Experiments about the Generalization Ability of Common Vector based methods for Face Recognition *

Marcelo Armengot, Francesc J. Ferri, and Wladimiro Díaz

Dept. d'Informàtica, Universitat de València
Dr Moliner, 50 46100 Burjassot, Spain
francesc.ferri@uv.es

Abstract. This work presents some preliminary results about exploring and proposing new extensions of common vector based subspace methods that have been recently proposed to deal with very high dimensional classification problems. Both the common vector and the discriminant vector approaches are considered. The different dimensionalities of the subspaces that these methods use as intermediate step are considered in different situations and their relation to the generalization ability of each method is analyzed. Comparative experiments using different databases for the face recognition problem are performed to support the main conclusions of the paper.

1 Introduction

There is a number of pattern classification methods that are supposed to deal with very high dimensional data by means of projecting the original data in appropriate subspaces fulfilling certain properties. In particular, some of these methods rely on the fact that dimensionality is (much) larger than the number of available samples in each class. This is usually known as the small sample size problem[1]. This is the case of the face recognition problem when faces are processed as vectors of pixels and then each face can be seen as a point in a very high dimensional space where features are supposed to correspond to specific parts of faces.

Classical approaches using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) and extensions have been applied to face recognition. One of the first methods that are worth mentioning is known as the Eigenfaces approach [2] in which PCA is used to project faces in a lower dimensional space where noise and small variations are discarded. Using LDA instead of PCA has been also proposed to keep discriminant information instead. As directly applying LDA is seldom possible in practice, many different ways of making this feasible gives rise to different approaches as Fisherfaces [3], the Null Space method [4] or the PCA+NULL method [5].

* This work has been partially supported by Spanish project DPI2006-15542-C04-04

Methods based on the concept of common vectors have been recently proposed for classification problems in high dimensional spaces like speech and face recognition [6–8]. Common vector methods can be thought of as projection methods that make all samples of a particular class to collapse on a single point (the common vector). In this subspace, the classification of test samples can be based on their (usually Euclidean) distance to the common vector in the appropriate subspace. The reason by which this distance is a good candidate to represent the degree of membership to a particular class has never been deeply studied.

The goal of this work is to analyze the two main common vector based methods and compare different options to define classification rules from common vectors and compare the results with some other subspace-based methods. The paper is organized as follows. The next section briefly presents the technicalities of the methods involved. The section 2.2 includes the different options to define the corresponding classification rules. The experimental comparison is carried out in Section 3 and final conclusion and further work is given in Section 4.

2 Subspace Methods and Common Vectors

Let suppose we have given m_k samples in $\mathbb{R}^n$ (with $m_k \leq n$) corresponding to the k-th class, $\{x_i^k\}_{i=1}^{m_k}$ and $k = 1, \ldots, c$.

Let us refer now to a particular class and drop the superscript k. It is possible to represent each x_i as the sum

$$x_i = x_o + \hat{x_i}$$

in which the common vector, x_o, represents the invariant properties common to all samples of the class and $\hat{x_i}$, called the remaining vector, represents the particular trends of this particular sample.

The decomposition above corresponds to the projection of x_i onto two orthogonal subspaces whose direct sum gives the whole representation space, $\mathbb{R}^n$.

These projections can be written in terms of the orthonormal projection operator P and its orthogonal complement $P^\perp$, where $P = UU^T$ and $U = [u_1, \ldots, u_r]$ is the $n \times r$ matrix formed with the r eigenvectors corresponding to nonzero eigenvalues of the k-th class covariance matrix, ϕ_k.

In other words, x_o is the projection of x_i onto the $n-r$-dimensional *null* space of ϕ_k and $\hat{x_i}$ is the projection onto the corresponding *range* space (of dimension r).

Instead of computing x_o as $P^\perp x_i$, it is possible to use the $P = UU^T$ that usually involves much less eigenvectors as

$$x_o = x_i - UU^T x_i$$

There is also a more convenient way of obtaining U by using the difference subspace of each class and the Gram-Schmidt ortho-normalization procedure instead of eigenanalysis [7].

The above equation projects data onto a linear subspace, Null(ϕ_k), and gives the same common vector, x_o for all samples x_i in the training set. When an unseen k-th class sample, x is projected onto the same subspace, it is assumed that a vector relatively close to x_o will be obtained.

The Common Vector approach (CVA) consists of projecting test samples to the null spaces of covariance matrices of each class and measure Euclidean distances from these projections and each one of the common vectors in each class and then assign the class as

$$\arg \min_{k=1,\ldots,c} (||x - P^k x - x_o^k||)$$

This classifier obviously gives 100% accuracy with the training set but its generalization ability depends on how unseen k-th class samples are scattered in the corresponding null spaces and how other class samples are distributed in these subspaces.

Figure 1 shows four common vectors corresponding to one of the experiments performed in Section 3.

Fig. 1. Common Vectors corresponding to 4 of the classes in the AR dataset used.

2.1 The discriminant common vector approach

The CVA is similar to the Eigenfaces method because of the fact that the projection tries to represent the information in each class in the best way. No explicit discriminant information is taken into account.

If we consider the within-class scatter matrix, S_w, the between-class scatter matrix, S_b and the total scatter matrix, $S_t = S_w + S_b$, we can think of maximizing the modified Fisher criterion [8] in the following way:

1. First project data onto the null space of S_w and obtain a (discriminant) common vector corresponding to each class.
2. Apply PCA to the common vectors in the corresponding null subspace to maximize its scatter.

In this case, a unique projection and corresponding subspace is obtained for all classes. Also, test samples are projected onto the same subspace and the corresponding (Euclidean) distances from each common vector are supposed to be representative of their class. By construction, the final subspace obtained in this case has at most $c-1$ dimensions which makes the method more appealing and opens the possibility of easily visualizing the classification problem in particular cases.

Figure 2 shows four (discriminant) common vectors obtained in the experiments performed in Section 3.

Fig. 2. Discriminant Common Vectors corresponding to 4 of the classes in the AR dataset used.

2.2 Generalization Ability of Common Vector based methods

From the formulation of common vector based methods and its properties [7, 8], it is clear that a 100% recognition rate is obtained on the training set (apart from degenerate cases and linear dependences). Once the original dimension has been decreased (explicitly or implicitly) and all training samples have collapsed to their common vectors, classification is based on the Euclidean distance or variations using angles between subspaces [7, 9]. In the Euclidean case, this relies on the (strong) assumption that test samples will be isotropically distributed around its common vector and that all of them will be closer to it than the projected samples from other classes.

The situation is very different in the two common vector based methods. In the CVA, there is a different subspace for each class. Moreover, usually the dimension of these subspaces is quite large (the number of zero eigenvalues or roughly the original dimension minus the number of training samples) which can potentially make the use of Euclidean distance useless. In the DCV there is only one subspace in which all common vector from all classes lie. Also, this subspace is of relatively low dimension (the number of classes minus one as all Fisher-based methods). This fact makes possible to illustrate its behavior for a small 3-class subproblem using data from the AR database that will be introduced in Section 3. In Figure 3 the (discriminant) common vector (bold cercles) are shown along with the decision boundaries induced by the minimum distance classifier. Several test samples per class are also shown using different symbols.

It is quite clear from the above explanations and the illustrative example, that the hypothesis of isotropic distribution is not fulfilled in general (as in most cases when LDA-based methods are successfully applied). What is worth studying is to which extent this can be a practical problem and also to start proposing ways to circumvent this kind of problems.

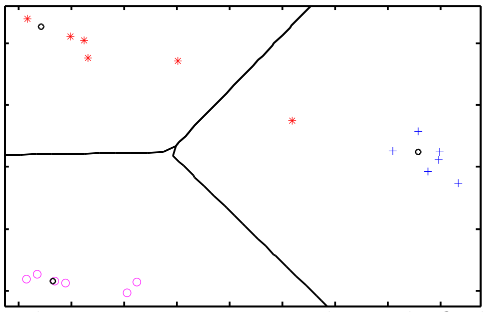

Fig. 3. Test samples and common vectors projected onto the final 2-dimensional subspace corresponding to a 3-class subproblem using the AR database.

3 Experimental Results

Three different standard and publicly available databases have been considered in this work. First, the Olivetti Research Lab (ORL) face database [10], consisting of 10 different views of 40 different individuals is considered. The images have been taken at different times and with different lighting conditions, facial expressions and details. The image resolution is 92×112 and all images are reasonably aligned.

Second, the Yale face database [11] contains a total of 165 frontal face images from 15 different people (11 images per person). Images include also different lighting conditions, expressions and other details. The resolution is 243×320 and images are approximately aligned. No preprocessing of this database has been made.

Last, the AR face database [12] is considered. This database contains 26 different views of a number of people from which 20 men and 18 women have been selected for the experiments. Faces wearing sunglasses or scarfs have been discarded, keeping only variabilities due to lighting conditions, expressions and time between pictures. The final size considered is 14 different views of 28 different people. The images have been down-sampled at 150×115 and the margins of the images with no face information have been trimmed. One example from each one of the databases considered is shown in Figure 4.

All databases have been used in the same way, different number of training faces (3, 4, 6, 8 and 9) have been considered and the remaining ones have been used for testing to obtain holdout estimates of the recognition accuracy of each method. Holdout experiments have been repeated 8, 7, 6, 5 and 5 times (depending on the number of training faces) and the results have been averaged. All accuracy results shown correspond then to averaged holdout estimates of the expected accuracy of each method.

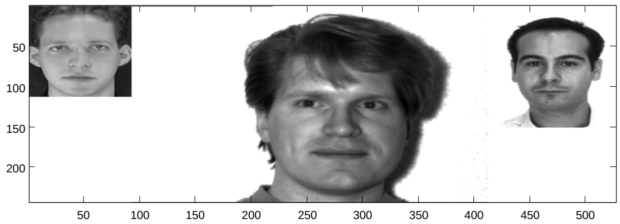

Fig. 4. Example faces from each of the databases considered in this work. From left to right, ORL, Yale, and AR.

Apart from the Common Vector approach (CVA) and the Discriminant Common Vector method (DCV), other well-known related methods have been considered for the experiments. In particular, Eigenfaces [2], LDA [11] and Fisherfaces [3] have been implemented and tested on the same data for comparison purposes. The data has been normalized to have zero mean and unit variance and 95% of the total energy is preserved when using the Eigenfaces method [11]. The dimension has been reduced to $c(m-1)$ using PCA in the Fisherfaces method and standard regularization techniques have been used to apply the LDA. The nearest neighbor method using the Euclidean distance has been used in the reduced subspaces to assign the definitive label in all three methods.

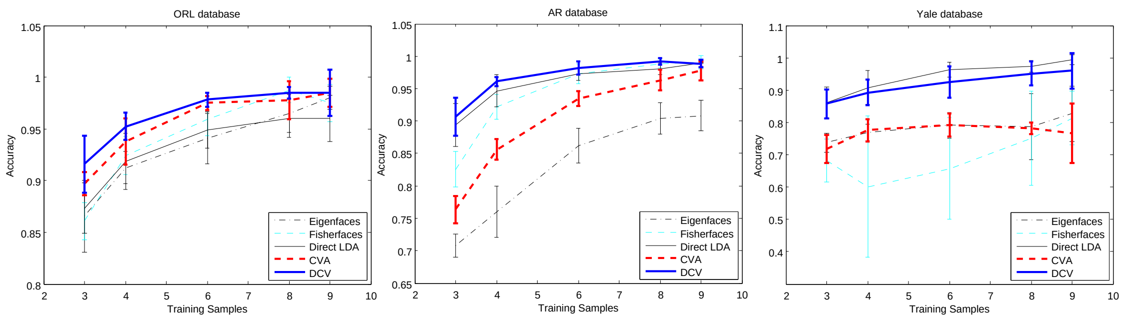

Fig. 5. Accuracies obtained with the three databases for the different methods considered and different number of training samples.

The results obtained with the five subspace based methods considered on the three standard databases in the conditions explained above are shown in Figure 5 in one graph per database. Standard deviations of the averaged holdout estimates of the accuracy are also displayed.

As can be seen in the figure, all methods give very similar results (no significant differences) in the case of the (easy) ORL database.

A significant difference can be observed between the two methods based on common vectors in the other two databases. As it could be expected, the results confirm that discriminant common vectors consistently gives better results than the plain common vector approach.

Relatively surprising is the fact that some of the standard methods like (regularized) LDA and Fisherfaces (in AR database only) give very similar (insignificantly better for the Yale database) results than the DCV method.

From the computational point of view, the methods considered have very different behaviors both in training and testing phases. Although the implementations used (using a very well known matrix based prototyping software system) are by no means optimized or programmed in a consistent way that allows strict comparison, approximate CPU times have been measured for illustrative purposes. These time measurements are shown in Figure 6. The most outstanding trend in this graph is the high computational efficiency of the DCV method both in training and testing. The CVA approach shows the more asymmetric behavior: it is the most efficient in training but by far the slower in testing.

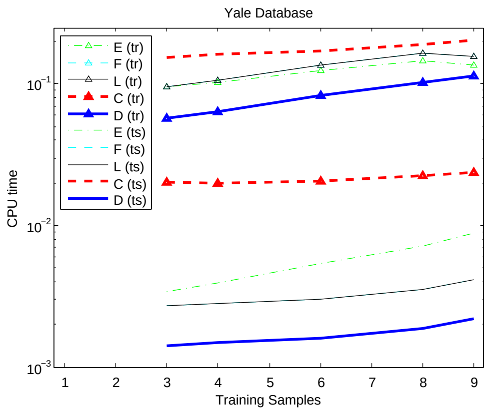

Fig. 6. Approximate CPU times for training (tr) and testing (ts) phases of each of the methods considered: Eigenfaces (E), Fisherfaces (F), LDA (L), Common Vectors (C) and Discriminant Common Vectors (D).

To study the ability of the different methods to properly generalize, the above experiments have been repeated but adding increasing amounts of noise to the images. In particular, salt and pepper noise and random (small) rotations and blurring have been considered only with the AR database. The results obtained are shown in Figure 8 only for 6 training samples. Note that the results for 0 level noise in both cases are just another random realization of the ones shown in Figure **??**. In the case of salt and pepper noise, the noise level means the probability of having a corrupted pixel. In the second case, noise level means the amount of degrees (either positive or negative) that a random rotation of the image can have. Combined with the maximum rotation, a plain average mask of sizes ranging from 2 to 5. The particular settings used for the results in Figure 8 are shown in Table 1, and an example of an original image from the AR database and some maximally corrupted images is shown in Figure 7.

Table 1. Settings to generate corrupted images for the experiments to assess generalization ability.

Noise type	Noise level					
Salt and pepper	0	0.1	0.2	0.3	0.4	0.5
Rot.& blur (degrees)	0	3	6	9	12	
Rot.& blur (mask size)	1	2	3	4	5	

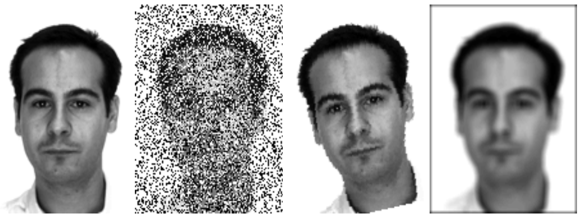

Fig. 7. Example of maximally corrupted images used in the experiment. From left to right: original, salt and pepper (0.5), rotated (12^o) and blurred (5×5 mask).

From this experiment, it is worth noting the ability of the DCV method to deal with noise. Even when all other methods exhibit a dramatic drop in their accuracy (with a very very high level of salt and pepper noise), the DCV still gives very similar results. The differences among methods when a combination of blurring and small random rotations are are used is not as significant.

4 Concluding Remarks

In this work, a comparative experimentation of common vector based approaches and related subspace based methods for face recognition have been considered. The main features, advantages and drawbacks of these methods have been put forward and their potential generalization ability has been studied. The experimental setting has been designed on one hand to compare the behavior and computational efficiency of all methods in several standard databases, and also to open the way to study the generalization ability of the different approaches. Experiments with increasing levels of different types of noise have been also conducted to study the way in which the accuracy degrades. The main preliminary conclusion is that the DCV method is the best option from the different point of views. From the particular point of view of generalization, the experiments performed in this work are clearly insufficient. More experimentation and some particular proposals to introduce more robust classification rules in the transformed subspaces are currently on its way.

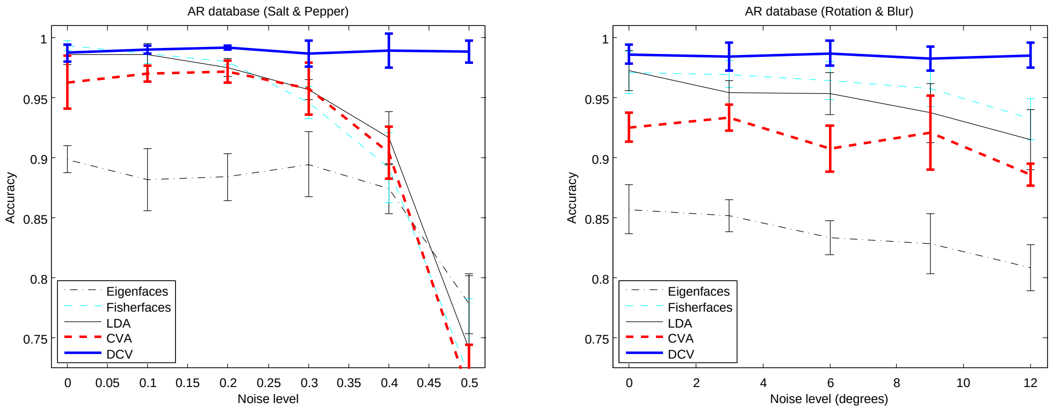

Fig. 8. Results obtained with the AR database with increasing levels of a) salt and pepper and b) rotation and ll blur noise using all the classification methods considered.

References

1. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press (1990)
2. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience **3** (1991) 71–86
3. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19** (1997) 711–720
4. Chen, L., Liao, H.M., Ko, M., Lin, J., Yu, G.: A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognition **33** (2000) 1713–1726
5. Huang, R., Liu, Q., Lu, H., Ma, S.: Solving the small sample size problem of lda. In: 16th International Conference on Pattern Recognition. Volume 3. (2002) 29–32
6. Gulmezoglu, M.B., Dzhafarov, V., Keskin, M., Barkana, A.: A novel approach to isolated word recognition. IEEE Transactions on Speech and Audio Processing **7** (1999) 620–628
7. Gulmezoglu, M.B., Dzhafarov, V., Barkana, A.: The common vector approach and its relation to principal component analysis. IEEE Transactions on Speech and Audio Processing **9** (2001) 655 – 662
8. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: Discriminative common vectors for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 4 – 13
9. He, Y.H., Zhao, L., Zou, C.R.: Face recognition using common faces method. Pattern Recognition **39** (2006) 2218–2222
10. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. In: 2nd IEEE Workshop on Applications of Computer Vision. (1994)
11. Swets, D., Weng, J.: Using discriminant eigenfeatures for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **18** (1996) 831–836
12. Martinez, A., Benavente, R.: The AR face database. Technical Report 24, Computer Vision Center, Barcelona (1998)