic, 1975.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional for data mining applications," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Seattle, WA, Jun. 1998, pp. 94–105.

[3] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[4] G. H. Bearman, D. Cabid, and R. M. Levenson, "Spectral imaging: Instrumentation, applications and analysis," in *Proc. SPIE*, vol. 3920, 2000.

[5] K. D. Bollacker and J. Ghosh, "Linear feature extractors based on mutual information," in *Proc. 13th Int. Conf. Pattern Recognit.* Vienna, Austria, 1996, vol. B, pp. 720–724.

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

[7] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, 2004.

[8] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.

[9] C. J. Huberty, *Applied Discriminant Analysis*. Hoboken, NJ: Wiley, 1994.

[10] P. Groves and P. Bajcsy, "Methodology for hyperspectral band and classification model selection," presented at the IEEE Workshop Adv. Tech. Anal. Remotely Sensed Data. An Honorary Workshop for Prof. David A. Landgrebe, Washington, DC, 2003.

[11] A. K. Jain and W. G. Waller, "On the optimal number of features in the classification of multivariate gaussian data," *Pattern Recognit.*, vol. 10, pp. 365–374, 1978.

[12] L. Jimenez and D. Landgrebe, "Supervised classification in high dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Trans. Syst., Man, Cybern. C*, vol. 28, no. 1, pp. 39–54, Feb. 1998.

[13] Y. Kim, W. N. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Dis. Data Mining*, Boston, MA, 2000, pp. 365–369.

[14] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Proc. 7th Eur. Conf. Mach. Learn.*, Catania, Italy, 1994, pp. 171–182.

[15] D. Korycinski, M. M. Crawford, and J. W. Barnes, "Adaptive feature selection for hyperspectral data analysis using a binary hierarchical classifier and tabu search," in *Proc. 2003 Int. Geosci. Remote Sens. Symp.*, Toulouse, France, Jul. 2003, pp. 297–299.

[16] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognit.*, vol. 33, no. 1, pp. 25–41, Jan. 2000.

[17] A. Kulcke, C. Gurschler, G. Spöck, R. Leitner, and M. Kraft, "On-line classification of synthetic polymers using near infrared spectral imaging," *J. Near Infrared Spectrosc.*, vol. 11, pp. 71–81, 2003.

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

[19] M. Last, A. Kandel, and O. Maimon, "Information-theoretic algorithm for feature selection," *Pattern Recognit. Lett.*, vol. 22, pp. 799–811, 2001.

[20] D. Landgrebe, "Hyperspectral image data analysis as a high dimensional signal processing problem," *IEEE Signal Proc. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.

[21] M. H. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[22] R. Leitner, H. Mairer, and A. Kercek, "Real-time classification of polymers with NIR spectral imaging and blob analysis," *Real-Time Imag.*, vol. 9, no. 4, pp. 245–251, 2003.

[23] F. Masulli and G. Valentini, "Mutual information methods for evaluating dependence among outputs in learning machines," Tech. Rep. TR-01-02, DISI Dipart. Inf. Sci. Inf., Univ. Genova, Italy, 2007.
[24] H. Matsuda, "Physical nature of higher-order mutual information: Intrinsic correlations and frustration," *Phys. Rev. E*, vol. 62, no. 3, pp. 3096–3102, 2000.
[25] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.
[26] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Incorporating connected region labelling into automated registration using mutual information," in *Mathematical Methods in Biomedical Image Analysis*, A. A. Amini, F. L. Bookstein, and D. C. Wilson, Eds. Los Alamitos, CA: IEEE Comput. Soc. Press, 1996, pp. 23–31.
[27] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, 2003.
[28] G. D. Tourssari, E. D. Frederick, M. K. Markey, and C. E. Floyd, Jr., "Applications of mutual information criterion for feature selection in computer-aided diagnosis," *Med. Phys.*, vol. 28, no. 12, pp. 2394–2402, 2001.
[29] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM J. Res. Develop.*, vol. 4, pp. 66–82, 1960.
[30] B. Yu, I. M. Ostland, P. Gong, and R. Pu, "Penalized discriminant analysis of *in situ* hyperspectral data for conifer species recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 5, pp. 2569–2577, Sep. 1999.

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.

Currently, he is an Assistant Lecturer at the Department of Programming Languages and Computer Systems, Jaume I University, Castellón de la Plano, Spain. He has collaborated in different projects, most of them in medical application of computer science, and has published more than 25 scientific papers in national and international conferences, books, and journals. His main research interests include pattern recognition and biomedical applications, including image pattern recognition, hyperspectral data, structured light, and feature extraction and selection.

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

He has been a Visiting Scientist at the Silsoe Research Institute, the University of Surrey, the University of Bristol, CEMAGREF, the University of Genoa, and the Instituto Superior Tcnico, Lisbon Portugal. He is a Full Professor at the Department of Programming Languages and Information Systems, Jaume I University, Castellón de la Plano, Spain, and is currently the Group Coordinator at the Computer Vision Laboratory. He has authored more than 90 scientific papers in the fields of computer vision and pattern recognition. He has also been a co-editor of two books and acted as a 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.

Dr. Pla is a Member of the International Association for Pattern Recognition.

José Salvador Sánchez (S'95–A'98–M'00) received the B.Sc. degree in computer science from the Technical University of Valencia, Valencia, Spain, in 1990 and the Ph.D. degree in computer science engineering from Jaume I University, Castellón de la Plano, Spain, in 1998.

Since 1992, he has been an Associate Professor at the Department of Programming Languages and Information Systems, Jaume I University, and is currently the head of the Pattern Recognition Section, Computer Vision Laboratory. He is author or coauthor of more than 60 scientific publications and co-editor of two books. His current research interests include pattern recognition and machine learning domains, including classification, feature and prototype selection, ensembles of classifiers, and decision tree induction.

Dr. Sánchez is a Member of IEEE Signal Processing Society, the IEEE Neural Networks Society, the IEEE Information Theory Society, IAPR, AERFAI (Spanish Association of Pattern Recognition and Image Analysis), ECCAI, and AEPIA (Spanish Association for Artificial Intelligence).