Comparison of Supervised Feature Extraction Methods for Multispectral Images

J.M. Leiva-Murillo, R. Santiago-Mozos, F. Pérez-Cruz and A. Artés-Rodríguez
Department of Signal Processing and Communications
Universidad Carlos III de Madrid*
Avda. Universidad 30, 28911 Leganés (Madrid) SPAIN
leiva@ieee.org rsmozos@ieee.org fernandop@ieee.org antonio@ieee.org

ABSTRACT

In this paper, a new feature extraction method is proposed. This procedure focuses on the particular case of the detection of small objects in images. Despite of its simplicity, it outperforms some of the most popular feature extraction methods currently used. This is achieved by taking advantage of some particular characteristics of this particular kind of applications.

1 INTRODUCTION

The problem that motivates this work is the detection of small combustion sources in infrared multispectral images under hard run-time constraints. The processing is based on a sliding window that explores the image and a binary (combustion source/background) non-linear classifier. Sitting on a given point, this window must be wide enough to capture both the spatial and chromatic contrast needed to discriminate between combustion sources and the background. This detector has to be implemented using dedicated hardware in order to meet hard real-time constraints. This characteristic imposes some limitations on the detection scheme, such as:

- The input to the non-linear classifier must be a low dimensional vector for reducing the non-linear functions to be calculated. According to this, a feature extraction step must be included prior to the classifier. Additionally, a low dimension input space can improve the generalization ability of the classifier (Schölkopf, 02).

- The feature extraction must be linear by the same reasons.

Figure 1 shows the scheme of the proposed detector.

The aim of this communication is to evaluate different learning-based feature extraction algorithms that are designed taking into account the labels of each sample of the training set (i.e. supervised learning). Several linear schemes have been developed with the aim of

*This work has been partially supported by CICYT grant TIC2000-0380-C03-03.

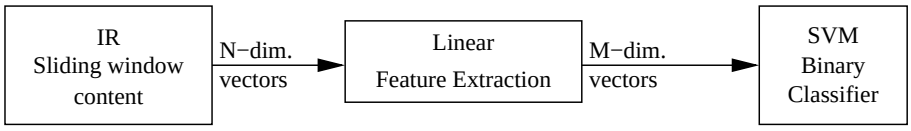

Figure 1: *Proposed scheme for detection.*

reducing the number of features without losing any relevant information for the classifier. However, supervision is needed for enhancing the discernment between the classes.

In a problem as this one, a linear classifier would be needless due to the inherent distribution of the components: pdfs (probability density functions) corresponding to both classes are expected to be overlapped (and, in fact, they are in our application), and high disparities between the variances are our main resource for a efficient decision. This is why a non-linear classifier must be utilized. To be exact, a binary SVM using RBF (Radial Basis Function) kernel has been used.

2 Feature Extraction Procedures

With the aim of reducing the dimensionality of the raw feature space, several methods have been proposed in order to transform the original vector of signals into another, with lower dimension, without missing out discriminative information to be used by the classifier. Four procedures are analyzed in this paper: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Maximization of Mutual Information (MMI) and a new variant of PCA that will be called "Supervised-PCA" (S-PCA).

PCA deals with finding the directions of the input space in which the components have the higher energy. These directions are orthogonal, and are obtained by the simple process of eigen-decomposition of the covariance matrix. This is achieved by means of the transformation

$$\tilde{x} = D^{-1/2}Vx \tag{1}$$

with x the original sample set, D the diagonal matrix containing the eigen-values of the auto-correlation matrix $R = E\{xx^T\}$ and V the set of its eigen-vectors. This gives us a set of normalized, decorrelated signals that, besides, match the principal directions. Moreover, PCA offers several properties that make it not only an interesting tool for dimension reduction itself, but it can also be used by any procedure that needs a whitened set of signals.

ICA goes further on PCA, and tries to guaranty independence in addition to decorrelation. Thus, ICA finds a rotation on the whitened signals so that the output consists of a set of statistically independent components. Some ways have been described to carry this out. By minimizing the mutual information between the outputs (Bell, 95), we can obtain a set of independent signals. This has been proved to be equivalent to the Maximum Likelihood solution, so both methods gives the same solution. Either by using high-order statistics or non-linear functions, one can also characterize the pdf of the output signals

and so try to make them as different from a gaussian distribution (they are assumed not to carry useful information) as possible (Cardoso, 97), (Hyvarinen, 00). Finally, some efforts are being focused on learning the transformation taking into account the correlation between projections applied to both the input and output signals (canonical correlations) (Borga, 97).

As it can be easily seen, these methods are very powerful for finding a small set of features from a bigger one in order to have the highest amount of useful information to face up the classifier. However, aspects related to the difference between the classes has not been considered. That is why other methods, more concerned about finding the most discriminative directions, are being intensively studied.

Information Theory (IT) provides a powerful formulation that allows us to relate output signals with their corresponding classes. Thus, MMI (Maximization of Mutual Information) tries to maximize the MI between the output components and the input labels in order to reach the highest disparity between the classes (Principe, 00), (Torkkola, 00). Nevertheless, IT is versatile enough to open interesting ways of research in this topic.

Finally, as a part of this family of supervised feature extraction methods, we propose and analyze another simple and effective procedure based on PCA, that we call "supervised PCA". As the sample set is very asymmetric, i.e. the quantity of negative samples (heat sources) is far higher than the quantity of positive ones (background), both the mean and variance of the background data are almost equal to those corresponding to the whole set. This is why the background samples are, after whitening, contained into the unit hypersphere centered at the origin. Our aim is now to find, by applying PCA again on the positive samples, their most powerful directions, i.e those along which the data are more scattered, and so easier to be separated (Fig. 2).

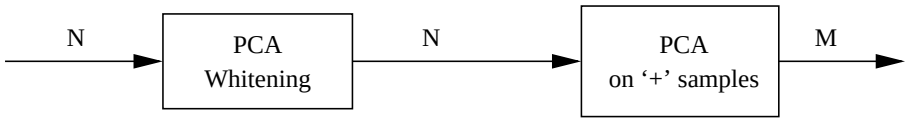

Figure 2: *Supervised PCA scheme.*

Therefore, the second stage after applying PCA as described in Ec. 1 consists of looking for another transformation that finds the principal components only in the positive samples (Fig. 3).

$$\tilde{x'} = \tilde{V}^{+}\tilde{x} \tag{2}$$

Here, $\tilde{V}^{+}$ is the set of eigenvectors obtained from the auto-correlation matrix $R^{+} = E\{\tilde{x^{+}}\tilde{x^{+}}^{T}\}$ positive samples. Normalization based on eigenvalues is not applied now because any orto-normal projection will make the sample set keep normalized.

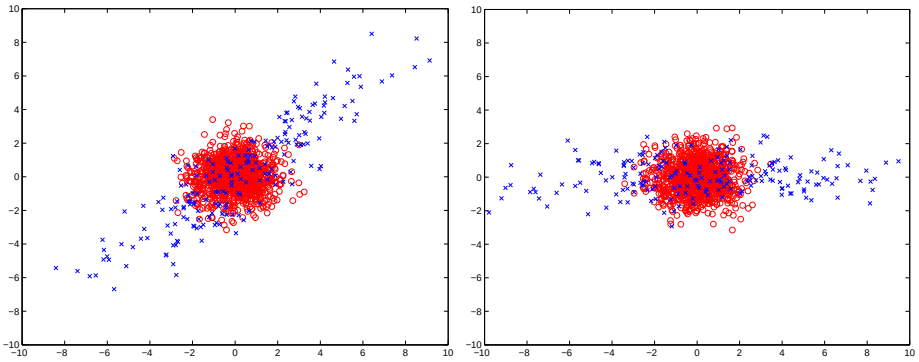

Figure 3: *Example of trivial transformation carried out by S-PCA on a 2D sample set. In the figure on the right, the most relevant direction for the classifier has been found.*

3 Results

In order to evaluate the performance of these methods, an analytic criterion as well as a classification-based one have been used.

With the aim of not depending only on the classifier for deciding the optimal components to be used (training the machine is a computational costly, slow process, specially if several components are involved and are to be iteratively chosen), a measure of the a-priori "usefulness" of each component is required. In this case, the rate $\frac{\sigma^+}{\sigma^-} \approx \frac{\sigma^+}{\sigma} = \sigma^+$ (σ^+, σ^- and σ are, respectively, the standard deviation of positive, negative and whole samples) is used for ranking the components.

As neither PCA nor ICA make use of the labels at first, supervision boils down to finding those output components which bring us the higher disparity between the classes, i.e. the higher value for the rate described above.

In order to carry out the feature selection, a "growing" process has been applied by the classifier, i.e. the components that provide the best classification results are iteratively chosen and added to the set. This mechanism have been proved to select the components that provide the higher values for σ^+ and so agrees with the analytic approach.

The test has been applied to a set of seven 90×256 infrared images with 3 bands. As the window has a size of 5×5, 75 components per pixel are generated. Feature extraction will reduce the dimension to 6. Two of the images have been used for training the classifier and the rest have been used for testing. The whole set of images have been utilized in the feature-extraction module.

In this case, every heat source is assumed to comprise an area of 5×5 pixels. Due to the windowing, 81 alarms are generated for each of these sources. Three sources and thus 243 positive vectors are available per image.

Table 1 shows the rate σ^+ for the best six components obtained by each method. As it can be seen, ICA and S-PCA seem to provide the most useful components since the values are higher than those corresponding to PCA and MMI.

However, it is in Figure 4 where a complete ROC curve with classification results from

Method	1^{st} comp.	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
PCA	6.3053	4.6980	4.3494	3.9182	3.6883	3.3277
ICA	6.5834	5.8176	5.6654	5.6600	4.3545	3.9186
MMI	6.9752	5.5872	5.5247	4.7088	4.3764	3.6436
S-PCA	7.0503	5.9657	5.6724	5.4670	3.9341	3.8613

Table 1: *Rate σ^+ of the selected components by each method.*

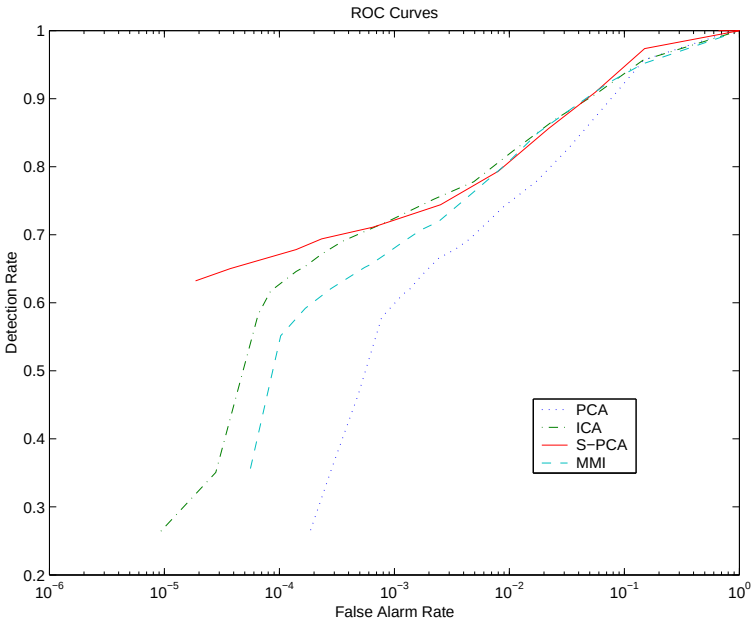

Figure 4: *Classification results. ROC curves of each method.*

each set of components is displayed. The curves have been obtained by sweeping the threshold of the resulting SVM-classifier. The limitations of PCA become obvious in the diagram, whereas the other methods yield similar performance at medium and high false alarm rates. However, if low false-alarm rates are required, S-PCA is proved to provide the best results.

4 Conclusions

ICA shows its utility and robustness even with such a complex sample set as this. If MMI does not reach its capability is due to the difficulty of building an accurate model of both the background and source pdfs as well as the problems that numerical methods are to face up in such a high dimensional space. Limitations of PCA arises from the fact that, in general, a principal component (for the whole sample set) do not have to provide a high deviation rate and so a good classification ability.

Although S-PCA outperforms these methods at low false alarm rates, its usefulness is limited by the very particular characteristics of this kind of applications. The main

assumption is the asymmetry between the quantity of positive classes and negative ones. We want the whole set of samples to be centered at the origin, drawing a unit-radius hypersphere that allows us to easily find the principal components of the positive samples. In applications related to the detection of small sources from an image, this condition usually holds.

An important advantage that must be stressed is the fact that we can always control the shape of the resulting sample set, and so the hyperparameters of the classifier become easier to establish and less sensitive to the feature extraction.

References

A.J. Bell and T.J. Sejnowski, "An Information-Maximization approach to blind separation and blind deconvolution" *Neural Computation*, 7:1129-1159, 1995.

M. Borga, H. Knutsson and T. Landelius, "Learning Canonical Correlations" *Neural Computation*, 7:1129-1159, 1995.

J-F Cardoso, "Infomax and Maximum Likelihood for Source Separation" *IEEE Signal Processing Letters*, 4:112-114, 1997.

A. Hyvärinen and E. Oja, "Independent Component Analysis: A Tutorial " *Neural Networks*, 13(4-5) 411-430, 2000.

J.C. Principe, D. Xu and J.W. Fisher III, "Information theoretic learning," In Simon Haykin, editor, *Unsupervised Adaptative Filtering*. Wiley, New York, NY, 2000.

B. Schölkopf and A. Smola , "Learning with Kernels". MIT Press, Cambridge, MA, 2002.

K. Torkkola and and W. M. Campbell, "Mutual Information in Learning Feature Transformations ," *Proc. 17th International Conf. on Machine Learning*, 2000.