Automatic Band Selection in Multispectral Images Using Mutual Information-Based Clustering

Adolfo Martínez-Usó, Filiberto Pla, Pedro García-Sevilla, and J.M. Sotoca

LSI Department, Jaume I University. Castellón, Spain
{auso, pla, pgarcia}@uji.es
http://www.vision.uji.es

Abstract. Feature selection and dimensionality reduction are crucial research fields in pattern recognition. This work presents the application of a novel technique on dimensionality reduction to deal with multispectral images. A distance based on mutual information is used to construct a hierarchical clustering structure with the multispectral bands. Moreover, a criterion function is used to choose automatically the number of final clusters. Experimental results show that the method provides a very suitable subset of multispectral bands for pixel classification purposes.

1 Introduction

Works in multispectral imaging are producing many emerging applications in several disciplines. Multi or hyperspectral sensors acquire information from a range of wavelengths in the spectrum and, unquestionably, they have produced an important improvement of the results obtained from just one or three bands in some demanding application fields, like remote sensing, medical imaging, product quality inspection, fine arts, etc. The work we present here is not focused on a specific field and could be applied to any kind of multispectral images. However, due to the lines of work that we follow, we are strongly interested in SAR images as well as in fruit quality inspection tasks.

Obviously, from the point of view of pixel classification tasks, a very desirable step when we have a large amount of input spectral information is a process to reduce this initial information without losing classification accuracy in a significant way. This reduction could be done in two different ways: feature extraction [9,7] or feature selection [2]. 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. In this work we will focus on feature selection rather than feature extraction due to the fact that in feature extraction the total amount of information is needed to obtain the new set of input bands. On the other hand, selecting the relevant range of wavelengths in the spectrum, where the process obtains better results, allows the acquisition step to deal with a reduced set and makes the analysis simpler.

In multispectral applications, the question is how to select the correct bands from the multispectral range to characterise the problem. In this case, regarding to feature selection for pixel classification, this question could be addressed using information theory and, more concretely, by measures based on the mutual information concept [11].

In recent years, clustering techniques are becoming more popular, being hierarchical clustering one of the most used approaches. Important advances have been made

J.F. Martínez-Trinidad et al. (Eds.): CIARP 2006, LNCS 4225, pp. 644–654, 2006.

in different fields as segmentation [1] [10], text classification [3] or even in semantically meaningful grouping [12]. A comprehensive analysis of these methods can be also found in [4]. In our work, we take advantage of this representation because it is a very intuitive way to group the input data in order to progressively reduce the amount of information.

The methodology of the algorithm presented in this work can be summarised as follows. A similarity space is defined among image bands, where a dissimilarity measure is defined based on the mutual information between a pair of bands. From the initial set of bands that form a multispectral image, the process starts with a hierarchical clustering in the defined dissimilarity space. In order to progressively construct a hierarchical family of derived clusters the method uses a linkage strategy with an inter-cluster distance as the objective function to optimise. The number of final clusters is calculated automatically by means of a functional. The maximum values in this function indicate which number of clusters is suitable in order to form an accurate partition. Finally, for each of the final clusters, a band representing the cluster is chosen, providing the final bands selected, which are considered the most relevant.

2 Band Selection Algorithm

In this section the dimensionality reduction algorithm is introduced. To this end, the method proposed tries to identify the subset of bands that are as much independent as possible among them. It is known that independence between bands [9] is one of the key issues to obtain relevant subsets of bands for classification purposes. As we will show in the experimental results, the resulting bands obtained by means of our method produce very satisfactory classification rates with respect to other feature selection approaches.

To find the subset of K bands that are as much independent as possible among them, our approach defines a dissimilarity space based on mutual information between bands. In this dissimilarity space, a clustering process is performed. As a result of the clustering, bands are grouped according to the amount of information they share. Therefore, all the bands in the same cluster are highly dependent among them. In a final stage, a band representing each cluster is chosen, in such a way that the band selected will be the band that share as much information with respect to the other bands in the cluster. Eventually, the K selected bands from the final K clusters will have a significant degree of independence, and therefore, will provide an adequate reduced representation that will provide satisfactory classification results.

2.1 Mutual Information-Based Distance

Let us calculate mutual information from entropy measures according to the well-known expression $I(X,Y) = H(X) + H(Y) - H(X,Y)$, where $H(X)$, $H(Y)$ are the entropies of random vectors X, Y respectively and $H(X,Y)$ is the joint entropy. I is an absolute measure of common information between two sources, however, as we can infer from the previous equation, I by itself would not be a suitable distance measure. The reason is that it can be low because the X, Y variables present a weak relation (such as it should be desirable) or because the entropies are small (in such case, the variables contribute with few information). Thus, it is convenient to obtain a proper measure so that it works independently from the marginal entropies and also measures the statistical dependence as a distance.

Let us consider a set of n bands $X_1, ..., X_n$ from a multispectral image and let us suppose that each band represents a random variable. From this input data, we shall employ a measure of similarity between any two random images, $NI(X_i, X_j) = \frac{2 \cdot I(X_i, X_j)}{H(X_i)+H(X_j)}$, which is a normalised measure of I. This measure is used to calculate distance $D_{NI} = \left(1 - \sqrt{NI(X_i, X_j)}\right)^2$. Both D_{NI} and NI had been proposed in [5].

2.2 Hierarchical Clustering

The hierarchical structures are commonly represented by a tree diagram or *dendrogram* with a nested set of partitions. In this representation, called hierarchical clustering, the sequence of disjoint partitions is obtained using only the information contained in a distance matrix. This matrix of dissimilarities calculates the distance D_{NI} for each pair of groups and is used to decide how to link nested clusters in consecutive levels of the hierarchy.

There are several linkage strategies that we can use as the rule to decide how the distance matrix has to be updated [6]. Different linkage strategies create different tree structures. The algorithm here proposed uses an agglomerative strategy, that is, it starts with n initial clusters and, at each step, merges the two most similar groups to form a new cluster. Thus, the number of groups is reduced 1 by 1 until there is just one cluster. Our hierarchical clustering algorithm is based on a Ward's linkage method [13]. Ward's linkage method has the property of producing minimum variance partitions. Thus, this method is also called minimum variance method because it pursues to form each possible group in a manner that would minimise the loss associated with each grouping (internal cohesion). To this end, the hierarchical grouping merges the pair of clusters that minimise the increment in the square error of the whole partition. The error used to this calculation is the intra-cluster dispersion. In addition to several studies that conclude that this method outperforms other hierarchical clustering methods [6], the process helps us to form groups with not much variance in their level of independence, that is, clusters with similar D_{NI} distances will be joined together.

2.3 Fully Automated K-Assessment

Most of the applications that imply a band selection process suffer from a lack of an $automatic-K-selection$, that is, the final number K of selected bands is not chosen automatically. This drawback is usually solved by a manual introduction of the K value or by determining a threshold value in order to control the progression of certain functional [4]. Therefore, a method with an $automatic-K-selection$ would be desirable in order to finish correctly the hierarchical process and make the method completely unsupervised.

In this paper, we introduce a functional that automatically calculates how many clusters would be desirable as a final subset, that is, the K number. In our work, we have compared this result with the classification rates for each number of final clusters in order to check the validity of the K taken. Thus, a valid K number would be a value from which the classification accuracy does not improve or even become worse.

The developed functional will begin in the stage where each band, from the N starting bands, is a single cluster and will finish in the stage with only two clusters. Let us suppose certain stage where the bands are grouped in n clusters (where $n \leq N$) forming a partition $C = \{c_0, c_1, ..., c_n\}$, that is, a set of clusters. Let us also suppose that we have a cluster of bands c_i and $\bar{C}_i = \{c_0, c_1, ..., c_n\}$ is the complementary subset of clusters where $c_i \notin \bar{C}_i$, $C \equiv \{c_i\} \cup \bar{C}_i$.

Let us define $I(c_i)$ as the average of the *internal* distances[1] among the bands b_i belonging to c_i. We shall also define $E(c_i)$ as the average of the *external* distances between the bands belonging to c_i and the bands in $\bar{C}_i$. Thus,

$$I(c_i) = \frac{1}{\|c_i\|^2} \sum_{b_i \in c_i} \sum_{b_j \in c_i} D_{NI}(b_i, b_j) \qquad E(c_i) = \frac{1}{\|C_i\| \cdot \|\bar{C}_i\|} \sum_{b_i \in c_i} \sum_{b_j \in \bar{C}_i} D_{NI}(b_i, b_j)$$

Note that $I(c_i)$ calculates an intra-cluster average difference whereas $E(c_i)$ calculates the inter-cluster average difference. Both of them use mutual information among bands (as described in 2.1) and are related to a particular cluster c_i. Hence, we will define $P_I(C)$ and $P_E(C)$ as the global average measures among all the clusters in a particular partition C as follows:

$$P_I(C) = \frac{1}{\|C\|} \sum_{c_i \in C} I(c_i) \qquad P_E(C) = \frac{1}{\|C\|} \sum_{c_i \in C} E(c_i)$$

In an ideal partition, we would hope the inter-cluster value $P_E(C)$ to be very large, and the intra-cluster value $P_I(C)$ to be very small. Thus, left side of figure 1 plots the function $P_E(C) - P_I(C)$ against the number of clusters in C. As we can see, the maximum difference between $P_E(C)$ and $P_I(C)$ is obtained when each band is considered as an independent cluster. Since our aim is a band reduction, this measure is not enough by itself. Hence, we plot the linear function that ranges from 0 to $P_E(C_0) - P_I(C_0)$ where C_0 is the initial partition when each band is considered as a single cluster. This linear behaviour would be the graph we will expect if all steps in the clustering process would provide the same variation in the values of $P_E(C)$ and $P_I(C)$.

Taking into account the two described functions, we shall consider the resulting functional from the difference between the first one, which could be considered as the *real* behaviour, and the second one, which could be considered as the expected *linear* behaviour. The maximum value in this $K\_functional$ is considered the better K value for the final number of clusters. For some application, other local peaks around the maximum of the $K\_functional$ could also be taken into account. Plotting the functional values for each number of clusters (Fig. 1 on the right) we can see that the graph draws an increasing function from right (where each band is an independent cluster) to left (where several bands are grouped in a few clusters) until a maximum value which is the selected K number. In short, the functional equation has the following form:

$$K\_func(C) = (P_E(C) - P_I(C)) - \|C\| \frac{P_E(C_0) - P_I(C_0)}{\|C_0\|}$$

[1] It is important to point out that when we talk about distance we are referring to D_{NI}.

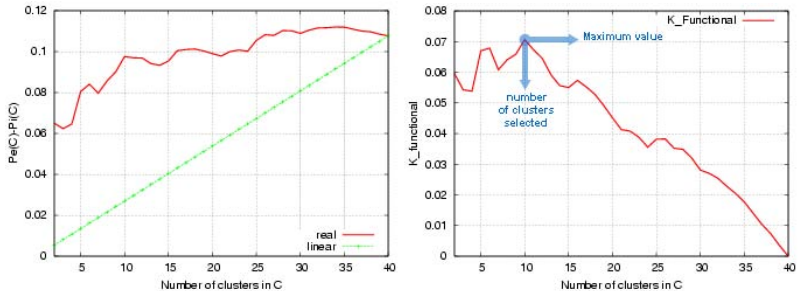

Fig. 1. Left picture shows *real* and expected *linear* graphs for NIR database. Resulting $K\_functional$ is plotted on the right.

2.4 Choosing the Cluster Representatives

After the distance matrix is initialised, the algorithm looks for the two most similar clusters that will have the minimum distance value in the matrix. These two clusters are merged into one and the matrix is updated using Ward's linkage method. Of course, the rows/columns corresponding to the merged clusters are deleted and a row/column for the new cluster is added.

The described process is repeated until the stage with just one band. After that, by means of the previous described functional, the algorithm selects the K number of final clusters. The resulting mutually exclusive groups represent groups of highly correlated bands, and bands from two different clusters will have low correlation. Thus, let us consider now the resulting cluster c_i with n bands. The weight of each band $i \in c_i$ is calculated as $W_i = \frac{1}{n}\sum_{j\in c_i, j\neq i}\frac{1}{\epsilon+D(i,j)^2}$ where ϵ is a very small value to avoid singular values, and function $D(i,j)$ returns the distance value between bands i,j. The representative band from each group is selected as the band with the highest W of the cluster. A low value of W_i means that the band i has an average large distance with respect to the other bands in the cluster, that is, in this case, the band i will have an average low correlation with respect to the other bands in the cluster. In a reverse way, a high value of W_i means that band i has, in average, a high correlation with respect to the other bands in the cluster. Thus, choosing the band in the cluster with the highest average correlation (mutual information) with respect to the other bands in the cluster, what we are doing is choosing the band that better predicts the information content of the other bands, since the more mutual information two random variables have, the more can predict one of the variable about the other one.

As a result of the algorithm, K bands will be selected that represent K clusters. These K bands will be significantly separated in the dissimilarity space defined, thus, having a low correlation and, therefore, having a high degree of independence among them.

3 Results

To test the proposed approach, several multispectral images from different databases are used in the experimental results:

1. Multispectral images of oranges obtained by an imaging spectrograph (RetigaEx, Opto-knowledge Systems Inc., Canada). This database has two groups, VIS collection (400-720 nm in the visible) and NIR collection (650-1050 nm in the near infrared). In both cases, the camera has a spectral resolution of 10 nm. The database includes several kinds of orange defects. It has eight classes, obtaining 1463346 labelled pixels from VIS and 1491888 labelled pixels from NIR.
2. The $92AV3C$ source of data corresponds to a spectral image (145 X 145 pixels, 220 bands, 17 classes) acquired with the AVIRIS data set and collected in June 1992 over the Indian Pine Test site in Northwestern Indiana[2].
3. $DAISEX'99$ project provides useful aerial images about the study of the variability in the reflectance of different natural surfaces. This source of data corresponds to a spectral image (700 X 670 pixels, 6 classes) acquired with the 128-bands HyMap spectrometer during the DAISEX-99 campaign (http:/io.uv.es/projects/daisex/).

In addition to the previous description, images in Fig. 2 show some instances of the database collections used. These images are presented as RGB compositions.

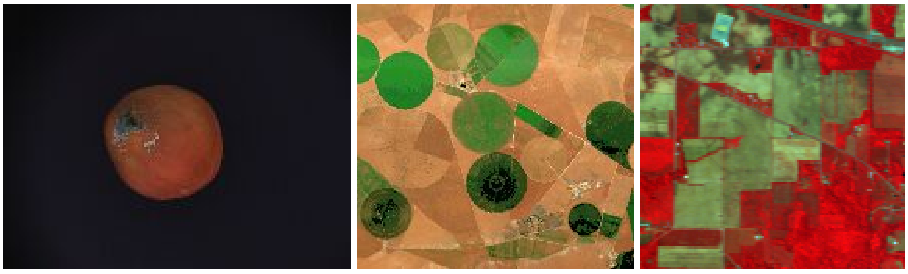

Fig. 2. Examples of RGB composition. First for an orange image in the Visible spectrum, second for HyMap spectrometer and third for AVIRIS (92AV3C)

On $92AV3C$ and $DAISEX'99$ databases, because of the labelled "background", which corresponds to pixels with an undetermined class, we can divide each database into two groups, one with background and another without background.

Since we perform the Ward's linkage method using a distance based on Mutual Infor-mation, we shall name hereafter $WaLuMI$ to our proposed algorithm. It has been tested with these six databases described, that is, the VIS and NIR collections from the database of oranges, $92AV3C$ database with and without background and $DAISEX'99$ database with and without background.

In order to assess the performance of the method, a Nearest Neighbour (NN) classifier was used to classify pixels into the different classes. The 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.

To analyse the accuracy of the ranking of bands obtained by the proposed approach, two supervised filter feature selection methods were also tested. Thus, the band selection process was considered as a supervised feature selection approach, in this case

[2] http:/dynamo.ecn.purdue.edu /∼biehl/MultiSpec

using the labelled data set for the feature selection process. The main motivation about comparing the proposed method with supervised approaches is that the labelled data contains information about the distribution of classes existing in the hyperspectral data, and they allow the search for relevant feature subsets. By comparing the performance with those approaches, we can measure the capability to obtain subsets of relevant features (image bands) by the introduced algorithm without a prior knowledge of the class distributions in the multispectral image, allowing the labelling of data.

The first method is the well-known *ReliefF* algorithm [8] based on pattern distances. This algorithm initialises every feature weight to zero and then iterates m times looking for a set of feature weights that optimises a criterion function.

The second technique is related to divergence measures between classes. One of the best-known distance measures used for feature selection in multi-class problems is the average Jeffries-Matusita (JM) distance [2].

In terms of class separability, the higher is the JM distance between two classes, the more separability between them. To obtain suboptimal subsets of features, we have applied a search strategy based on a Sequential Forward Selection applying this distance $((SFS)JMdistance)$. This technique starts from an empty feature subset and adding one feature at a time, reaching a feature subset with the desired cardinality.

3.1 Performance Evaluation Including Background Pixels

During the image labelling process, there are always pixels in an image that are not assigned to any class of interest, mainly because they are pixels that either do not clearly belong to some of the predefined classes or they are assigned to a complementary class. The pixels that have not been assigned to any class are labelled as "background" class. In this subsection, we include the background information in the databases for its evaluation.

In order to increase the statistical significance of the results, the experimental results shown in this section about the classification rates correspond to the average classification accuracy obtained by the NN classifier over five random partitions. The samples in each partition were randomly assigned to the training and test set with equal sizes as follows: VIS = 43902 pixels, NIR = 44758 pixels, HyMap = 37520 pixels, 92AV3C = 2102 pixels.

On the other hand, given the huge size of the data sets and the trouble in computational cost to apply the supervised approaches, particularly in the case of VIS, NIR and HyMap, the following independent partitions with respect to the data sets were randomly extracted maintaining the prior probability of the classes: VIS = 87805 pixels, NIR = 89516 pixels, HyMap = 93804 pixels and 92AV3C = 10512 pixels. Using these databases, the supervised approaches and the proposed method were applied in order to obtain a ranking of relevance of the features, that is, of bands.

Figures 3, 4 (left) and 5 (left) represent in their top row the classification rates with respect to the subset of N bands selected by each method for each database. In all cases, we show the performance of the *NN* classifier with respect to the number of features obtained by $WaLuMI$, $(SFS)JMdistance$ and $ReliefF$. Note that the proposed method obtained better performance with respect to the rest of methods in all databases. It is worthwhile mentioning that the $WaLuMI$ approach has a good behaviour in all cases when choosing the smaller sets of bands (one to ten), where the decision is more critical.

$ReliefF$ performs poorer with respect to the other approaches except with HyMap image, where the performance of $(SFS)JMdistance$ is worse.

Therefore, regarding to the band selection problem, where there exists high correlation among different features (image bands), the principle of looking for non-correlated bands from the different regions of the spectrum, by reducing the mutual information in the ensemble of image bands, has proven to be an effective approach to obtain subsets of selected image bands that also provide satisfactory results from the classification accuracy point of view.

3.2 Performance Evaluation Without Background Pixels

The hyperspectral data assigned to the "background class" are usually very scattered and overlapped with other classes, and this fact damages the classification accuracy. Moreover, the elimination of this information supposes a supervised knowledge to detect those regions of the image.

These regions are very difficult to detect with precision from unsupervised information. Therefore, the goal of this experiment is analysing the advantages that suppose the knowledge of the class distribution without the noise that the *background* class can introduce. In this case, we will focus on HyMap and 92AV3C hyperspectral data, where the background information is much more undefined.

In the case of HyMap, we added the *background* class to the training set and validation set: training = 26190 pixels and validation = 65479 pixels. The test set contains all classes except the *background* class. The total number of test samples is 327336 pixels. Thus, the experiment classifies the test using the ranking of relevance of the features obtained by the validation set with the proposed method and the supervised methods used in the comparison.

The image 92AV3C only contains 10366 instances without the *background* class. Therefore, we apply a holdout partition, where the training and the validation set have the same size with 5181 pixels and the rest of pixels represent the test set = 5185 pixels.

Figures 4 and 5 represent, in their right side of the top row, similar classification results than the previous subsection, but without "background class". The best performance is obtained by $WaLuMI$, even better in the first bands where the decision is more critical.

3.3 Selection of the K Final Clusters

For the experiments shown in previous subsections, we also present the values of the $K\_functional$ described in section 2.3. Figures 3, 4 and 5 show the results[3] of this $automatic - K - selection$ process in their bottom rows. We can realise how the maxima of the functional approximately fit with K values which provide satisfactory classification rates. Of course, we could make this comparison each time we carry out a

[3] Note that, in order to achieve a clearly graph, figures only show the 30 last partitions in VIS and NIR databases and the 50 last partitions in $HyMap$ and $92AV3C$ databases.

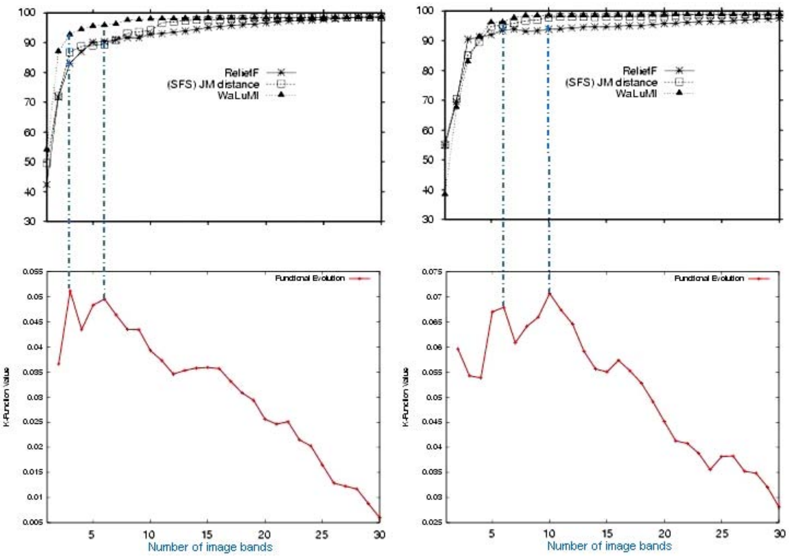

Fig. 3. $Automatic - K - selection$ (x-axe) in VIS/NIR databases. Top row shows the classification rates. Bottom row shows the corresponding K_functional. Left column shows the results for VIS multispectral image. Right column shows the results for NIR multispectral image.

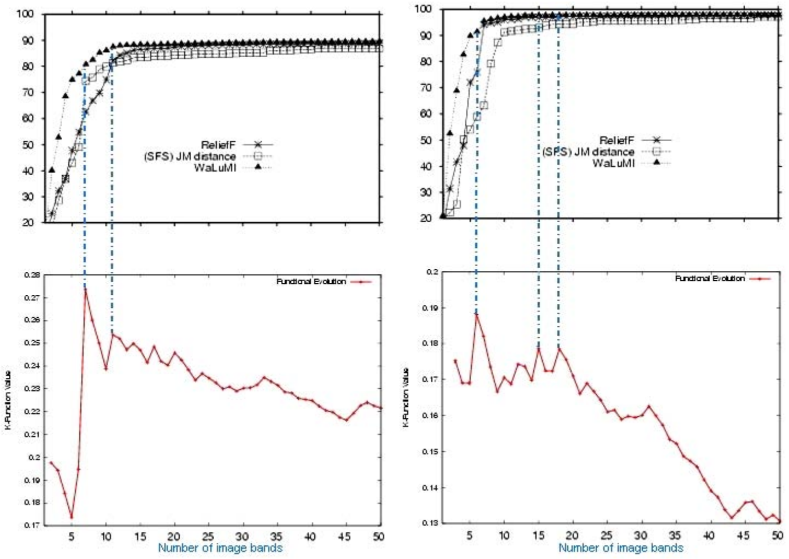

Fig. 4. $Automatic - K - selection$ (x-axe) in HyMap database. Top row shows the classification rates. Bottom row shows the corresponding K_functional. Left column shows the results for HyMap multispectral image. Right column shows the same results, but this time the image is considered without "background class".

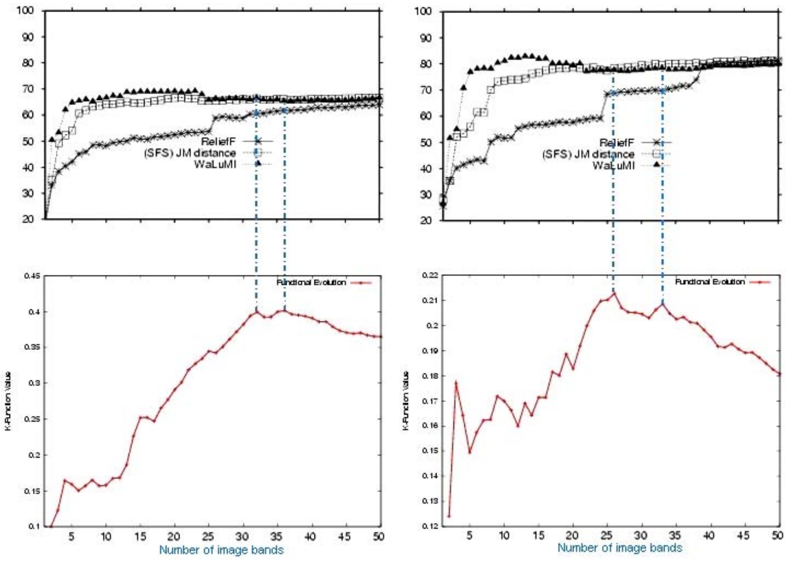

Fig. 5. $Automatic - K - selection$ (x-axe) in 92AV3C database. Top row shows the classification rates. Bottom row shows the corresponding K_functional. Left column shows the results for 92AV3C multispectral image. Right column shows the same results, but this time the image is considered without "background class".

band reduction, that is, testing the classification rates for each possible number of final clusters. The problem is the high computational/temporal cost it involves. So, in order to avoid this expensive process, we provide an automated method that selects as good partitions as we would choose manually according to classification rates but without having to run the classification experiments.

Figures 3 and 4 show in their bottom row the $K\_functional$ values for the oranges and HyMap databases respectively. In both databases a good number of clusters has been selected according to the classification accuracy graph. On the other hand, figure 5 does not achieve as good results as would be desirable because the selected K is a little bit away from the values that we manually would have chosen. However, taking into account that this image starts from 220 clusters (image bands), the K selected is a reasonable band reduction.

4 Conclusions

An unsupervised approach to select image bands in multispectral images based on mutual information measures has been introduced. The method uses a clustering process to group bands correlated among them, and selecting a subset of bands with a high degree of independence in a completely unsupervised way.

The results obtained from the point of view of pixel classification in multispectral images provide experimental evidence about the importance that independence among bands plays in the problem of classification. The method here presented is computationally affordable, avoiding the problem of labelling, and providing very satisfactory classification results with respect to other well known supervised feature selection criteria. In addition, an $automatic - K - selection$ process contributes to achieve a fully unsupervised algorithm that improves our previous work.

References

1. Jean-Marie Beaulieu and Morris Goldberg. Hierarchy in picture segmentation: A stepwise optimization approach. *IEEE Transactions on PAMI*, 11:150–163, 1989.
2. L. Bruzzonne, F. Roli, and Serpico S.B. An extension to multiclass cases of the jeffreys-matusita distance. *IEEE Trans on GRS*, 33:1318–1321, 1995.
3. I. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *JMLR*, 3:1265–1287, 2003.
4. Chris Ding and Xiaofeng He. Cluster merging and splitting in hierarchical clustering algorithms. *ICDM'02*, 1:139–146, 2002.
5. Raquel Dosil, Xosé R. Fdez-Vidal, and Xosé M. Pardo. Dissimilarity measures for visual pattern partitioning. *LNCS*, (3523):287–294, 2005.
6. A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data.* Prentice-Hall, 1988.
7. L. Jimenez and D. Landgrebe. Supervised classification in high dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on System, Man, and Cybernetics*, 28, Part C(1):39–54, 1998.
8. I. Kononenko. Estimating attributes: analysis and extensions of relief. *In Proceedings of 7th European Conference on Machine Learning, Catania, Italy*, pages 171–182, 1994.
9. S. Kumar, J. Ghosh, and M.M. Crawford. Best basis feature extraction algorithms for classification of hyperspectral data. *IEEE Trans on GRS*, 39(7):1368–1379, 2001.
10. E. Sharon, A. Brandt, and R. Basri. Fast multiscale image segmentation. *CVPR*, 1:70–77, 2000.
11. G.D. Tourssari, E.D. Frederick, M.K. Markey, and C.E.Jr. Floyd. Applications of mutual information criterion for feature selection in computer-aided diagnosis. *Machine Learning Research*, 3:2394–2402, 2001.
12. A. Vailaya, M. Figueiredo, A.K. Jain, and H.J. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10:117–130, 2001.
13. John H. Ward. Hierarchical grouping to optimize an objective function. *American Statistical Association*, 58(301):236–244, 1963.