In Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 215–220, 2002

Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods

Ming-Hsuan Yang
Honda Fundamental Research Labs
Mountain View, CA 94041
myang@hra.com

Abstract

Principal Component Analysis and Fisher Linear Discriminant methods have demonstrated their success in face detection, recognition and tracking. The representations in these subspace methods are based on second order statistics of the image set, and do not address higher order statistical dependencies such as the relationships among three or more pixels. Recently Higher Order Statistics and Independent Component Analysis (ICA) have been used as informative representations for visual recognition. In this paper, we investigate the use of Kernel Principal Component Analysis and Kernel Fisher Linear Discriminant for learning low dimensional representations for face recognition, which we call Kernel Eigenface and Kernel Fisherface methods. While Eigenface and Fisherface methods aim to find projection directions based on second order correlation of samples, Kernel Eigenface and Kernel Fisherface methods provide generalizations which take higher order correlations into account. We compare the performance of kernel methods with classical algorithms such as Eigenface, Fisherface, ICA, and Support Vector Machine (SVM) within the context of appearance-based face recognition problem using two data sets where images vary in pose, scale, lighting and expression. Experimental results show that kernel methods provide better representations and achieve lower error rates for face recognition.

1 Motivation and Approach

Subspace methods have demonstrated their success in numerous visual recognition tasks such as face detection, face recognition, 3D object recognition, and tracking. In particular, Principal Component Analysis (PCA) [30] [17], and Fisher Linear Discriminant (FLD) [6] methods have been applied to face recognition with impressive results. While PCA aims to extract a subspace in which the variance is maximized (or the reconstruction error is minimized), some unwanted variations (due to changes in lighting, facial expressions, viewing points, etc.) may be retained (See [8] [10] for examples). It has been observed that in face recognition the variations between the face images of the same person due to illumination and viewing direction are almost always larger than image variations due to the changes in face identity [1]. Therefore, while the PCA projections are optimal in a correlation sense (or for reconstruction from a low dimensional subspace), these eigenvectors or bases may be suboptimal from the classification viewpoint.

Representations of Eigenface [30] (based on PCA) and Fisherface [6] [32] [27] (based on FLD) methods encode pattern information based on second order dependencies, i.e., pixel-wise covariance among the pixels, and are insensitive to the dependencies among multiple (more than two) pixels in the samples. Higher order dependencies in an image include nonlinear relations among the pixel intensity values, such as the relationships among three or more pixels in an edge or a curve, which may capture important information for recognition. Several researchers have conjectured that higher order statistics may be crucial to better represent complex patterns. Recently, Higher Order Statistics (HOS) have been applied to visual learning problems. Rajagopalan et al. used HOS of the images of a target object to get a better density estimation. Experiments on face detection [22] and vehicle detection [21] showed comparable, if not better, results than PCA-based methods.

The concept of Independent Component Analysis (ICA) maximizes the degree of statistical independence among output variables using contrast functions such as Kullback-Leibler divergence, negentropy and cumulants [11]. A neural network algorithm (i.e., infomax learning rule) to carry out ICA was proposed by Bell and Sejnowski [7], and was applied to face recognition [3]. Although the idea of extracting higher order (nonlinear) statistical dependencies in the ICA-based face recognition method is attractive, the assumption that the face images comprise of a set of independent basis images (or factorial code) is not intuitively clear. In [3] Bartlett et al. showed that ICA-based approach outperform PCA-based method in face recognition using a subset of frontal view FERET face images. However, Moghaddam showed that ICA-based

approach does not provide significant advantage over PCA-based method [16]. The experimental results suggest that seeking non-Gaussian and independent components may not necessarily provide a better representation for face recognition.

One reason for the recent success of Support Vector Machine (SVM) algorithms is the kernel trick which provides an efficient way to compute nonlinear features of samples, thereby yielding a rich representation (or one can view this as projecting samples from an input space to a higher dimensional feature space). Nevertheless, it is clear that not all these nonlinear features are essential for recognition or classification purpose in most applications (See also [31] on feature extraction for SVM). In [25], Schölkopf et al. extended the classical PCA to Kernel Principal Component Analysis (KPCA). Empirical results on digit recognition using MNIST data set and object recognition using a chair database showed that Kernel PCA is able to extract nonlinear features and thus provided better recognition results. Baudat and Anouar [5], Roth and Steinhage [23], and Mika et al. [15] applied the kernel trick to FLD and proposed Kernel Fisher Linear Discriminant (KFLD) method. Their experiments showed that KFLD is able to extract the most discriminant features in the feature space, which is equivalent to extracting the most discriminant nonlinear features in the original input space.

In this paper we seek a method that not only extracts higher order statistical dependencies of samples as features, but also maximizes the class separation when we project these features to a lower dimensional space for efficient recognition. Since much of the important information may be contained in the high order dependencies among pixels of a face image, we investigate the use of Kernel PCA and Kernel FLD for face recognition, which we call Kernel Eigenface and Kernel Fisherface methods, and compare their performance against the standard Eigenface, Fisherface, ICA-based and SVM-based methods. In the meanwhile, we explain why kernel methods are suitable for visual recognition tasks such as face recognition.

2 Kernel Principal Component Analysis

Given a set of m centered (zero mean, unit variance) samples $\mathbf{x}_k$, $\mathbf{x}_k = [x_{k1}, \ldots, x_{kn}]^T \in R^n$, PCA aims to find the projection directions that maximize the variance of a subspace which is equivalent to finding the eigenvalues from the covariance matrix, C,

$$\lambda \mathbf{w} = C\mathbf{w} \tag{1}$$

for eigenvalues $\lambda \geq 0$ and eigenvectors $\mathbf{w} \in R^n$. In Kernel PCA, each vector $\mathbf{x}$ is projected from the input space, R^n, to a high dimensional feature space, R^f, by a nonlinear mapping function: $\Phi : R^n \rightarrow R^f, \quad f > n$. Note that the dimensionality of the feature space can be arbitrarily large. In R^f, the corresponding eigenvalue problem is

$$\lambda \mathbf{w}^\Phi = C^\Phi \mathbf{w}^\Phi \tag{2}$$

where C^Φ is a covariance matrix. All solutions $\mathbf{w}^\Phi$ with $\lambda \neq 0$ lie in the span of $\Phi(\mathbf{x}_1)$, ..., $\Phi(\mathbf{x}_m)$, and there exist coefficients α_i such that

$$\mathbf{w}^\Phi = \sum_{i=1}^{m} \alpha_i \Phi(\mathbf{x}_i) \tag{3}$$

Denoting an $m \times m$ matrix K by

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \tag{4}$$

, the Kernel PCA problem becomes

$$m\lambda K\boldsymbol{\alpha} = K^2\boldsymbol{\alpha} \quad \equiv \quad m\lambda\boldsymbol{\alpha} = K\boldsymbol{\alpha} \tag{5}$$

where $\boldsymbol{\alpha}$ denotes a column vector with entries $\alpha_1, \ldots, \alpha_m$. The above derivation assume that all the projected samples $\Phi(\mathbf{x})$ are centered in R^f. See [25] for a method to center the vectors $\Phi(\mathbf{x})$ in R^f.

Note that classical PCA is a special case of Kernel PCA with first order polynomial kernel. In other words, Kernel PCA is a generalization of classical PCA since different kernels can be utilized for different nonlinear projections.

We can now project the vectors in R^f to a lower dimensional space spanned by the eigenvectors $\mathbf{w}^\Phi$, Let $\mathbf{x}$ be a test sample whose projection is $\Phi(\mathbf{x})$ in R^f, then the projection of $\Phi(\mathbf{x})$ onto the eigenvectors $\mathbf{w}^\Phi$ is the nonlinear principal components corresponding to Φ:

$$\mathbf{w}^\Phi \cdot \Phi(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})) = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \tag{6}$$

In other words, we can extract the first q ($1 \leq q \leq m$) nonlinear principal components (i.e., eigenvectors $\mathbf{w}^\Phi$) using the kernel function without the expensive operation that explicitly projects samples to a high dimensional space R^f. The first q components correspond to the first q non-increasing eigenvalues of (5). For face recognition where each $\mathbf{x}$ encodes a face image, we call the extracted nonlinear principal components Kernel Eigenfaces.

3 Kernel Fisher Linear Discriminant

Similar to the derivations in Kernel PCA, we assume the projected samples $\Phi(\mathbf{x})$ are centered in R^f (See [25] for a method to center the vectors $\Phi(\mathbf{x})$ in R^f),

we can formulate the equations that use dot products for FLD only. Denoting the within-class and between-class scatter matrices by S_W^Φ and S_B^Φ, and applying FLD in kernel space, we need to find eigenvalues $\boldsymbol{\lambda}$ and eigenvectors $\mathbf{w}^\Phi$ of

$$\lambda S_W^\Phi \mathbf{w}^\Phi = S_B^\Phi \mathbf{w}^\Phi \tag{7}$$

, which can be obtained by

$$W_{OPT}^\Phi = \arg\max_{W^\Phi} \frac{|(W^\Phi)^T S_B^\Phi W^\Phi|}{|(W^\Phi)^T S_W^\Phi W^\Phi|} = [\mathbf{w}_1^\Phi \dots \mathbf{w}_m^\Phi] \tag{8}$$

where $\{\mathbf{w}_i^\Phi | i = 1, 2, \dots, m\}$ is the set of generalized eigenvectors corresponding to the m largest generalized eigenvalues $\{\lambda_i | i = 1, 2, \dots, m\}$.

Consider a c-class problem and let the r-th sample of class t and the s-th sample of class u be $\mathbf{x}_{tr}$ and $\mathbf{x}_{us}$ respectively (where class t has l_t samples and class u has l_u samples), we define the kernel function:

$$(k_{rs})_{tu} = k(\mathbf{x}_{tr}, \mathbf{x}_{us}) = \Phi(\mathbf{x}_{tr}) \cdot \Phi(\mathbf{x}_{us}) \tag{9}$$

Let K be a $m \times m$ matrix defined by the elements $(K_{tu})_{u=1,\dots,c}^{t=1,\dots,c}$, where K_{tu} is a matrix composed of dot products in the feature space R^f, i.e.,

$$K = (K_{tu})_{u=1,\dots,c}^{t=1,\dots,c} \text{ where } K_{tu} = (k_{rs})_{s=1,\dots,l_u}^{r=1,\dots,l_t} \tag{10}$$

Note K_{tu} is a $l_t \times l_u$ matrix, and K is a $m \times m$ symmetric matrix. We also define a matrix Z:

$$Z = (Z_t)_{t=1,\dots,c} \tag{11}$$

where (Z_t) is a $l_t \times l_t$ matrix with terms all equal to $\frac{1}{l_t}$, i.e., Z is a $m \times m$ block diagonal matrix. The between-class and within-class scatter matrices in a high dimensional feature space R^f are defined as

$$S_B^\Phi = \sum_{i=1}^{c} l_i \boldsymbol{\mu}_i^\Phi (\boldsymbol{\mu}_i^\Phi)^T, \quad S_W^\Phi = \sum_{i=1}^{c} \sum_{j=1}^{l_i} \Phi(\mathbf{x}_{ij}) \Phi(\mathbf{x}_{ij})^T \tag{12}$$

where μ_i^Φ is the mean of class i in R^f, l_i is the number of samples belonging to class i. From the theory of reproducing kernels, any solution $\mathbf{w}^\Phi \in R^f$ must lie in the span of all training samples in R^f, i.e.,

$$\mathbf{w}^\Phi = \sum_{p=1}^{c} \sum_{q=1}^{l_p} \alpha_{pq} \Phi(\mathbf{x}_{pq}) \tag{13}$$

It follows that we can get the solution for (13) by solving:

$$\boldsymbol{\lambda} K K \boldsymbol{\alpha} = K Z K \boldsymbol{\alpha} \tag{14}$$

Consequently, we can write (8) as

$$\begin{aligned} W_{OPT}^\Phi &= \arg\max_{W^\Phi} \frac{|(W^\Phi)^T S_B^\Phi W^\Phi|}{|(W^\Phi)^T S_W^\Phi W^\Phi|} \\ &= \arg\max_{W^\Phi} \frac{|\boldsymbol{\alpha} K Z K \boldsymbol{\alpha}|}{|\boldsymbol{\alpha} K K \boldsymbol{\alpha}|} = [\mathbf{w}_1^\Phi \dots \mathbf{w}_m^\Phi] \end{aligned} \tag{15}$$

We can project $\Phi(\mathbf{x})$ to a lower dimensional space spanned by the eigenvectors $\mathbf{w}^\Phi$ in a way similar to Kernel PCA (See Section 2). Adopting the same technique in the Fisherface method (which avoids singularity problems in computing W_{OPT}^Φ) for face recognition [6], we call the extracted eigenvectors in (15) Kernel Fisherfaces.

4 Experiments

We tested both kernel methods against ICA, SVM, Eigenface, and Fisherface methods using the publicly available AT&T and Yale databases. The face images in these databases have several unique characteristics. While the images in the AT&T database contain facial contours and vary in pose as well as scale, the face images in the Yale database have been cropped and aligned. The face images in the AT&T database were taken under well controlled lighting conditions whereas the images in the Yale database were acquired under varying lighting conditions. We used the first database as a baseline study and then used the second one to evaluate face recognition methods under varying lighting conditions.

The minimum number of components in Eigenface, Kernel Eigenface, and ICA-based methods were empirically determined to achieve the lowest error rates (See Figures 2 and 3). For Fisherface and Kernel Fisherface methods, we projected all the samples onto a subspace spanned by the $c-1$ largest eigenvectors. We utilized both polynomial and Gaussian kernels in the kernel Eigenface, kernel Fisherface, and SVM-based methods. The types of kernel and corresponding parameters (e.g., polynomial degree) were also empirically determined to achieve the best results. Typically, second or third order polynomial kernels suffices to achieve good results with less computation than Gaussian kernels. We present more results of kernel methods with polynomial kernels.

All experiments were performed using the "leave-one-out" strategy: To classify an image of person, that image is removed from the training set of $(m-1)$ images and the projection matrix is computed. All the m images in the training and test sets were projected to a reduced space using the computed projection matrix $\mathbf{w}$ or $\mathbf{w}^\Phi$ and recognition was performed based on a nearest neighbor classifier except SVM-based methods. We adopted the "one-against-the-rest" scheme in training SVMs for experiments. In other words, we trained 400

SVMs for AT & T database and 165 SVMs for the Yale database.

4.1 Variation in Pose and Scale

The AT&T (formerly Olivetti) database contains 400 images of 40 subjects. To reduce the computational complexity, each face image was downsampled to 23×28 pixels. We represented each image by a raster scan vector of the intensity values, and then normalize them to be zero-mean vectors. The mean and standard deviation of Kurtosis of the face images are 2.08 and 0.41, respectively (the Kurtosis of a Gaussian distribution is 3). Figure 1 (top) shows images of two subjects. In contrast to images of the Yale database, the images include facial contours, and variations in pose as well as scale. However, the lighting conditions remain relatively constant.

Figure 1. Face images in the AT&T database (Top) and the Yale database (Bottom).

The experimental results are shown in Figure 2. Among all the methods, the Kernel Fisherface method with Gaussian kernel and second order polynomial kernel achieve the lowest error rate. Furthermore, the kernel methods perform better than their classical counterparts, respectively. The kernel methods also perform better than SVM-based and ICA-based methods. Though our experiments using ICA seem to contradict to the good empirical results reported in [4] [3] [2], a close look at the data sets reveals a significant difference in pose and scale variations of the face images in the AT&T database, whereas a subset of frontal view FERET face images with change of expression was used in [3] [2]. Furthermore, the comparative study on classification with respect to PCA in [4] (Table 1, pp. 819) and the errors made by two ICA algorithms in [2] (Figure 2.18, pp. 50) seem to suggest that ICA methods do not have clear advantage over other approaches in recognizing faces with pose and scale variations.

4.2 Variation in Lighting and Expression

The Yale database used in our experiments contains 165 closely cropped face images of 11 subjects that include variations in both facial expression and lighting (See Figure 1). For computational efficiency, each image was downsampled to 29×41 pixels, and then represented by a centered vector of normalized intensity values. The mean and standard deviation of Kurtosis of the face images are 2.68 and 1.49, respectively. Figure 1 shows 22 closely cropped images of two subjects which include internal facial structures such as the eyebrow, eyes, nose, mouth and chin, but do not contain facial contours.

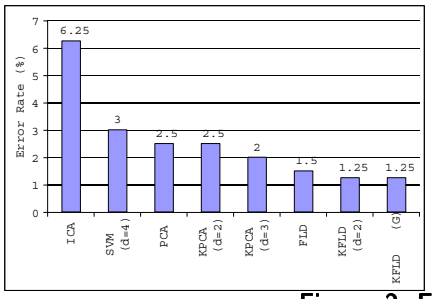
Figure 3 shows the experimental results. Both kernel methods perform better than standard methods using ICA and their classical counterparts, whereas SVM-based method performs better than Kernel Eigenface method. Notice that the improvements by the kernel methods are significant (more than 15% reduction in error rate). Notice also that kernel methods consistently perform better than classical methods for both databases. The performance achieved by the ICA method indicates that face representation using independent sources is not effective when the images are taken under varying lighting conditions.

Figure 4 shows the training samples of the Yale database projected onto the first two eigenvectors extracted by the Kernel Eigenface and Kernel Fisherface methods. The projected samples of different classes are smeared by the Kernel Eigenface method whereas the samples projected by the Kernel Fisherface are separated quite well. In fact, the samples belonging to the same class are projected to the same position by the largest two eigenvectors. This example provides an explanation to the good results achieved by the Kernel Fisherface method.

4.3 Discussion

Our experimental results show that Kernel Eigenface and Fisherface methods are able to extract nonlinear features and achieve lower error rates. One explanation for some of the superior performance of Kernel Fisherface method over SVM-based method may be attributed to the fact that Kernel Fisherface method uses all training samples to extract the most discriminant (nonlinear) features in the solution, not all (possibly infinite number of) the features of a subset set of samples, i.e., the support vectors, for recognition.

The performance of the proposed kernel methods may be improved by using other classifiers such as k-nearest neighbor and perceptrons. The performance of SVM-based method may be improved by adopting "one-against-one" strategy. downside is that a large number of classifiers need to be trained (See also [20] for a multiclass SVM method). Another potential improvement is to use the extracted nonlinear features and a linear Support Vector Machine (SVM) to construct a decision surface. Such a two-stage approach



Method	Reduced Space	Error Rate (%)
Eigenface	40	2.50 (10/400)
Fisherface	39	1.50 (6/400)
ICA	80	6.25 (25/400)
SVM, d=4	N/A	3.00 (12/400)
LLE, # neighbor=70	70	2.25 (9/400)
Isomap, ϵ=10	30	1.75 (7/400)
Kernel Eigenface, d=2	40	2.50 (10/400)
Kernel Eigenface, d=3	40	2.00 (8/400)
Kernel Fisherface (P)	39	1.25 (5/400)
Kernel Fisherface (G)	39	1.25 (5/400)

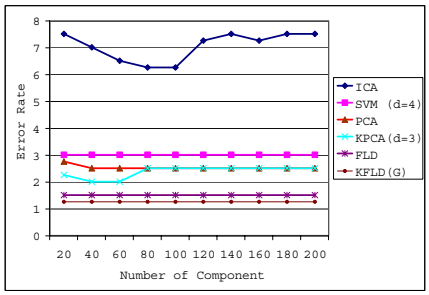
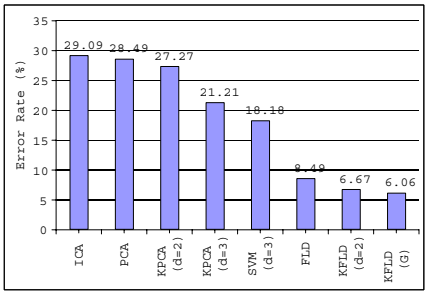


Figure 2. Experimental results on AT&T database.



Method	Reduced Space	Error Rate (%)
Eigenface	30	28.48 (47/165)
Fisherface	14	8.48 (14/165)
ICA	100	28.48 (47/165)
Isomap, ϵ=20	60	27.27 (45/165)
LLE, # neighbor = 10	30	26.06 (43/165)
SVM, d=3	N/A	18.18 (30/165)
Kernel Eigenface, d=2	80	27.27 (45/165)
Kernel Eigenface, d=3	60	24.24 (40/165)
Kernel Fisherface (P)	14	6.67 (11/165)
Kernel Fisherface (G)	14	6.06 (10/165)

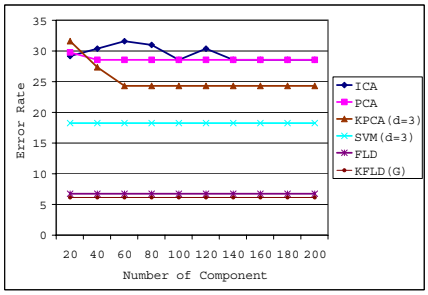
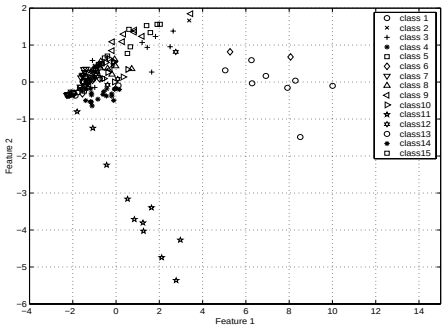
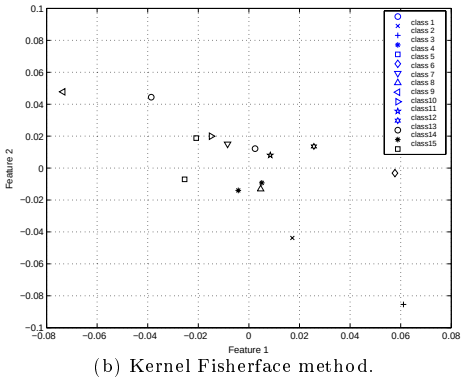


Figure 3. Experimental results on Yale database.



(a) Kernel Eigenface method.



(b) Kernel Fisherface method.

Figure 4. Samples projected by Kernel PCA and Kernel Fisher methods.

is, in spirit, similar to nonlinear SVMs in which the samples are first projected to a high dimensional feature space where a hyperplane with largest margin is constructed. In fact, one important factor of the recent success in SVM applications for visual recognition is due to the use of kernel methods. The superior results achieved by SVMs than the classical Eigenface method in the second experiment can be attributed to the rich feature representation and a decision surface with large margin. However, our experimental results also suggest that the feature representation of SVMs does not have as much discriminative power as Fisherface or Kernel Fisherface method.

It is difficult, if not impossible, to estimate a true distribution of face images (or a subset of face images). Thus it is difficult to justify whether the density function of face images is Gaussian or not. Nevertheless, kurtosis is often used as an index of non-Gaussianity of samples. The computed kurtoses suggest that the face images are not non-Gaussian (in fact, sub-Gaussian). In general the more non-Gaussian the data, the better ICA can be estimated. This may explain why ICA-based methods do not perform well in these experiments (We used the fixed point algorithm [12] to extract independent components.). One potential improvement of ICA-based method is to select a subset of independent components by class discriminability as suggested in [2].

In terms of execution time, our experiments (with Matlab implementations) show that the ratio of computation loads required by these methods is, on the

average, ICA: SVM: KFLD: KPCA: FLD: PCA = 8.7: 5.1: 3.3: 3.2: 1.3: 1.0 (averaged over all the experiments).

5 Conclusion and Future Work

The representations in the classical Eigenface and Fisherface approaches are based on second order statistics of the image set, i.e., covariance matrix, and do not use high order statistical dependencies such as the relationships among three or more pixels. For face recognition, much of the important information may be contained in the high order statistical relationships among the pixels. Using the kernel tricks that are often used in SVMs, we extend the classical methods to kernel space where we can extract nonlinear features among three or more pixels. We investigated Kernel Eigenface and Kernel Fisherface methods, and demonstrated that they provide a more effective representation for face recognition. Compared to other techniques for nonlinear feature extraction, kernel methods have the advantages that they do not require nonlinear optimization, but only the solution of an eigenvalue problem. Experimental results on two benchmark databases show that Kernel Eigenface and Kernel Fisherface methods achieve lower error rates than the ICA, Eigenface and Fisherface approaches in face recognition. The performance achieved by the ICA method also indicates that face representation using independent basis images is not effective when the images contain pose, scale or lighting variations. Our future work will focus on analyzing face recognition methods using other kernel machines in high dimensional space [18] , nonlinear subspace algorithms [16] [29] [24] [9], evolutionary pursuit [13], and generative methods [28] using FERET [19], AR [14] and CMU PIE [26] databases.

References

[1] Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.

[2] M. S. Bartlett. *Face Image Analysis by Unsupervised Learning and Redundancy Reduction*. PhD thesis, University of California at San Diego, 1998.

[3] M. S. Bartlett, H. M. Lades, and T. J. Sejnowski. Independent component representations for face recognition. In *Proceedings of SPIE Symposium on Electronic Imaging: Science and Technology; Human Vision and Electronic Imaging III*, vol. 3299, pp. 528–539, 1998.

[4] M. S. Bartlett and T. J. Sejnowski. Viewpoint invariant face recognition using independent component analysis and attractor networks. In *Advances in Neural Information Processing Systems 9*, page 817, 1997.

[5] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.

[6] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[7] A. J. Bell and T. J. Sejnowski. An information - maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[8] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[9] C. M. Bishop and J. M. Winn. Nonlinear bayesian image modeling. In *Proceedings of the Sixth European Conference on Computer Vision*, vol. 1, pp. 3–17, 2000.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, 2001.

[11] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, 2001.

[12] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[13] C. Liu and H. Wechsler. Evolutionary pursuit and its application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):570–582, 2000.

[14] A. M. Martinez and A. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

[15] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K.-R. Müller. Invariant feature extraction and classification in kernel spaces. In S. Solla, T. Leen, , K.-R, and Müller, editors, *Advances in Neural Information Processing Systems 12*, pp. 526–532. MIT Press, 2000.

[16] B. Moghaddam. Principal manifolds and bayesian subspaces for visual recognition. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pp. 1131–1136, 1999.

[17] B. Moghaddam and A. Pentland. Probabilistic visual learning for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[18] P. J. Phillips. Support vector machines applied to face recognition. In *Advances in Neural Information Processing Systems 11*, pp. 803–809, 1998.

[19] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1034, 2000.

[20] J. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pp. 547–553. MIT Press, 2000.

[21] A. N. Rajagopalan, P. Burlina, and R. Chellappa. Higher order statistical learning for vehicle detection in images. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1204–1209, 1999.

[22] A. N. Rajagopalan, K. S. Kumar, J. Karlekar, R. Manivasakan, and M. M. Patil. Finding faces in photographs. In *Proceedings of the Sixth IEEE International Conference on Computer Vision*, pp. 640–645, 1998.

[23] V. Roth and V. Steinhage. Nonlinear discriminant analysis using kernel functions. In S. Solla, T. Leen, , K.-R, and Müller, editors, *Advances in Neural Information Processing Systems 12*, pp. 568–574. MIT Press, 2000.

[24] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2000.

[25] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[26] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database of human faces. Technical Report CMU-RI-TR-01-02, Carnegie Mellon University, 2001.

[27] D. L. Swets and J. Weng. Hierarchical discriminant analysis for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):386–401, 1999.

[28] Y. W. Teh and G. E. Hinton. Rate-coded restricted Boltzmann machines for face recognition. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pp. 908–914. MIT Press, 2001.

[29] J. B. Tenebaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2000.

[30] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[31] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pp. 668–674. MIT Press, 2001.

[32] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, pp. 336–341, 1998.