Face Recognition with Kernel Correlation Filters on a Large Scale Database

Jingu Heo, Marios Savvides, Ramzi Abiantun, Chunyan Xie and B.V.K. Vijayakumar
ECE Department, Carnegie Mellon University, USA
{jheo@, msavvid@ri, raa@, chunyanx@, kumar@ece}.cmu.edu

ABSTRACT

Recently, *Direct Linear Discriminant Analysis (D-LDA)* and *Gram-Schmidt LDA* methods have been proposed for face recognition. By also utilizing some of the null-space of the *within-class scatter matrix*, they exhibit better performance compared to *Fisherfaces and Eigenfaces*. However, these linear subspace methods may not discriminate faces well due to large nonlinear distortions in the face images. Redundant class dependence feature analysis (CFA) method exhibits superior performance compared to other methods by representing nonlinear features well. We show that with a proper choice of kernel parameters used with the proposed Kernel Correlation Filters within the CFA framework, the overall face recognition performance is significantly improved. We present results of this proposed approach on a large scale database from the Face Recognition Grand Challenge (FRGC) which contains over 36,000 images.

1. INTRODUCTION

Machine recognition of human faces from still-images and video frames is an active research area due to the increasing demand for authentication in commercial and law enforcement applications. Despite the research advancement over the years, face recognition is still a highly challenging task in practice due to large nonlinear distortions caused by natural facial appearance distortions such as expressions, pose and ambient illumination variations. Two well-known popular algorithms for face recognition are *Eigenfaces* [1] and *Fisherfaces* [2]. The Eigenfaces method generates features that capture the holistic nature of faces through the Principal Component Analysis (PCA), which determines a lower-dimensional subspace that offers the minimum mean squared error approximation to the original high-dimensional face data. Instead of seeking a subspace that is efficient for *representation*, the Linear Discriminant Analysis (LDA) method seeks projection directions that are more optimal for *discrimination*. In practice, Fisherfaces first performs PCA to overcome the singularities in the within-class scatter matrix (S_w) by reducing the dimensionality of the data and then applies traditional LDA in this lower-dimensional subspace.

Recently [3], it has been suggested that the null space of the S_w matrix is important for discrimination. The claim is that applying PCA in Fisherfaces may discard discriminative information since the null space of S_w contains the most discriminative information. Fueled by this insight, Direct LDA (DLDA) [3] and Gram-Schmidt LDA (GSLDA) [4] methods have been proposed by utilizing part of the null-space of the S_w. However, these linear subspace methods may not discriminate faces well due to large nonlinear distortions in the faces. In such cases, the proposed kernel correlation filter approach may be attractive because of its ability to tolerate some level of distortions [5]. One of the recent advances in advanced correlation filters is redundant class dependence feature analysis (CFA) [6] which proposes a novel feature extraction method using advanced correlation filter methods to provide superior performance. We will show that with a proper choice of nonlinear kernel parameters with our proposed kernel correlation filter, the performance can be significantly improved over previous work. Our experimental evaluation includes results from CFA, GSLDA, Fisherfaces, Kernel LDA (KDA), Eigenfaces and Kernel PCA on a large scale database from the face recognition grand challenge (FRGC) dataset collected by the University of Notre Dame [12].

2. BACKGROUND

The PCA finds the minimum mean squared error linear transformation that maps from the original N -dimensional data space into an M-dimensional feature space (where $M << N$) to achieve dimensionality reduction by using the eigenvectors corresponding to the largest eigenvalues. These resulting basis vectors are obtained by finding the optimal $\mathbf{W}$ vectors that maximizes the total variance of the projected data

$$\mathbf{W}_{\mathbf{opt}} = \arg\max |W^T S_T W| = [w_1 \; w_2 \; \ldots \; w_m] \qquad (1)$$

where S_T denotes the total scatter matrix. Figure 1 shows examples of Eigenfaces generated from the generic training images of FRGC data.

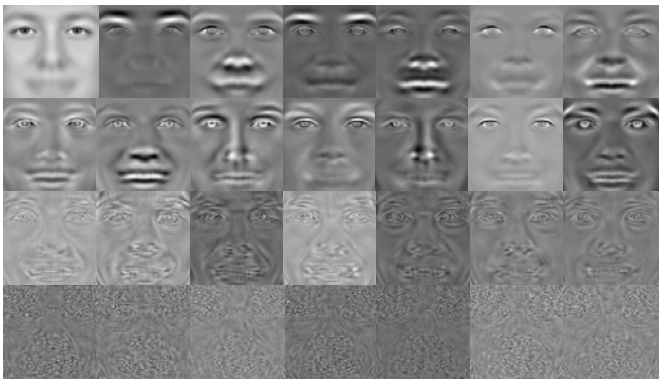

Figure 1: Eigenfaces from the FRGC data sorted by the largest eigenvalues; 1st and 2nd row images show first 14 eigenvectors, 3rd row images show 201 ~ 207 eigenvectors and 4th row images show 501~507 eigenvectors in descending order.

LDA is another commonly used method which seeks to find a set of discriminant projection vectors that maximize the ratio of the between-class scatter and the within-class scatter in the projected space. The optimal basis vectors can be denoted as

$$W_{opt} = \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|} = [w_1 \; w_2 \; ... \; w_m] \quad (2)$$

where S_B and S_W indicate the between-class scatter matrix and the within-class scatter matrix, respectively. The solution can be obtained by solving the following generalized eigenvalue problem,

$$S_B w_i = \lambda_i S_W w_i, \quad i = 1,2,...m \quad (3)$$

If S_w is invertible, the above generalized eigenvalue problem simplifies to the following regular eigenvalue problem.

$$S_W^{-1} S_B w_i = \lambda_i w_i \quad (4)$$

Due to the fact that the number of training images is usually significantly smaller than the number of pixels, the within-class scatter matrix S_w is singular causing problems for LDA. Fisherfaces overcomes this singularity problem of LDA by first performing PCA to reduce the dimensionality (to overcome this singular-matrix issue) and then applies LDA in this lower-dimensional subspace. The projection vectors from Fisherfaces can be found by optimizing the following figure of merit

$$W_{pca} = \arg\max |W^T S_T W| = [w_1 \; w_2 \; ... \; w_{N-c}]$$
$$\mathbf{W_{opt}} = \arg\max_{W} \frac{|W^T W_{pca}{}^T S_B W_{pca} W|}{|W^T W_{pca}{}^T S_W W_{pca} W|} = [w_1 \; w_2 \; ... \; w_{c-1}] \quad (5)$$

where c is defined as the total number of classes.

On the other hand, the Direct LDA (DLDA) derives eigenvectors after simultaneous diagonalization [8]. Unlike previous approaches, the DDLA simultaneously diagonalizes S_B first and then diagonalizes S_W which can be expressed as follows.

$$W^T S_B W = I, \quad W^T S_W W = \Lambda \quad (6)$$

The eigenvectors with very small (close to zero) eigenvalues in the S_B can be discarded since they contain no discriminative power, while simultaneously keeping the eigenvectors with small eigenvalues in the S_W, especially the null-space. Another method is the GSLDA approach which avoids computing the inverse of the within-class scatter matrix or performing the diagonalization needed in LDA. The GSLDA approach calculates the orthogonal basis vectors in

$$\overline{S_T(0)} \cap S_W(0) \quad (7)$$

where $S_T(0)$.and $S_w(0)$ indicate the corresponding null spaces of each scatter matrix and $\overline{S_T(0)}$ indicates the orthogonal complement spaces of $S_T(0)$. The GSLDA method has been reported to offer better performance over Fisherfaces and other LDA methods [4], which in turn outperform PCA based methods [2]. Figure 2 shows examples of the LDA basis vectors generated from the generic training images of the FRGC data.

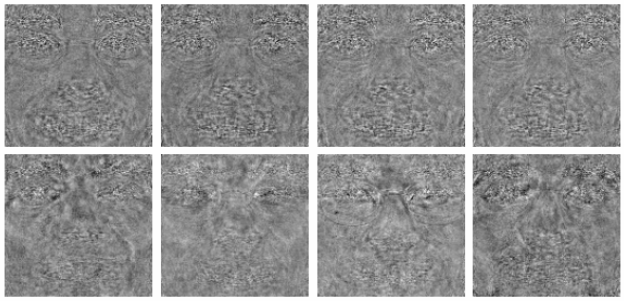

Figure 2: The LDA basis vectors; 1st row images are examples of the Fisherfaces, and 2nd row images are examples of the GSLDA eigenvectors.

Correlation filter approach represents the data in the frequency domain using derived closed form solution to specific optimization criteria. One of the most popular correlation filters is the minimum average correlation energy (MACE) [7] filter. This is designed to minimize the average correlation plane energy resulting from the training images, while constraining the correlation peak value at the origin to pre-specified values. Correlation outputs from MACE filters typically exhibit sharp peaks, making the peak detection and location relatively easy and robust. The closed form expression for the vectorized MACE filter **h** is

$$\mathbf{h} = \mathbf{D^{-1}X(X^{+}D^{-1}X)^{-1}u} \quad (8)$$

where **X** is a d^2xN complex valued matrix (where N is the number of training images and d^2 is the number of pixels in each image) and its ith column contains the lexicographically re-ordered version of the 2-D Fourier transform of the ith training image**.** **D** is a d^2xd^2 diagonal matrix containing the average power spectrum of the training images along its diagonal and **u** is a column vector containing N pre-specified correlation values at the origin. Optimally trading off between noise tolerance and peak sharpness produces the optimal trade-off filters (OTF). OTF filter vector is given by

$$\mathbf{h} = \mathbf{T^{-1}X(X^{+}T^{-1}X)^{-1}u} \quad (9)$$

where $\mathbf{T} = \left(\alpha \mathbf{D} + \sqrt{1-\alpha^2}\mathbf{C}\right)$, $0 \le \alpha \le 1$, and $\mathbf{C}$ is a $d^2 x d^2$ diagonal matrix whose diagonal elements $\mathbf{C}(k,k)$ represent the noise power spectral density at frequency k. Correlation filters are well suited for biometric verification application and have been shown to exhibit robustness to illumination variations and other distortions [5].

The LDA based methods offer the potential to outperform Eigenfaces. However, LDA based vectors may not have generalization power to the unseen classes. This problem also occurs when we apply the correlation filters since the typical design of correlation filters is based on only using the gallery images. The class-dependence feature analysis (CFA) is proposed to generalize the correlation filters and allow the use of the generic dataset to produce a novel feature extraction method as explained in the next section.

3. CLASS-DEPENDENCE FEATURE ANALYSIS (CFA)

In CFA approach, one filter (e.g., MACE filter) is designed for each class in the generic training set. Then a test image y is characterized by the inner products of that test image with the n MACE filters, i.e.,

$$\mathbf{x} = \mathbf{H}^T y = [\mathbf{h}_{\text{mace-1}}\ \mathbf{h}_{\text{mace-2}} \cdots\ \mathbf{h}_{\text{mace-n}}]^T y \qquad (10)$$

where $h_{\text{mace-n}}$ is a filter is trained to give a small correlation output (close to 0) for all classes except for class-n as shown in Figure 3. For example, the number of filters generated by the FRGC generic training set is 222 since it contains 222 classes (or subjects). Then each input image y is projected onto those basis vectors to yield a 222 dimensional feature vector x. The similarity of the probe image to the gallery image is done in this 222 dimensional feature space.

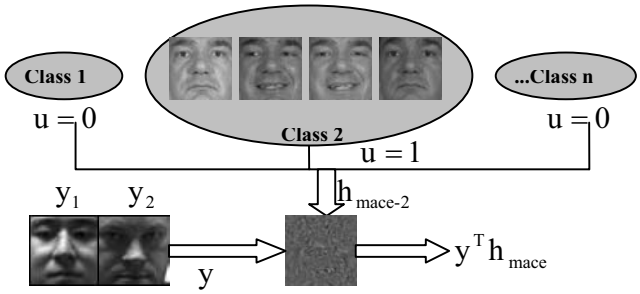

Figure 3: The CFA algorithm; the filter response of y_1 and $h_{\text{mace-2}}$ can be distinctive to that of y_2 and $h_{\text{mace-2}}$

Due to the nonlinear distortions in human faces, the linear subspace methods have not performed well in real face recognition applications. As a result, the PCA and LDA algorithms are extended to represent nonlinear features efficiently by mapping onto a higher dimensional space. Since nonlinear mappings increase the dimensionality rapidly, kernel trick methods are used for computational efficiency as they enable us to obtain the necessary inner products in the higher-dimensional feature space without computing the higher-dimensional feature mapping. Examples are Kernel Eigenfaces ,Kernel Fisherfaces [13] and Support Vector Machines (SVM)[9][10][11]. The mapping function can be denoted as follows.

$$\Phi : R^N \rightarrow F \qquad (11)$$

Kernel functions defined by $K(x,y) = <\Phi(x), \Phi(y)>$ can be used without having to form the mapping as long as kernels form an inner product and satisfy Mercer' theorem. Examples of kernel functions are: Polynomial kernel ($K(a,b) = (<a,b>+1)^p$), Radial Basis Function kernel ($K(a,b) = \exp(-(a-b)^2/2\sigma^2)$), and Neural Net Kernel ($K(a,b) = \tanh(k<a,b>-\delta)$ are known.

4. KERNEL CORRELATION FILTERS

The Kernel Correlation Filters can be extended from the linear advanced correlation filters using the kernel trick. The correlation output of a filter **h** and an input y can be expressed as

$$\begin{aligned} y \cdot \mathbf{h} &= y \cdot \mathbf{D}^{-1}\mathbf{X}(\mathbf{X}^+\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{u} \\ &= (\mathbf{D}^{-0.5}y)\cdot(\mathbf{D}^{-0.5}\mathbf{X})(\mathbf{D}^{-0.5}\mathbf{X}\cdot\mathbf{D}^{-0.5}\mathbf{X})^{-1}\mathbf{u} \\ &= (y'\cdot\mathbf{X}')(\mathbf{X}'\cdot\mathbf{X}')^{-1}\mathbf{u} \end{aligned} \qquad (12)$$

where $\mathbf{X}' = \mathbf{D}^{-0.5}\mathbf{X}$ indicates pre-whitened version of $\mathbf{X}$. From now on, we assume the $\mathbf{X}$ is already pre-whitened. Now we can apply the kernel trick to yield the Kernel Correlation Filter:

$$\begin{aligned} \Phi(y)\cdot\Phi(\mathbf{h}) &= (\Phi(\mathbf{y})\cdot\Phi(\mathbf{X}))(\Phi(\mathbf{X})\cdot\Phi(\mathbf{X}))^{-1}\mathbf{u} \\ &= K(y,x_i)K(x_i,x_j)^{-1}\mathbf{u} \end{aligned} \qquad (13)$$

We can use these filters on the same CFA framework to obtain the same 222 dimensional feature vector representation which we refer to as the Kernel CFA (KCFA) method. Figure 4 shows the comparative experimental results using Eigenfces (PCA), GSLDA, CFA, and KCFA of the experiment 4 of the FRGC. The performance of Eigenfaces is the benchmark provided by NIST [12].

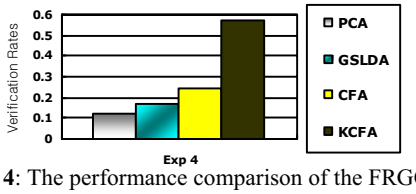

Figure 4: The performance comparison of the FRGC experiment 4 at 0.1 % FAR. The Kernel CFA shows the best results over all linear methods.

The similarity or distance measure between gallery image and probe image is important. Commonly used distance measures are L1-norm, L2-norm, *Mahalanobis distance* [6] and normalized cosine distance (given below) which exhibits the best results on the CFA and KCFA.

$$\mathbf{d(x,y)} = -(\mathbf{x}\cdot\mathbf{y})/(\|\mathbf{x}\|\|\mathbf{y}\|) \qquad (14)$$

where **d** denotes the similarity (or distance) between **x** and **y**. The PCA and GSLDA use the L2-norm while the CFA and KCFA use the distance in eq. 14, and the resulting performance is shown in Figure 4.

4. DISTANCE MEASURE IN SVM SPACE

A direct use of the SVM as a classifier may produce worse performance under those distortions since only small number of training images are allowed to build the SVM. Also due to the large dimensionality of the data, we want to apply SVMs in a reduced dimensional feature space. Instead of using the SVM as a classifier directly, we use the projection coefficients in the SVM space. Figure 5 shows an example of the decision boundary and distance measure in the SVM space.

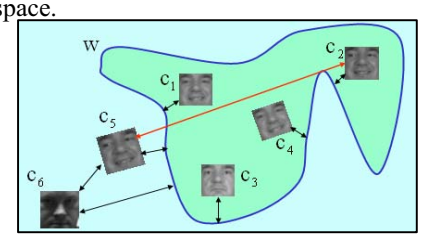

Figure 5: The decision boundary of a class and distance measure in the SVM space; a direct use of the SVM may falsely indicate that image C_5 as not the same person with those images inside the decision boundary.

The *L2-norm* distance without the SVM decision boundary **w** may be large between the same classes causing poor performance. In this case, the *L2-norm* distance of C_2 and C_5 is greater than that of C_6 and C_5 causing a misclassification. However, if we project C_5 on to **w**, the projection coefficients among the same classes will be small and we can change the threshold distance depending on FAR and FRR. Thus this approach may lead to flexibility of varying thresholds and better performance in classification. We design 466 SVMs (in a one-against all framework) using the gallery set of the FRGC data. The probe images are then projected on the class-specific SVMs which will provide a similarity score. As shown in Figure 6, the new distance measure in the SVM space produces better results than using normalized cosine distance. We also compared the different kernel approaches such as KPCA and KDA with different distance measure showing the SVM based KCFA methods have superior to other kernel approaches as shown in Figure 7.

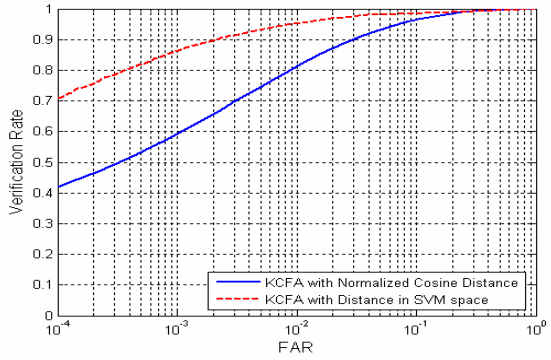

Figure 6: VR vs FAR for FRGC experiment 4 using KCFA with different distance measure.

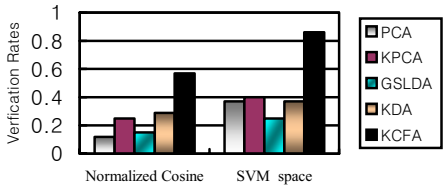

Figure 7: VR vs FAR for FRGC experiment 4 for different methods using normalized cosine distances and SVM space.

5. CONCLUSIONS

Due to nonlinear distortions and poor quality face images, linear approaches such as PCA, LDA, and CFA may not be suitable to represent or discriminate facial features efficiently. By using the kernel trick, the proposed Kernel Correlation Filter within the CFA framework exhibits better performance over all linear and other kernel approaches. We show that by further incorporating the SVM in this reduced dimensional feature space, an additional significant performance gain is achieved.

6. REFERENCES

[1] M. Turk and A. Pentland, "Eigenfaces for Recognition," Journal of *Cognitive Neuroscience*, Vol. 3, pp.72-86, 1991.

[2] P.Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. PAMI*, Vol.19. No.7, pp.711- 720, 1997.

[3] L.F. Chen, H.Y.M. Liao, M.T. Ko, J.C. Lin, and G.J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, Vol. 33, pp. 1713-1726, 2000

[4] W. Zheng, C. Zou, and L. Zhao, "Real-Time Face Recognition Using Gram-Schmidt Orthgonalization for LDA," *IEEE Conf. Pattern Recognition (ICPR)* , pp.403-406, 2004

[5] M. Savvides, B.V.K. Vijaya Kumar and P. Khosla, "Face verification using correlation filters," *Proc. of Third IEEE Automatic Identification Advanced Technologies*, Tarrytown, NY, pp.56-61, 2002.

[6] C. Xie, M. Savvides, and B.V.K. Vijaya Kumar, "Redundant Class-Dependence Feature Analysis Based on Correlation Filters Using FRGC2.0Data",*Proc.of IEEE Computer Vision and Pattern Recognition (CVPR)*-Workshops, Vol 3, San Diego, June 2005

[7] A. Mahalanobis, B.V.K. Vijaya Kumar, and D. Casasent, "Minimum average correlation energy filters," *Appl. Opt*. 26, pp. 3633-3630, 1987.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition* (2nd Edition), *New York: Academic Press*, 1990

[9] V. N. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag, 1995.

[10] B. Scholkopf, *Support Vector Learning*, Munich, Germany: Oldenbourg-Verlag, 1997.

[11] P. J. Phillips, "Support vector machines applied to face recognition," *Advances in Neural Information Processing Systems 11*, M. S. Kearns, S. A. Solla, and D. A. Cohn, eds., 1998.

[12] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W.Worek, "Overview of the Face Recognition Grand Challenge," *IEEE Conf. Computer Vision and Pattern Recognition(CVPR)*, 2005

[13] M.H Yang, "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition using Kernel Methods," *IEEE Conf. Automatic Face and Gesture Recognition*, pp. 215-220, 2002