

A COMPARISON OF SUBSPACE ANALYSIS FOR FACE RECOGNITION

Jian Li, Shaohua Zhou, and Chandra Shekhar

Center for Automation Research
ECE Department, University of Maryland
College Park, MD 20742
{lij, shaohua, shekhar}@cfar.umd.edu

ABSTRACT

We report the results of a comparative study on subspace analysis methods for face recognition. In particular, we have studied four different subspace representations and their 'kernelized' versions if available. They include both unsupervised methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), and supervised methods such as Fisher Discriminant Analysis (FDA) and probabilistic PCA (PPCA) used in a discriminative manner. The 'kernelized' versions of these methods provide subspaces of high-dimensional feature spaces induced by non-linear mappings. To test the effectiveness of these subspace representations, we experiment on two databases with three typical variations of face images, i.e. pose, illumination and facial expression changes. The comparison of these methods applied to different variations in face images offers a comprehensive view of all the subspace methods currently used in face recognition.

1. INTRODUCTION

Subspace analysis is often used in signal processing and computer vision problem as an efficient method for both dimension reduction and finding the direction of the projection with certain properties. Usually, the vector constructed by raster-scanning the face image is considered to lie in a high-dimensional vector space. In the context of face recognition [1], we attempt to find some basis vectors in that space serving as directions of projection, and hopefully the projected data are clustered according to their class labels. Traditional linear subspace representations, namely Principal Component Analysis (PCA) [2], Fisher Discriminant Analysis (FDA) [3, 4], and Independent Component Analysis (ICA) [5, 6] have been implemented, where PCA and ICA are unsupervised methods and FDA is supervised. However, all these methods are linear, which limits their applicability. Recently, their 'kernelized' versions, namely

Kernel PCA (KPCA) [7], Kernel FDA (KFDA) [8] and Kernel ICA (KICA) [9], have appeared in the literature where the limitation of linearity is overcome by construction of a high-dimensional feature space induced by a nonlinear mapping. Probabilistic PCA (PPCA) can be regarded as a revised PCA with some probabilistic flavor incorporated. Even though PPCA in its original form is an unsupervised approach, we implement it in a supervised manner by introducing inter-personal space [10]. While all of the above methods are commonly found in the face recognition literature [2, 3, 4, 6, 10, 11], they have not been discussed in a framework of subspace analysis for comparison. This paper attempts to accomplish this task in both theory and application.

The rest of the paper is as follows. Section 2 reviews the underlying theories of each subspace analysis technique and Section 3 describes the 'kernelized' subspace methods. Section 4 gives the experimental methodology. Section 5 presents the experimental results using two different databases with three variations, namely pose, illumination and facial variations. Section 6 concludes the paper.

2. SUBSPACE METHODS

The basic framework of subspace analysis is as follows. Suppose we have n d -dimensional training vectors, forming a matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$. Later we will see that in most cases, these training vectors are formed by raster-scanning the training face images, but in PPCA, the concept of intra-personal space is introduced, and the vectors are formed from the difference of face images belonging to one person. These vectors are preprocessed to have zero-mean and unit-variance. Since original vector dimension d is usually very large, we attempt to find m basis vectors ($m < d$) forming a matrix $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_m]$ in \mathbb{R}^d , such that the new representation,

$$\mathbf{Y} = \mathbf{W}^T \mathbf{X}, \quad (1)$$

satisfies certain properties. And different properties give rise to different kinds of analysis methods such as PCA,

Partially supported by the DARPA Grant N00014-00-1-0908.

ICA, FDA and PPCA.

Since the ‘eigenface’ approach was proposed by Turk and Pentland [2], PCA has emerged as a popular technique in the computer vision community. Variants of PCA techniques have been studied and used [7, 11, 13]. Linear PCA is the simplest version. It decomposes the available data into uncorrelated directions, along which there exist the maximum variations. In other words, it tries to minimize the representation error $\|\mathbf{W}\mathbf{Y} - \mathbf{X}\|$. Towards this goal, a total scatter matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ is defined and the optimal matrix \mathbf{W} is formed by the eigenvectors corresponding to the m largest eigenvalues of \mathbf{S} .

In contrast to PCA which makes a decomposition into uncorrelated components, ICA [5, 6] decomposes the data into statistically independent components. Usually a contrast function measuring the statistical dependence of the new representation y_1, \dots, y_m is defined and minimized. ICA turns out to be a non-linear minimization problem which requires a lot of computations.

While component analysis is oriented towards representing the data, discriminant analysis keeps in mind the classification task. It attempts to maximize the between-class scatter while minimizing the within-class scatter. In FDA [4, 3], two scatter matrices are defined: between-class scatter matrix \mathbf{S}_B and within-class scatter matrix \mathbf{S}_W [12]. In linear FDA, we want to maximize

$$J(\mathbf{W}) = \frac{|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|}. \quad (2)$$

And the optimal \mathbf{W} formed by generalized eigenvectors that correspond to the largest eigenvalues in

$$\mathbf{S}_B \mathbf{W}_i = \mu_i \mathbf{S}_W \mathbf{W}_i, \quad i = 1, \dots, C - 1, \quad (3)$$

where C is the number of classes.

As a variant of PCA, PPCA has some probabilistic flavor. It assumes there is some latent variable \mathbf{y} related to sample vector \mathbf{x} through

$$\mathbf{x} = \mathbf{u} + \mathbf{W}\mathbf{y} + \mathbf{e}, \quad (4)$$

where \mathbf{u} is sample mean, \mathbf{W} is the *loading matrix*, the latent variable $\mathbf{y} \sim \mathbf{N}(0, \mathbf{I})$, and $\mathbf{e} \sim \mathbf{N}(0, \sigma^2 \mathbf{I})$ is measurement noise vector. This model is a special case of Factor Analysis (FA) [12]. The unknown \mathbf{W} and σ^2 are estimated using the Maximum Likelihood principle [13]. The resulting estimate of \mathbf{W} and σ^2 obey the coupled equations:

$$\mathbf{W} = \mathbf{U}_m (\mathbf{D}_m - \sigma^2 \mathbf{I})^{1/2} \mathbf{R}, \quad (5)$$

$$\sigma^2 = \frac{1}{d - q} \sum_{i=q+1}^d \lambda_i, \quad (6)$$

where \mathbf{R} is any orthogonal matrix, $\mathbf{D}_m = \text{diag}[\lambda_1, \dots, \lambda_m]$ is a diagonal matrix whose diagonal elements are the m

largest eigenvalues of the total scatter matrix \mathbf{S} . The \mathbf{U}_m matrix is formed by the eigenvectors corresponding to those eigenvalues. A suboptimal approach [10] is to set $\sigma^2 = 0$ and $\mathbf{R} = \mathbf{I}$ in Eq. (5). Also in this approach, if for certain x its principal component is $\hat{x} = \mathbf{W}^T x = [\hat{x}_1, \dots, \hat{x}_m]^T$, it can be shown that,

$$-2 \log(P(x)) \propto \sum_{i=1}^m \frac{\hat{x}_i^2}{\lambda_i} + \frac{\epsilon^2}{\sigma^2}, \quad (7)$$

where ϵ^2 is the MSE of using \hat{x} to represent x and λ_i is the i -th largest eigenvalue of the scatter matrix. So in PPCA, PCA is used to find the optimal load matrix, and $P(x)$ will be used for classification. We have seen clearly that FDA is a supervised method because it incorporates the class information. PPCA can be operated in a supervised fashion by introducing the concept of intra-personal space (IPS) [10], which is constructed by collecting all pixel-wise difference of face images belonging to the same person. The PPCA density is then fitted on top of the IPS. So given a probe image, we can first calculate the difference images between the probe image and the gallery images, then $P(x)$'s are computed for all difference images as in Eq. (7), and finally the classifier associates the identity with the one yielding the largest $P(x)$.

3. KERNELIZED SUBSPACE METHODS

Kernel-based methods utilize the fact that the cost functions mentioned in section 2 can be expressed in terms of dot product ($x_i \cdot x_j = x_i^T * x_j$), and by replacing the dot product with kernel function $K(x_i, x_j)$, the original data is mapped non-linearly into a high-dimensional or infinite dimensional features space. However, explicit knowledge on the non-linear mapping is not required because it is embedded in the kernel function. In our experiment, we use the Gaussian kernel function

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right). \quad (8)$$

The detailed description of KPCCA [12], KICA [9], KFDA [8] can be found in references.

4. EXPERIMENTAL SETTING

Our experiments are carried out on two face databases: AT&T and FERET, with a special emphasis on testing the effects of various changes in face images, such as variations of pose, illumination and facial expressions. We also test the generality of different methods by making training and testing sets non-overlapping.

For the AT&T database, instead of using the ‘leave-one-out method’ for testing the effectiveness of the subspace



Fig. 1. Examples of AT&T database. Downsampled to 28x23. The database contains 40 classes, 10 images per class.

methods, we randomly take 5 images from each class as the training data and leave the rest 5 images as the probe. Such test is run five times and we take the average of the results for comparison. Example images are shown in Fig. 1.

For the FERET database, two experiments are done, one for facial expression and one for illumination. Examples images are shown in Fig. 2. To avoid the overlap between training and probe sets, we randomly divide the 600 images into 2 sets, each with 300 images belonging to 100 classes. Two different IPS's are constructed separately for facial expression and illumination. Classification is then performed according to the procedure described in the end of Section 2.

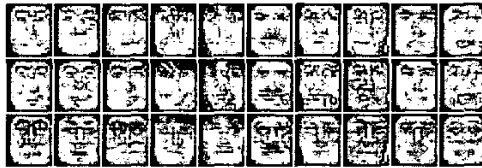


Fig. 2. Examples of FERET database. Downsampled to 24x21. The database contains 200 classes, 3 images per class. First row are images of neutral face, followed by a row of images with expression variations. The last row are images with illumination changes.

5. EXPERIMENTAL RESULTS

5.1. Performance Comparison

We use the cumulative match score (CMS) curve and the recognition rate within the top one match as criteria in comparison. Recognition rate is shown in Fig. 3. CMC curves of these subspace methods under three different variations are shown in the Fig. 4.

Recognition Rate (%)	PCA	KPCA	ICA	KICA	FDA	KFDA	PPCA
Pose	92.2	90.1	88.2	88	89.1	95.3	81.4
Facial Expression	61	61	53	46	72	73	68
Illumination	57	61	57	59	75	65	72

Fig. 3. Recognition rate under each condition.

In pose variation case, the results are much better than the other two cases because the class which the probe image

belongs to has five training samples in stead of one in the other two cases. In this case, it can be seen in performance $PCA > FDA > ICA > PPCA$, and for the kernel methods, $KFDA > KPCA > KICA$. PCA outperforms FDA since in pose variation case, the within class scatter is hard to minimize. PPCA does well in the situation that variations similar to those in the probe images are learned in the training process. However, it is outperformed here partly because pose variations may take very different forms for every image.

For facial expression changes, the CMS curves for all linear methods almost overlap, though their top one matches are slightly different. For the kernel method, $KFDA > KPCA > KICA$.

As to illumination changes, the CMS curves show that $FDA > PPCA > PCA > ICA$ and $KFDA \geq KICA > KPCA$. PPCA does well in this case, since during training, illumination variation has been learned from the first half of the database. Here supervised methods exhibits its advantage when the number of samples for each class is small.

The recognition rates of the four subspace methods, averaging the kernel/nonkernel versions, are compared in Fig. 5. The average recognition rates for the overall kernel-based methods versus non-kernel methods are also shown in Fig. 5. We can observe that kernel-based methods produce similar results as non-kernel ones.

To sum up, overall method-wise comparison shows $ICA < PCA < PPCA < FDA$. PPCA and FDA are better partly because of their discriminative power embedded in training stage. We also observe that kernel-based methods are not necessarily better than non-kernel methods, which might imply that second-order statistics are enough in a face recognition problem. A final note is that since our experiment is set to test the generality of different methods, our recognition rates are not as high as those obtained by other methods such as 'leave-one-out'.

5.2. Comparison of Computational Load

This concerns training time and testing time. For most methods, although training might be time-consuming, such as in ICA because of non-linear optimization involved, testing time is rather short since it only requires simple matrix calculations.

For training time, $PCA < LDA < PPCA < ICA$ and $KPCA < KLDA \ll KICA$. Testing time: $PCA = LDA = ICA = PPCA$ and $KPCA = KLDA = KICA$.

6. CONCLUSION

We have compared several subspace methods commonly used in the face recognition literature. Their performances under different variations are also shown. The concept of intra-personal space is introduced during the application of

PPCA. Currently most subspace methods are done over the original space of face images, instead of over intra-personal space. Subspace methods over intra-personal or inter-personal spaces need to be studied.

7. ACKNOWLEDGEMENT

We are grateful to Prof. Rama Chellappa for helpful discussions.

8. REFERENCES

- [1] W. Y. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips, "Face recognition: A literature survey," *UMD CfAR Technical Report CAR-TR-948*, 2000.
- [2] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 72–86, 1991.
- [3] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of Optical Society of America A*, pp. 1724–1733, 1997.
- [4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, 1997.
- [5] A. Hyvarinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.
- [6] M.S. Barlett, H. M. Lades, and T. J. Sejnowski, "Independent component representations for face recognition," *Proc. SPIE 3299*, pp. 528–539, 1998.
- [7] B. Schelkopf, A. Smola, and K. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.
- [8] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. 1999, pp. 41–48, IEEE.
- [9] F. Bach and M. I. Jordan, "Kernel independent component analysis," *Technical Report CSD-01-1166, Computer Science Division, University of California, Berkeley*, 2001.
- [10] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *IEEE Trans. PAMI*, vol. 24, no. 6, pp. 780–788, 2002.
- [11] M.-H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods," *Proc. of Intl. Conf. on Face and Gesture Recognition*, 2002.
- [12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2001.
- [13] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

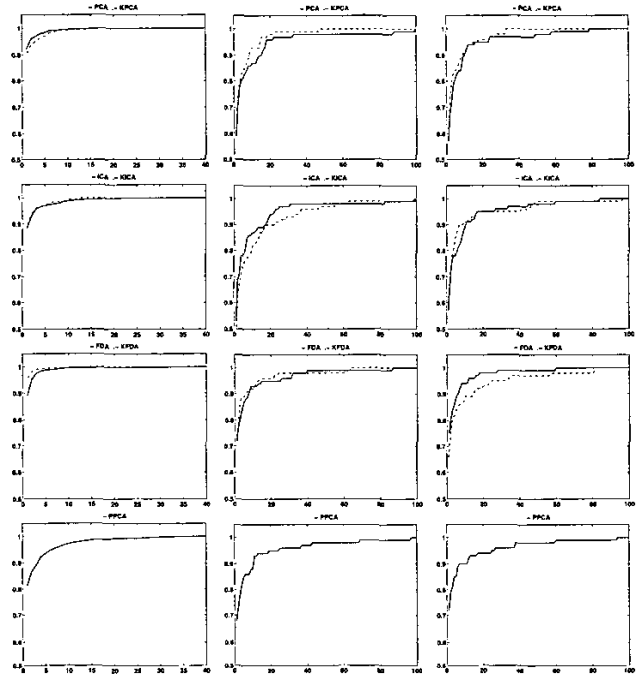


Fig. 4. CMS curves for pose (left column), facial expression (middle column) and illumination (right column) variations. 1st row: PCA and KPCA. 2nd row: ICA and KICA. 3rd row: FDA and KFDA. 4th row: PPCA.

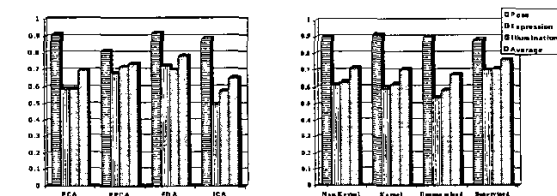


Fig. 5. Average recognition rate of (left) different subspace methods and (right) nonkernel/kernel, supervised/unsupervised methods, under different variations.