Hyperspectral data processing algorithm combining Principal Component Analysis and K Nearest Neighbours

P. Beatriz Garcia-Allende, Olga M. Conde, Marta Amado, Antonio Quintela, Jose M. Lopez-Higuera
Photonics Engineering Group, Universidad de Cantabria, Avda. Los Castros s/n, 39005 Santander, Spain

ABSTRACT

A processing algorithm to classify hyperspectral images from an imaging spectroscopic sensor is investigated in this paper. In this research two approaches are followed. First, the feasibility of an analysis scheme consisting of spectral feature extraction and classification is demonstrated. Principal component analysis (PCA) is used to perform data dimensionality reduction while the spectral interpretation algorithm for classification is the K nearest neighbour (KNN). The performance of the KNN method, in terms of accuracy and classification time, is determined as a function of the compression rate achieved in the PCA pre-processing stage. Potential applications of these hyperspectral sensors for foreign object detection in industrial scenarios are enormous, for example in raw material quality control. KNN classifier provides an enormous improvement in this particular case, since as no training is required, new products can be added in any time. To reduce the high computational load of the KNN classifier, a generalization of the binary tree employed in sorting and searching, *kd-tree*, has been implemented in a second approach. Finally, the performance of both strategies, with or without the inclusion of the kd-tree, has been successfully tested and their properties compared in the raw material quality control of the tobacco industry.

Keywords: Nearest Neighbours (KNN), Principal Component Analysis (PCA), kd-tree, Imaging spectroscopy, Hyperspectral spectrograph

1. INTRODUCTION

Aerospace remote sensing has been the most important application of hyperspectral imaging spectroscopy. However, its suitability for monitoring industrial applications has been also shown [1-4]. Potential application fields are subjected not only to the development of simple, small and low-cost spectrometers, such as those ones based on passive Prism-Grating-Prism (PGP) devices [1], but also to the investigation of novel spectral interpretation techniques that satisfy real-time operation constraints.

The classification capability of the simple K nearest neighbour (KNN) algorithm [5] has been demonstrated in a wide variety of scenarios, ranging from face recognition [6] to food industry [7]. The latter also shows the successful application of KNN to spectral data. The most significant advantage of the KNN classifier is that no training is required. When a new and unknown spectrum needs to be classified a comparison with the set of previously known spectra is performed. In this paper, the KNN algorithm is investigated as the spectral interpretation algorithm to classify images from a hyperspectral imaging sensor applied to material identification. The KNN classifier has, however, a high computational complexity [8]. To solve this problem the renown compression technique Principal component analysis (PCA) [7,9] is employed. The performance of the classification strategy consisting in the application of PCA and the KNN algorithm is measured, in terms of accuracy and classification time, as a function of the compression rate achieved by PCA.

As mentioned, the main disadvantage of the KNN algorithm when real time constraint is required is its computational overhead. Therefore, once the skills of the proposed image classification strategy are shown, another approach is proposed. It consists in the employment of a generalization of the *k*-dimensional binary tree employed in sorting and searching, *kd-tree* [10-11], which is an efficient method to perform spectrum comparison only over the more similar ones.

Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIV, edited by Sylvia S. Shen, Paul E. Lewis, Proc. of SPIE Vol. 6966, 69660H, (2008) · 0277-786X/08/$18 · doi: 10.1117/12.770298

2. EXPERIMENTAL ISSUES

An schematic description of the imaging spectroscopic sensor setup is depicted in Figure 1. The illumination system is composed by two halogen floodlights Tasley MX500, with a power rating of 500W each. To configure the hyperspectral system, the Imspector V10E PGP device has been connected to a Navitar objective lens Zoom 7000. As this ImSpector version is designed for standard 2/3” detectors (6.6 x 8.8 mm), the monochrome digital camera Pixelink PL-A741 (1280 x 1024 pixels) has been assembled to the other side of the PGP device. Under these conditions, the hyperspectral analysis works in the Vis-NIR range that approximately goes from 400 up to 1000 nm. Finally, the system is controlled by a desktop PC (Pentium IV with a 3 GHz processor and RAM-512 MB) which also performs all the processing tasks with Matlab® 7.0 (R14) [12].

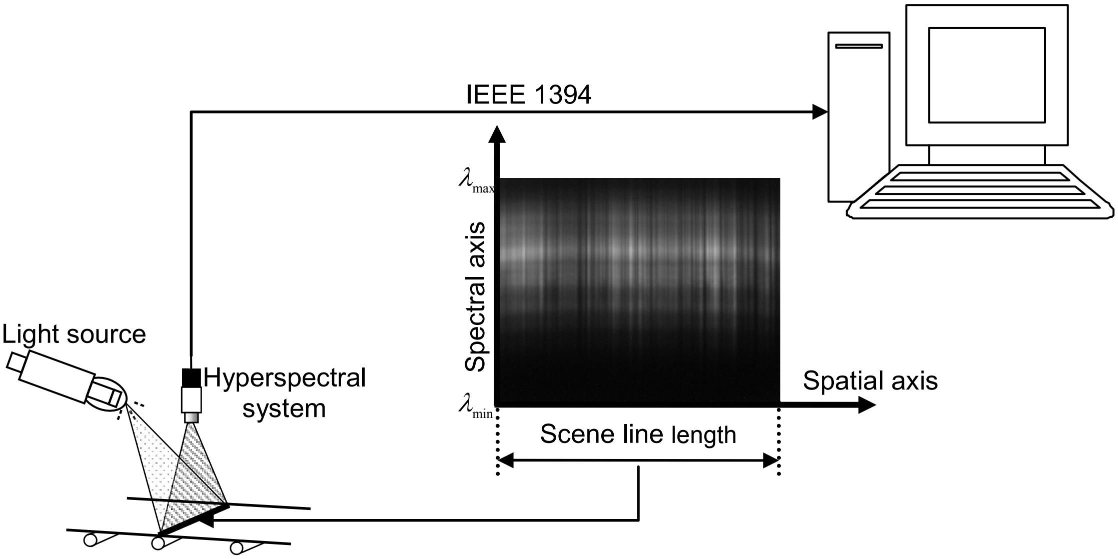

Fig. 1. Schematic diagram of the hyperspectral system.

Figure 1 also shows how the acquired hyperspectral image looks like. The vertical axis, along the 1024 pixels of the camera, covers the spectral range dispersed by the PGP device while the horizontal axis, along the 1280 pixels, images the line of vision of the lens. This line of vision is typically located over the conveyor belt which transports the mixed raw materials in an industrial production plant. Bright spots on the image stand for more diffused reflected radiation from the material under analysis.

Before operation the hyperspectral system must be properly calibrated, both in the spectral and in the spatial directions. The spectral or wavelength calibration is done using two different light sources with known emission wavelengths: one monochromatic He-Ne laser of 670 nm and a Hg-Ar lamp with multiple emission lines. Different algorithms have been used [13]: least squares regression, CDA (Centroid Detection Algorithm), GDA (Gaussian Detection Algorithm) and LPO (Linear Phase Operator). The minimum calibration error is attained with the LPO algorithm, after its application the main spectral specifications of the hyperspectral system become:

- Spectral range:	348,340 – 1021,877 nm.
- Spectral resolution-camera:	0,656 nm.
- Spectral resolution-V10E:	2,8 nm.
- Spectral resolution-system:	2,8 nm.

3. DATA ANALYSIS THEORETICAL FUNDAMENTALS

Qualitative analysis of the hyperspectral images consists in determining for each spatial pixel of the line of vision whether the sample falls within a defined range of allowed variability to determine if the material achieves the desired quality. This section outlines the theoretical fundamentals of the methods/algorithms that composed the proposed technique.

3.1 *K* nearest neighbour classifier

When performing hyperspectral image classification with application to material identification, each spatial pixel of image training set belongs to a known class or material. The aim is to determine the class of each spatial pixel of the scene line of a new and unknown image. A simple strategy could be considering that similar data should belong to the same class. The nearest neighbour classifier is based on this easy idea [5] and classification is performed by comparing the spectrum of the diffuse reflected radiation of the unknown pixel with all the previously known spectra that compose the training images. Similarity measurement can be performed by means of the euclidean distance, which is given by:

$$D(s_1, s_2) = \sum_{i=1}^{N} \left(s_{1i} - s_{2i}\right)^2 \tag{1}$$

where s_1 is the new unknown spectrum, s_2 is each one of the training spectra and *N* is the number of spectral bands where the spectrum of the diffused reflected radiation is measured, therefore in this case $N = 1024$. To avoid the influence of *outliers*, i.e. training spectra assigned to a wrong class, a variation of the method that consists in considering *K* neighbours instead of just one is employed. The unknown spectrum will be assigned to the most numerous class among the *K* nearest spectra. In this way, classification error probability is minimized. Therefore, this classifier has only one parameter: *K* or the number of neighbours to consider. In the proposed method different values of the parameter *K* are evaluated.

3.2 Principal Component Analysis

As indicated in the Introduction, PCA (Principal Component Analysis), also known as the *Hotelling transform* or the *Karhunen Loeve expansion*, is employed as a pre-processing stage. It is a widely known method that it is still being increasingly applied on a high-dimensional feature space (hyperspectral image of Section 2, 1024 spectral dimensions) to achieve dimensionality reduction. The main consequence of this reduction is the improvement of success and time performance of the subsequent classification. PCA provides this reduction together with redundancy removal, since the original data set is expressed on a new basis of vectors whose directions contain the most relevant information. Because of the assumption of PCA that variance means information, the first step of the analysis is subtracting the mean of the data. After that, the covariance matrix is computed and their eigenvalues $\left[f_1 \quad f_2 \quad \ldots \quad f_N\right]$ and eigenvectors $\left[\vec{e}_1 \quad \vec{e}_2 \quad \ldots \quad \vec{e}_N\right]$ are calculated, where *N* is the dimensionality of the input data space. These eigenvectors are also called variables or components, and they form the new basis of projection vectors. To achieve dimensionality reduction, the number of final components must be reduced to a lower value *M*, where *M<N.* The distinct versions of PCA differ in the criterion they employed to select the projection vectors [14]. The m-method, which is the classical and traditional version of PCA, is here used. The eigenvectors/components/features, $\vec{e}_i$, are selected as a function of the value of their corresponding eigenvalues, f_i. The first selected component is that with highest eigenvalue, the second that with the following eigenvalue, and so on. *M* is the threshold between ignored and selected components. The percentage of variance information that is kept in the transformation is denoted by:

$$I = \frac{\sum_{i=1}^{M} f_i}{\sum_{i=1}^{N} f_i} \cdot 100\% \tag{2}$$

where f_i stands for the *i*'th eigenvalue of the transformation; *N* is the dimensionality of the input space or data; *M* is the number of preserved eigenvectors, i.e. the dimensionality of the projected space.

A more detailed description of PCA application to hyperspectral images can be found in a previous work [15]. However, it is worth noting here that, since only the spectral axis is compressed, the total number of images remains the same in the way in which PCA is performed. In the spatial axis of the image, only an average of each 5 spatial pixels is carried out.

3.3 *K*-dimensional binary tree

The KNN classifier is extremely simple but it has several restrictions. One is the employment of the euclidean distance as similarity measurement. It can sometimes provide false results and other distances, such as the probabilistic ones, are required. In the application evaluated here, the worst limitation is, however, its high computational load. A first attempt to reduce this classification time has been done by means of PCA (whose figures will be discussed in Section 4). But this can be insufficient when dealing with strict real-time operation constraints in raw material quality control in some industrial processes. The *k*-dimensional binary tree [10-11], which is a generalization of the simple binary tree employed in sorting and searching, provides an efficient mechanism to evaluate only the closest data to the query one. It is investigated as a possible solution to be able to operate under real-time conditions.

A kd-tree is a data structure to store a finite subset of points in a *k*-dimensional space [10]. Each node of the tree has an associated discriminant coordinate that divides into two subsets the data set associated to the node. The root represents the entire data set, while each non-terminal node has two *sons* or *successors nodes* that represent the two data subsets defined by the segmentation. Terminal nodes contain *buckets* which are mutually exclusive small subsets of data records. In addition, two boundary vectors, which indicate the variation range of the discriminant coordinate, are associated to each node.

The search algorithm is a recursive procedure, whose parameter is the node under investigation. The first invocation passes the root of the tree as this argument. The boundary vector of each node defines the region of the multidimensional space that contains its data subset. The volume of this region is smaller for subsets defined by deeper nodes. If the node under analysis is terminal, all data records in the bucket are evaluated. During the whole search, a list of the *m* closest records and their distance to the query record are constantly maintained.

To be really efficient the expected number of records examined with the search algorithm has to be minimized. This is the mission of the implemented kd-tree optimized version [10], where the parameter to be adjusted is the number of records contained in each terminal bucket.

4. RESULTS AND DISCUSSION

The features of the proposed strategies are presented and discussed in this section. First, classification time and accuracy attained without the inclusion of kd-tree as a function of the compression rate achieved by means of PCA are shown in Section 4.1. Section 4.2 analyzes the improvement in technique performance provided by the kd-tree.

The figures presented here have been obtained employing a hyperspectral data set consisting of the diffuse reflectance spectrum at each spatial point of 120 different samples (30720 spectra). These samples belong to two different classes related with raw material quality control. The first class is composed by hyperspectral images collected by the system of Section 2 for samples consisting of tobacco leaves. And the second class is composed by hyperspectral images associated to undesired materials that typically appear in tobacco industry production plants due to the manual harvest procedure of the leaves. This class includes wood, cardboard, different colored cellophanes, leather, foil, paper of sweets, textile threads and brown and green leaves of other vegetable material different from tobacco. From now on the two classes are designed as "*target*" and "*non-target*", respectively.

In addition, the *cross-validation* method [16] is used to quantify how the technique depends on the reference and test data sets. The data set described above is divided in three subsets of the same size, each one containing 15 images of the non-target class and 25 images of the target one. Provided processing time and classification errors are the mean average of simulations over the three data subsets.

4.1 PCA and KNN with application to hyperspectral images classification

The image spectral axis contains 1024 pixel intensity values between 348 and 1021 nm. This means that the problem has potentially 1024 dimensions which is a difficulty for the KNN classifier because redundant information affects its classification ability. The above described PCA is applied to the images to deal with this problem. Table 1 displays the number of preserved components (*M*) and their associated percentages (I_m) of maintained information as a function of the PCA threshold. This threshold between the eigenvectors ignored and selected by PCA is defined with respect to the maximum eigenvalue of the covariance matrix of the training data set, i.e. a threshold of *1e-2* means that those components whose corresponding eigenvalues are smaller than a hundredth of the maximum are rejected. Classification times attained with each threshold are depicted in Figure 2. The maximum number of features included in the investigation of classification time and error is 19, because for higher percentages of preserved information the time performance of the classifier decreases exponentially.

Table. 1. Percentage of maintained information as a function of the PCA threshold.

PCA threshold	M	I_m(%)
1e-2	5	98.76
1e-3	8	99.46
1e-4	16	99.76
5e-5	19	99.78

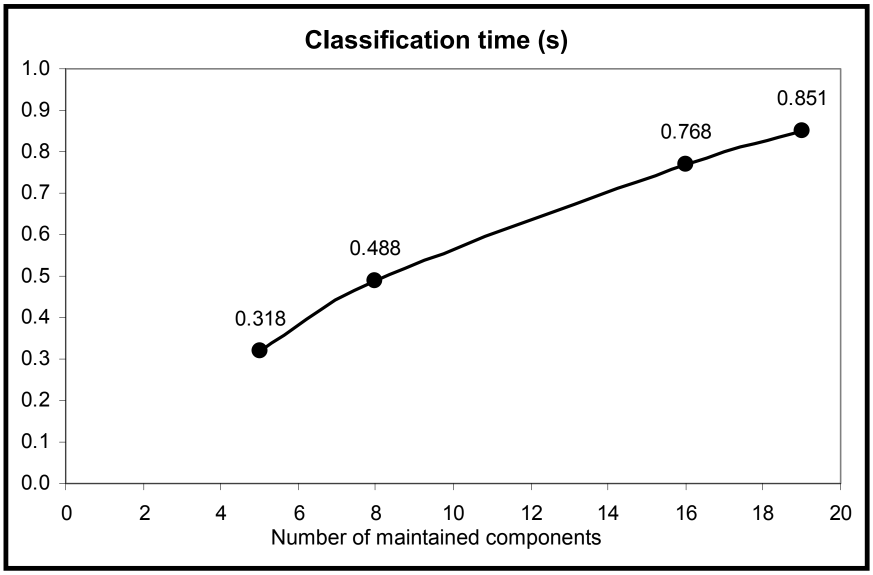

Fig. 2. Classification time performance of the KNN classifier as a function of the compression rate achieved in the pre-processing stage.

Classification time does not depend on the number of considered neighbours(*K*) because the assigned class is selected by voting. Its accuracy, however, does it. Figure 3 depicts simultaneously the dependencies of the percentages error on the number of preserved components and on the number of neighbours. Regarding the latter, 6 different values of the parameter $K\ (K = 1,3,5,10,15,20)$ are considered.

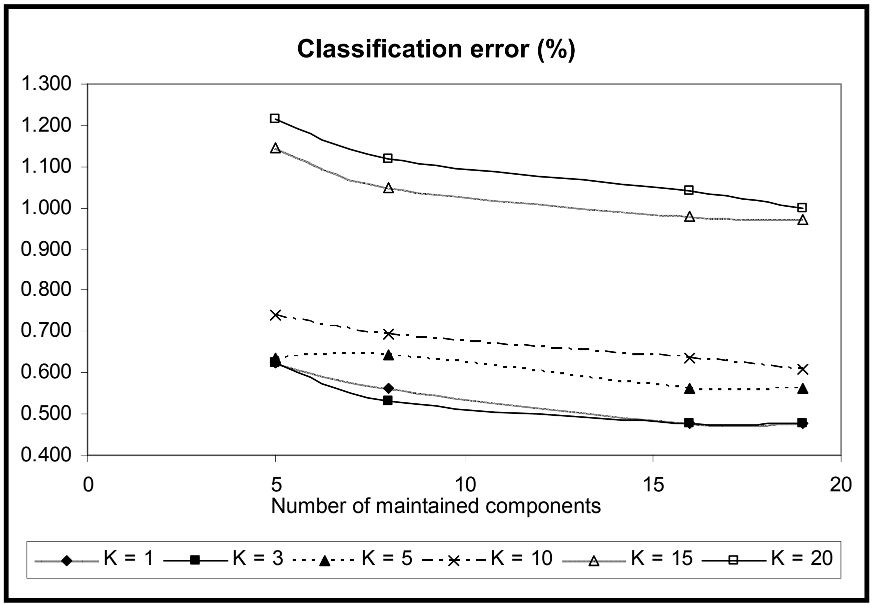

Fig. 3. Classification error performance of the KNN classifier as a function of the compression rate achieved in the pre-processing stage and the number of nearest neighbours (*K*).

As shown in Figure 3, the percentage of error decreases as the number of maintained components grows up to a PCA threshold of 1e-4. Therefore, 16 features are maintained by the PCA analysis. Besides, best classification accuracies are obtained with a number of neighbours ranging from 1 to 10, while percentages of error achieved with 15 or 20 are much higher. Maximum classification accuracy is attained with 3 nearest neighbours, and hence, this value, together with the 16 preserved components, are the chosen values for the adjustable parameters of the proposed data analysis technique. Figure 4 depicts the clustering degree of the target and non-target classes, when they are projected as a function of the first three features selected by PCA.

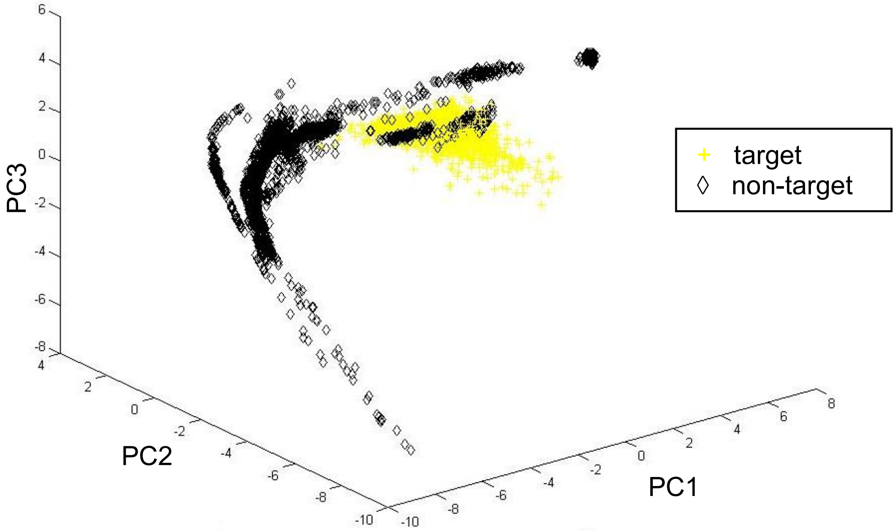

Fig. 4. Clustering of data based on the first three principal components of the reference data.

A mean processing time, including compression and classification, of 768 ms is provided by the proposed technique with the selected parameter values. With application to raw material quality control, in industrial environments this figure limits the maximum allowed speed of the conveyor belt of the production plant.

4.2 Technique performance improvement by means of the kd-tree

KNN classifier potentialities, such as its simplicity, have already been demonstrated in Section 4.1. However, its processing time can reduce heavily its performance in industrial environments. The feasibility of the kd-tree data structure will improve this restriction that usually prevents the employment of the PCA+KNN technique to classify hyperspectral images. The ability of the kd-tree to reduce KNN's classification times depends on the number of training records per terminal bucket. As described in Section 3.3, when a classification is to be made all the spectra in each terminal bucket have to be evaluated. Therefore, six different kd-tree containing respectively 2000, 1000, 500, 100, 50 and 20 spectra in their terminal nodes have been implemented. The previously determined data compression rate has been maintained and also the number of considered neighbours in the KNN classifier. The execution times achieved in this way, including PCA pre-processing and classification, are depicted in Figure 5, while Figure 6 shows that attained classification errors.

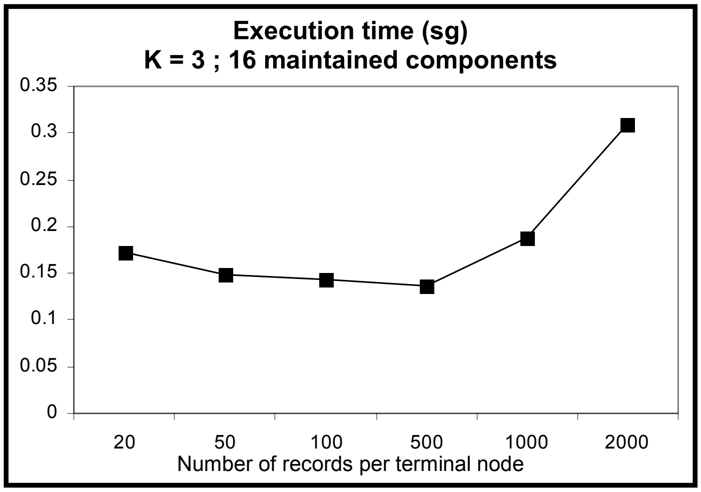

Fig. 5. Classification times with the inclusion of the kd-tree structure as a function of the number of records of each tearminal node of the tree.

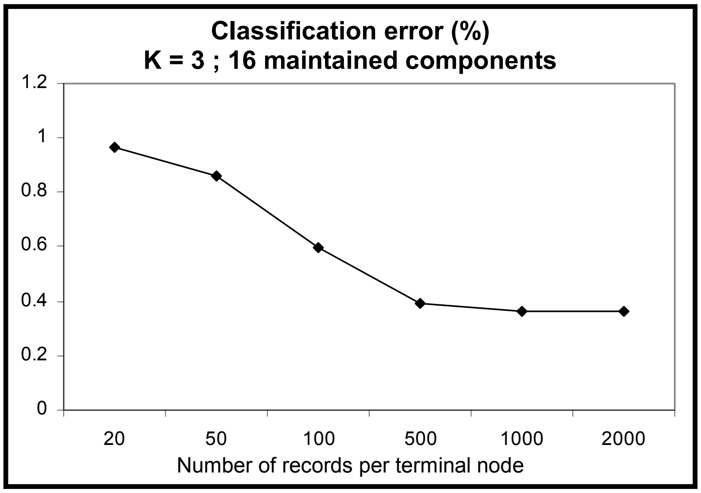

Fig. 6. Classification error with the inclusion of the kd-tree structure as a function of the number of records of each tearminal node of the tree.

Figures 5 and 6 show that no advantage is taken increasing from 1000 to 2000 the number of spectra per terminal node because classification accuracy does not improve and, however, execution time grows considerably. Therefore, the kd-tree implemented for this application will have this number of records in each bucket. While accuracy is maintained with respect to that achieved without the inclusion of the kd-tree data structure, a classification time reduction of 75.65% is obtained. In this way, the proposed interpretation technique will also be feasible when dealing with strict real-time constraints.

5. CONCLUSIONS

The feasibility of the KNN algorithm to classify hyperspectral images from a spectroscopic sensor with application to material identification has been demonstrated. The incentive for employing this technique lies on a potential that it provides in this particular application case. It consists in that, when it is compared with Artificial Neural Networks, for example, it does not require any training. This means that new products to be distinguished can be added at any time because when a classification is to be made the entire training set is examined. However, this advantage is achieved at the expense of a high computational load due to the same reason. To simultaneously take advantage of the benefit of KNN but satisfying real-time operation constraints, two possible solutions have been proposed. First, the widely known PCA has been employed to achieve execution times per image of approximately 768 ms and secondly the implementation of the kd-tree data structure has been included in the processing scheme, allowing a reduction in the execution time of 75.65%. The introduction of the kd-tree structure will depend on time constraints of the application.

ACKNOWLEDGEMENTS

This work has been co-supported by the Spanish TEC'2005-08218-C02-02 and TEC'2007-67987-C02-01 projects.

REFERENCES

[1] E. Herrala, T. Hyvarinen, O. Voutilainen and J. Lammasniemi, "An optoelectronic sensor system for industrial multipoint and imaging spectrometry", *Sensors and Actuators A*, 61, 335-338, (1997).

[2] S. Zavattini, S. Vecchi, R.M. Leahy, D.J. Smith, S.R. Cherry, "A hyperspectral fluorescence imaging system for biological applications", *IEEE Nuclear Science Symposium*, 2, 942-946, (Oct.19-25, 2003).

[3] B. Park, K.C. Lawrence, W.R. Windham, D.P. Smith, P.W. Feldner, "Hyperspectral imaging for food processing automation", *Proceedings of the SPIE*, 4816, 308-316, (2002).

[4] J. Xing, C. Bravo, Pál T. Jancsók, H. Ramon, J. Baerdemaeker, "Detecting Bruises on 'Golden Delicious' Apples using Hyperspectral Imaging with Multiple Wavebands", *Biosystems Engineering*, 90, 27-36, (2005).

[5] D. Barber, "Learning from Data. Nearest Neighbour Classification", http://www.anc.ed.ac.uk/~amos/lfd.

[6] C. Conde, A. Ruiz, E. Cabello, 'PCA vs. low resolution images in face verification', *Proceedings of the 12th Int. Conf. on Image Analysis and Processing* , 63–67, (Sept., 2003).

[7] M. O'farrell, E. Lewis, C. Flanagan, W. Lyons, N. Jackman, "Comparison of k-NN and neural network methods in the classification of spectral data from an optical fibre-based sensor system used for quality control in the food industry", *Sensors and Acturators B*, 111-112, 354–362, (2005).

[8] S. J. Baek, K.M. Sung, "Fast K-nearest-neighbour search algorithm for nonparametric classification", Electronic Letters, 36, 1821–1822, (2000).

[9] J. Workman Jr., A.W. Springsteen, "Applied Spectroscopy. A Compact Reference for Practioners", Academic Press Limited, London, 1st ed., 1998.

[10] J.H. Friedman, J.L. Bentley, R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time", ACM transactions on mathematical software, 3, No. 3, 209-226, (1997).

[11] A.W. Moore, "An introductory tutorial on kd-trees", Extract from Andrew Moore's Thesis: Efficient Memory-based Learning for Robot Control, University of Cambridge, 1991.

[12] Matlab® Reference Manual, The Mathworks Inc., MA, USA.

[13] J. Mirapeix, A. Cobo, C. Jaúregui, J.M. López-Higuera, "Fast algorithm for spectral processing with application to on-line welding quality assurance", *Measurement Science and Technology*, 17(10), 2623-2629, (2006).

[14] T.B. Moeslund, "Principal Component Analysis. An Introduction", Technical Report CVMT 01-02, Aalborg University, ISSN 0906-62333, 2001.

[15] P.B. García-Allende, O.M. Conde, A.M. Cubillas, C. Jáuregui, J.M. López-Higuera, “New raw material discrimination system base don a spatial optical spectroscopy technique”, Sens. Actuator A-Phys. 135 (2007) 605–612.

[16] “Cross Validation” http://research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf#search=%22lecture%2013%20validation%22