SUBJECT RECOGNITION USING A NEW APPROACH FOR FEATURE SELECTION

Àgata Lapedriza [1], David Masip [2], Jordi Vitrià [3]
[1] *Computer Vision Center,Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain*
[2] *Computer Vision Center,Universitat Oberta de Catalunya, Barcelona, Spain*
[3] *Computer Vision Center, Universitat de Barcelona, Barcelona, Spain*
{*agata,davidm,jordi*}*@cvc.uab.es*

Keywords: Face Classification, Feature Extraction, Feature Selection

Abstract: In this paper we propose a feature selection method that uses the mutual information (MI) measure on a Principal Component Analysis (PCA) based decomposition. PCA finds a linear projection of the data in a non-supervised way, which preserves the larger variance components of the data under the reconstruction error criterion. Previous works suggest that using the MI among the PCA projected data and the class labels applied to feature selection can add the missing discriminability criterion to the optimal reconstruction feature set. Our proposal goes one step further, defining a global framework to add independent selection criteria in order to filter misleading PCA components while the optimal variables for classification are preserved. We apply this approach to a face recognition problem using the AR Face data set. Notice that, in this problem, PCA projection vectors strongly related to illumination changes and occlusions are usually preserved given their high variance. Our additional selection tasks are able to discard this type of features while the relevant features to perform the subject recognition classification are kept. The experiments performed show an improved feature selection process using our combined criterion.

1 INTRODUCTION

Classification problems usually deal with data that lies in high dimensional subspaces. In this general case, it has been shown that the number of samples to estimate the parameters of reliable classifiers grows exponentially with the data dimensionality. This phenomenon is known as *the curse of dimensionality* (Bellman, 1961; Duda et al., 2001) and has been mitigated with the use of dimensionality reduction techniques.

In a first approach, dimensionality reduction techniques can be classified in: (i) feature selection, where only a subset of the original features are preserved and (ii) feature extraction, where a mixture of the original features is performed. Nevertheless, its common to find hybrid approaches, where a feature selection is performed on a previously extracted features set.

Principal Component Analysis (PCA) (Kirby and Sirovich, 1990) is one of the most used feature extraction techniques, given its simplicity and optimality under the L_2 reconstruction error criterion. Briefly, PCA finds a orthogonal projection vectors computed as the first eigenvectors of the data covariance matrix, which are sorted in order to preserve the maximum possible amount of data variance. The feature selection is usually performed taking the first eigenvectors with larger eigenvalue, minimizing the reconstruction error. Nevertheless, the optimal reconstruction does not guarantee optimal classification rate.

Supervised techniques such as FLD (Fisher, 1936) or NDA (Fukunaga and Mantock, 1983) have been developed taking the class membership into account. Also, other axis selection criteria such as the use of the mutual information (MI) can be used in order to add discriminability information to the PCA bases (Guyon and Elisseeff, 2003).

In addition, some applications suffer from artifacts that mislead the classifier estimation and the feature extraction process, specially when they involve a representative amount of data variance. For example, in the face recognition case, non neutral illumination conditions or partial occlusions can imply a non accurate feature extraction step. In section 2 we formally

define a new framework for classification problems where we have one partition of the data to learn (in that case the subject partition) and another partition that is independent from the firs one (in that case the artifacts partition.

In this paper, we propose a feature selection method to add discriminant information to the PCA algorithm using the mutual information measure that is suitable for this new defined classification framework. In contrast to previous works using mutual information in feature selection, our proposal allows to add independent selection criteria, in order to neutralize, in practise, possible artifact effects that are equally present in the whole data space. Misleading relevant components not related to the classification tasks are discarded, reducing the effect of the artifacts in the data. The proposed process is detailed and discussed in section 3.

We validate our proposal in a face recognition problem using the AR Face data set. To the feature selection for subject classification task, we consider also another data partition based on the light conditions and occlusions, obtaining a PCA representation that retains only the information useful for posterior classification, filtering the misleading components present in the data space. The experiments performed, that are detailed in section 4, show significant improvements in comparison to the classic PCA approach using the first eigenvectors with larger eigenvalue, and the mutual information methods found in recent literature.

Finally, in section 5 we discuss the proposed approach and conclude the work. Moreover we suggest some future research lines related with the proposed new framework for classification.

2 PROBLEM STATEMENT

Let be X a set. Suppose that we have two partitions of this set, C and K, that is

$$X = C_1 \bigcup ... \bigcup C_a = K_1 \bigcup ... \bigcup K_b \qquad (1)$$

where $C_\alpha \bigcap C_\beta = \emptyset$ and $K_\alpha \bigcap K_\beta = \emptyset$ for all α, β. Suppose also that they are *equidistributed* in the sense that $p(C_\alpha) = p(C_\beta)$ and $p(K_\alpha) = p(K_\beta)$ for all α, β.

We call they are *independent partitions* if they accomplish the following property:

$$p(C_\alpha | K_\gamma) = p(C_\beta | K_\gamma) \qquad (2)$$

for all α, β, γ. Notice that from the Bayes Rule we have the symmetric property

$$p(K_\alpha | C_\gamma) = p(K_\beta | C_\gamma) \qquad (3)$$

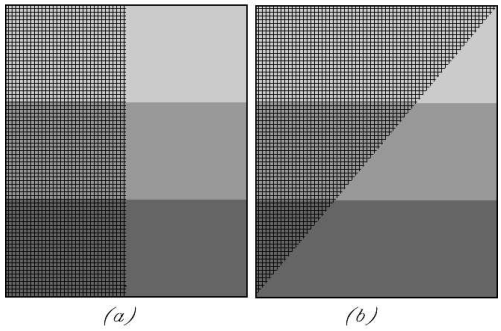

Figure 1: *(a)* Here we can see two independent partitions of the set: the one done by the grey labels (3 classes) and the other by the texture (2 classes). Notice that both are equidistributed and the property of the independent partitions' definition (also the symmetric one) is verified. *(b)* In this example two equidistributed partitions are also shown. However, in this case they are not independent. Notice, for instance, that $P(\text{rough texture}|\text{dark grey}) < P(\text{smooth texture}|\text{dark grey})$, what means that if we know information according to one of the partitions we have implicitly information according to the other one.

given that both partitions are equidistributed (in particular K).

An intuitive idea of this definition is the following: when we know the class of an element in X according one of the partitions, we do not have any information about its class according to the other partition. Figure 1 illustrates this *independence* concept for partitions in a 2-dimensional subspace.

Independent partitions of data can be found in real problems. For example, considering a set of manuscript symbols, they can be partitioned according to which symbol appears in the image (partition C) or according to the person who drew it (partition K). On the other hand, considering a set of face images having some kind of artifacts (scarfs, sunglasses, highlights or none) we can divide the set according to the subject that is in the image (partition C) or according to the appearing artifact (partition K). Then, assuming that the artefact do not depend on the subject, we have also two independent partitions of the set.

Let us focus in this second example of the faces set. In that case, subject classification is a usual task to explore in machine learning. The common procedure is to consider some labelled samples (according to C) and learn a classifier from this information.

However, suppose that we can have the training data also labelled according the partition of the arti-

facts (K). These labels are not used in the training step for subject recognition because they seem to be not interesting for the initial goal. Nevertheless, notice the following: given that the partitions C and K are independent, a source of information (a feature for instance) that is very relevant to perform a classification according C should not be relevant to perform a classification according K. The question is: can we use this information to improve the accuracy obtained when we consider only the information from the subject partition?

In this paper we explore this point and propose a feature extraction method that uses the labels from two independent partitions of the data to learn one of them.

3 FEATURE EXTRACTION AND MUTUAL INFORMATION FOR FEATURE SELECTION

In Principal Component Analysis (PCA) (Kirby and Sirovich, 1990) an orthogonal set of basis that preserve the maximum amount of data variance is obtained. In that case, the original $n-$ dimensional data can be reconstructed using d coefficients ($d \leq n$) minimizing the L_2 error. A common procedure to obtain this projection matrix is to compute the covariance matrix of the training data (previously centering each component) and find its eigenvalues and eigenvectors. By sorting the eigenvectors according to the module of they eigenvalues (largest first), we can create an ordered orthogonal basis with the first eigenvector having the direction of largest variance of the data. In that way, we have the directions in which the data set has the most significant amounts of energy and we can project our original data to a new d- dimensional subspace preserving as variance as possible from the inputs.

Although this procedure has been shown to be satisfactory in many occasions we propose in this section other possibilities for feature selection after this feature extractions process specially designed for independent partition frameworks. That is, we start from the entire set of projected features (n-dimensional data) obtained when the whole set of the eigenvectors is considered and propose to perform other feature selections taking into account not only the corresponding eigenvalues of the covariance matrix.

3.1 Mutual Information for Feature Ranking

The criteria we will use in this stage are based in the mutual information concept. The mutual information between two random variables X and Y is a quantity that measures the dependence of the two variables. We will use this statistic to define a relevance measure of the features that will be used to select them.

More concretely, consider a set of n examples $\{x_1,...,x_n\} \in X$ consisting of m input variables, $x_i = (x_{i1},...,x_{im})$ for all i, and one label per example c_i according to the partition C of the data. The mutual information between the j-th feature random variable (X_j) and the labels can be defined as

$$R(j) = \int_{X_j} \int_C p(x,c) log_2 \left(\frac{p(x,c)}{p(x)p(c)} \right) dxdc \quad (4)$$

where $p(x,c)$ is the joint probability density function of X_j and C, and $p(x)$ and $p(c)$ are the probability density functions of X_j and C respectively. Notice, the criterion $R(j)$ is a measure of dependency between the density of the variable X_j and the density of the targets C.

A common procedure for feature selection using this mutual information criterion is feature ranking (Guyon and Elisseeff, 2003). We can sort the features in a decreasing order of they mutual information with the target values and select the first d.

3.2 Proposed Feature Selection in Independent Partition Problems

Let us focus in the problem stated in section 2. Formally, we have two independent partitions of the data, C and K, and we want to select optimal features to classify this data according the criterion done by C. Suppose that we have the same framework as before but adding the labels $\{k_1,...,k_n\}$ according to the partition K. In that case we can compute for each j-th feature both mutual information values, according to C as before, $R_C(j)$ or according to K, $R_K(j)$. Given that *(a)* our goal is to classify according C and *(b)* we suppose C and K to be independent, we note that

1. a feature j_0 having high $R_K(j_0)$ is probably poorly useful for classifying according C
2. the most interesting features for classifying according C should have high R_C value and low R_K

Thus, from this observations we propose the following two approaches for selecting features for our problem

1. reject from the PCA components those features j having high $R_K(j)$ and select after this rejection the d components corresponding to the highest eigenvectors module
2. define a combined criterion for feature ranking using both R_C and R_K values and select the first d. This criterion should follow the property of being in some sense proportional to R_C and inversely proportional to R_K. In cases that the functions $R_C(j)$ and $R_K(j)$ have values of the same order we can use the substraction criteria

$$R(j) := R_C(j) - R_K(j) \tag{5}$$

Nevertheless, given that the order of the mutual information is not upper bounded and that this concept is not a distance between variables, we are not able define a general combined criterion valid for all data types and partitions.

3.3 Mutual Information estimation computation

The main drawback of the mutual information definition is that the densities are all unknown and are hard to estimate from the data, particulary when they are continuous. For this reason different approaches for mutual information estimation have been proposed in the literature (Torkkola, 2003; Bekkerman et al., 2003).

We follow in this paper a simplified estimation approach, discretizing the variables in bins. Concretely, to estimate $R_C(j)$, we discretize X_j in s bins, $\{B_1, ...B_s\}$, and performing frequencial counts on each bin. Therefore, if we have a possible classes, that is $C = C_1 \bigcup ... \bigcup C_a$, the computation is done by

$$R_C(j) = \sum_{\alpha=1}^{s} \sum_{\beta=1}^{a} p(B_\alpha, C_\beta) log \left(\frac{p(B_\alpha, C_\beta)}{p(B_\beta) p(C_\beta)} \right) \tag{6}$$

where the densities are estimated as:

$$p(C_\beta) = \frac{\#\{c_i = C_\beta\}_{i=1,...,n}}{n} \tag{7}$$

$$p(B_\alpha) = \frac{\#\{x_{ij}; x_{ij} \in B_\alpha\}_{i=1,...,n}}{n} \tag{8}$$

$$p(B_\alpha, C_\beta) = \frac{\#\{x_{ij} \in B_\alpha; c_i = C_\beta\}_{i=1,...,n}}{n} \tag{9}$$

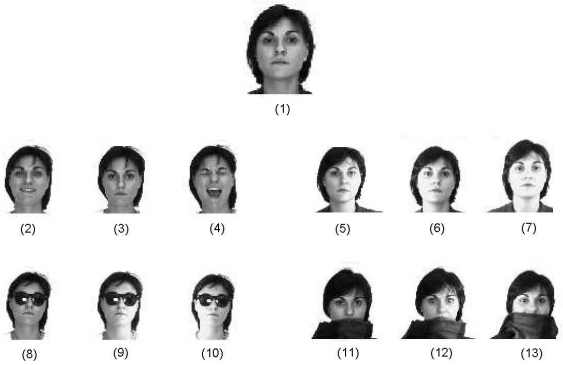

Figure 2: One sample from each of the image types in AR Face Database. The image types are the following: (1) Neutral expression, (2) Smile, (3) Anger, (4) Scream, (5) left light on, (6) right light on, (7) all side lights on, (8) wearing sun glasses, (9) wearing sun glasses and left light on, (10) wearing sun glasses and right light on, (11) wearing scarf, (12) wearing scarf and left light on, (13) wearing scarf and right light on.

4 EXPERIMENTS

The goal of this section is to illustrate and compare the proposed ideas in a real problem. To do this we perform subject recognition experiments using the publicly available ARFace database (Martinez and Benavente, 1998).

The ARFace Database is composed by 26 face images from 126 different subjects (70 men and 56 women). The images have uniform white background. The database has from each person 2 sets of images, acquired in two different sessions, with the following structure: 1 sample of neutral frontal images, 3 samples with strong changes in the illumination, 2 samples with occlusions (scarf and glasses), 4 images combining occlusions and illumination changes, and 3 samples with gesture effects. One example of each type is plotted in figure 2. We use in these experiments just the internal part of the face. The images have been aligned according the eyes and resized to be 33×33 pixels.

This database is specially appropriated to illustrate our idea given that two independent partitions of the set are done: subject classification (partition C) and image type classification (partition K). We have discussed in section 2 that these two partitions are independent.

We have performed 100 subject verification experiments following this protocol: for each subject, 13 images are randomly selected to perform the feature extraction step and the rest of the images are used to test.

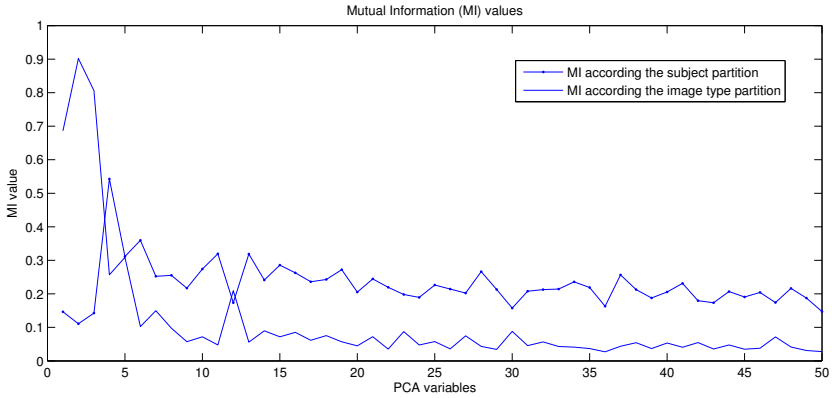

Figure 3: Mutual information of the first 50 principal components according to the subject partition (C) and image type partition (K) respectively

Three different criteria of feature extraction are considered:

- *Criterion* 1 (*CR1*): PCA and selection of the features corresponding to the high eigenvalues module
- *Criterion* 2 (*CR2*): PCA and selection of the features having more mutual information according to C (R_C)
- *Criterion* 3 (*CR3*): PCA and rejection of the features having more mutual information according to the image type (R_K)
- *Criterion* 4 (*CR4*): PCA and selection of features having higher score according the proposed criterion $R = R_C - R_K$

To estimate the mutual information we use a 4-Bin configuration (see section 3.3), that has been shown in to be an optimal discretization for this problem. After that, the classification is performed with the Nearest Neighbor.

To apply the fourth feature extraction system we need to ensure that the mutual information from the features according the subject criteria or the image type criteria have the same order. In figure 3 are plotted these values for the first 50 principal components of the original data. Notice that they are of the same order. Moreover, features having a special high value according one partition have low value according the other one, what is consistent with our hypothesis. Thus, this two observations support the use of the criterion $R = R_C - R_K$ for feature selection.

In figure 4 are plotted the mean accuracy obtained in the performed experiments at each dimensionality. On the other hand, table 4 shows the accuracies and the confidence intervals for some dimensionalities.

Notice that the most successful approach for this problem is to select principal component according to proposed combined criterion. We can see that the selection of variables that have high mutual information according the classification we want to learn, C, is specially appropriated for the initial components. However, if we use juts this criterion, from the 70th variable we are adding features that are not as suitable as the first ones. For this reason the accuracy begin to decrease. On the other hand, when we reject the features with high mutual information according the partition K it is more stable. However the results are lower in this case given that we preserve the PCA order for the components, rejecting features that are relevant to do the subject recognition task. Finally, when we use the classical PCA approach, we are preserving from the beginning features that are specially suitable to perform classification according K and rejecting some of them that are useful to classify according C. For this reason we can not achieve results as high as when we use the other approaches.

5 CONCLUSIONS AND FUTURE WORK

In this paper we introduce a new concept called *independent partitions* of a set and present a feature extraction method for a classification framework where two independent partitions of the data set are done. The method is based on Principal Component Analysis and uses the mutual information statistic to select features from the projected data. More concretely we propose a new system to rank the features obtained by the Principal Component Analysis considering the mutual information between the variables and both la-

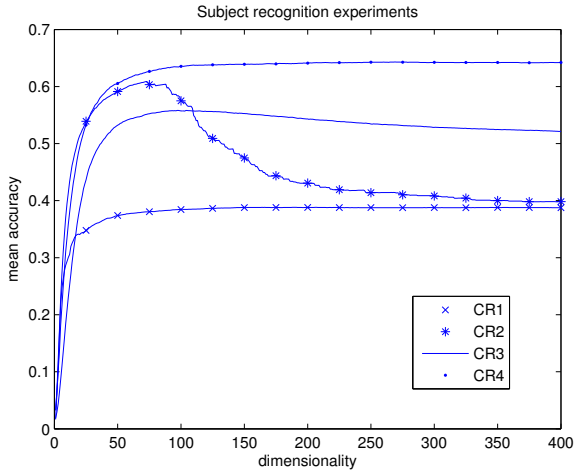

Figure 4: Mean accuracy in the performed subject recognition experiments at each dimensionality, considering the criterions for feature extraction specified above.

Table 1: Mean accuracy (in percentage) and confidence intervals of the 100 subject recognition experiments at 100, 200, 300, 400 dimensionalities respectively. The criterion of feature extraction are *CR1*, *CR2*, *CR3* and *CR4* specified above.

	100	200	300	400
CR1	38.45 ± 2.13	38.80 ± 2.31	38.78 ± 2.26	38.77 ± 2.30
CR2	57.50 ± 3.65	43.07 ± 3.24	40.83 ± 3.69	39.84 ± 3.68
CR3	55.73 ± 3.38	54.31 ± 3.35	52.85 ± 3.17	52.14 ± 3.26
CR4	63.54 ± 3.06	64.11 ± 2.89	64.24 ± 3.05	64.21 ± 3.12

bels of the data, according to each one of the independent partitions.

We perform subject recognition experiments to illustrate and test our idea and in that case the proposed method outperforms the original PCA approach in our tests.

Although the experimental results show the proposed system to be stable in that case there is future work to do in this research line. The proposed combined system, where the variables are ranked following a mutual information substraction criterion, is not enough general to be applied to any classification problem with independent partitions. Other combined criteria should be explored. Moreover, the mutual information is approximated by a simple approach. We plan to use more sophisticated algorithms to estimate it.

ACKNOWLEDGEMENTS

This work is supported by MEC grant TIN2006-15308-C02-01, Ministerio de Ciencia y Tecnologia, Spain.

REFERENCES

Bekkerman, R., El-Yaniv, R., Tishby, N., and Winter, Y. (2003). Distributional word clusters vs. words for text categorization. *J. Mach. Learn. Res.*, 3:1183–1208.

Bellman, R. (1961). *Adaptive Control Process: A Guided Tour*. Princeton University Press, New Jersey.

Duda, R., P.Hart, and Stork, D. (2001). *Pattern Classification*. Jon Wiley and Sons, Inc, New York, 2nd edition.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, 7:179–188.

Fukunaga, K. and Mantock, J. (1983). Nonparametric discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(6):671–678.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

Kirby, M. and Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108.

Martinez, A. and Benavente, R. (1998). The AR Face database. Technical Report 24, Computer Vision Center.

Torkkola, K. (2003). Feature extraction by non parametric mutual information maximization. *J. Mach. Learn. Res.*, 3:1415–1438.