Face Recognition using Canonical Correlation Analysis

Amit C. Kale  
Department of Electrical Engineering  
Indian Institute of Technology, Madras  
ee04s043@ee.iitm.ac.in

R. Aravind  
Department of Electrical Engineering  
Indian Institute of Technology, Madras  
aravind@tenet.res.in

Abstract

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are well known techniques for face recognition. Both PCA and LDA by themselves have good recognition rates. We propose Canonical Correlation Analysis (CCA) for combining two feature extractors to improve the performance of the system, by obtaining the advantages of both. CCA finds the transformation for each extractor dataset and maximizes the correlation between them. The feature extractors used in our experiments are combinations of PCA, LDA and Discrete Cosine Transform (DCT)-based PCA. We conduct tests on Yale Face Database, AR Face Database and FERET Database, and we show that the combination of two method works better than the individual methods.

1. Introduction

Face recognition is a widely studied research topic in the field of Computer Vision. It has received a lot of attention over past few years, due to availability of good computational facilities and requirement for robust biometric recognition systems for law enforcement and commercial applications. There are various face recognition algorithms which have been implemented, with main issues being pose and illumination variations [1].

A face image can be considered as a vector in a high dimensional space. This high dimensional vector has large variations between two images of the same person, and therefore it is not directly suited for the face recognition. Principal Component Analysis (PCA) [2] and Linear Discriminant Analysis (LDA)[4] are commonly used dimensionality reduction techniques, which give good recognition rates. Although PCA is a good low-dimensional representation for face images, it is not able to discriminate between variations due to illumination changes [6]. LDA solves the illumination change problem to some extent by finding the transformation such that it maximizes the inter-class separation and minimizes the intra-class variations [6]. A dimensionality reduction technique which we propose is Discrete Cosine Transform (DCT) based PCA, where the DCT of the whole image is computed and many of the DCT coefficients are discarded, thus reducing the number of dimensions. PCA of the remaining coefficients is performed to further reduce the dimensionality.

We employ by using Canonical Correlation Analysis (CCA) [9] to combine the vectors obtained by the dimensionality reduction techniques listed above. CCA finds a transformation for two feature data sets, such that the data covariance of the transformed data sets is maximized. The resulting vectors can then be combined either by concatenating or by adding. In our experiments, using CCA we combine PCA and LDA, PCA and DCT-based PCA, and LDA and DCT-based PCA.

We begin with a brief review of PCA, LDA and DCT-based PCA. Section 3 discusses CCA-based Feature Fusion Methods. Section 4 contains experimental results on various face databases; an improvement in recognition rates of upto 14% is observed with PCA and LDA, and LDA and DCT-based PCA combinations over the individual methods. Finally, conclusions are made in Section 5.

2. Review of PCA, LDA and DCT

In this section we review three different feature extractors which are used in CCA feature fusion, a brief description on each of these techniques is given. The main aim is to reduce the dimension of face vector without losing important information useful for recognition.

2.1. Principal Component Analysis

PCA uses the idea of representing a face vector as a weighted sum of basis vectors. It was introduced in [2], and was applied to face recognition [3]. As mentioned earlier we consider a face image as a point in a high dimensional space. The number of dimensions is equal to number of pixels in the face image. The set of training images form a face cluster in the high dimensional space. PCA finds out a subspace for the cluster such that, it has maximum energy stored. The directions corresponding to maximum data variations are the eigenvectors of the covariance matrix for the face cluster.

We consider there are \( L \) training images, with \( x_i \) being the vector form of the \( i \)th image in the training set. Let
$X = \{x_1, x_2, \ldots, x_L\}$ be the set of training images. The images are centered to zero by subtracting the grand mean $m = \frac{1}{L} \sum_{i=1}^{L} x_i$, to form a set $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_L\}$ with covariance matrix

$$C_{xx} = \frac{1}{L} \sum_{i=1}^{L} \tilde{x}_i \tilde{x}_i^T = \frac{1}{L} \tilde{X} \tilde{X}^T. \quad (1)$$

The eigenvectors and eigenvalues of $C_{xx}$ are given as $\{\tilde{w}_i\}$ and $\{\lambda_i\}$ respectively, and therefore $C_{xx} \tilde{w}_i = \lambda_i \tilde{w}_i$. The rank of $\{\lambda_i\}$ is $\leq L - 1$, so only a small number of eigenvalues are considered.

Let $\{\lambda_1, \lambda_2, \ldots, \lambda_{n_1}\}$ be the first $n_1$ eigenvalues sorted in decreasing order, where $n_1 \leq L - 1$ and let $W = [\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_{n_1}]$. The image vector $\tilde{x}_i$ projected on the eigenvectors $W$ leads to a low-dimensional representation of dimensionality $n_1$ and is given as

$$z_i = W^T \tilde{x}_i. \quad (2)$$

### 2.2. Fisher’s Linear Discriminant Analysis

Fisher’s Linear Discriminant Analysis or LDA [4] is a class specific method which discriminates between classes by finding projections such that, it maximizes the separation between two classes, under the constraint that the within-class variations are minimized.

In face recognition, each subject is treated as a class. When $c$ subjects or classes are present in the set of training images, we label class $i$ as $\chi_i$ and denote by $L_i$ the number of training images in $\chi_i$. The mean vector for class $\chi_i$ is $\mu_i$, and the mean vector over all training images is $\mu$, where $L = \sum_{i=1}^{c} L_i$. The between-class scatter matrix is given as

$$S_B = \sum_{i=1}^{c} L_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (3)$$

while the within-class scatter matrix is defined as

$$S_W = \sum_{i=1}^{c} \sum_{x_j \in \chi_i} (x_j - \mu_i)(x_j - \mu_i)^T. \quad (4)$$

If the matrix $S_W$ is non-singular, then the optimal projection $W_{opt}$ [4] is computed such that

$$W_{opt} = \arg \max_{W} \frac{W^T S_B W}{W^T S_W W} = [w_1, w_2, \ldots, w_m] \quad (5)$$

where $\{w_i| i = 1, 2, \ldots, m\}$ is a set of generalized eigenvectors of the two scatter matrices, $S_B$ and $S_W$. The upper bound on $m$ is $c - 1$, and it is the maximum rank of $S_B$. The problem of evaluating $W_{opt}$ is non-trivial, a detailed explanation is given in [4].

### 2.3. Discrete Cosine Transform-based PCA

We propose a DCT-based PCA as a feature extractor technique for face recognition. The DCT of an $N \times N$ image $\{f(x, y)\}$, as used in image compression [7], is defined as

$$F(u, v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) \cos \left( \frac{(2x + 1)u\pi}{2N} \right) \cos \left( \frac{(2y + 1)v\pi}{2N} \right). \quad (6)$$

Due to its energy compaction property, most of the energy is contained in a small number of low frequency components. DCT acts as a simple dimensionality reduction technique, thus allowing us to discard redundant information based on frequency, while not compromising on vital information which characterizes a face.

In DCT-based PCA we compute the DCT of the whole face image. The DCT image is scanned in a zig zag manner and we discard the low amplitude high frequency coefficients which lie at the tail. A PCA is performed on the set of DCT vectors corresponding to the training images. We project the DCT vectors on the Principal Components to obtain a low-dimensional representation of face vectors. For simplicity this section onwards we will refer DCT based PCA as DCT.

### 3. Face Recognition based on Canonical Correlation

In [14] CCA was used for the segmentation of functional Magnetic Resonance Images; here we apply CCA for the purpose of face recognition. Correlation analysis is useful to find a linear relationship between two sets of variables, and CCA creates new variables for the each set such that the correlation between these variables is maximized and independent of affine transformation.

In CCA, given two zero-mean random vectors $x$ and $y$, we wish to find vectors $\alpha$ and $\beta$ that maximizes $\text{corr}(\alpha^T x, \beta^T y)$ defined in Equation(8), i.e. the correlation between the projections. Let $n_1$ and $n_2$ be the length of the random vectors $x$ and $y$ respectively. The overall covariance matrix $C$ of $x$ and $y$ is given as

$$C = E \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} x^T \\ y^T \end{pmatrix} = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} \quad (7)$$

where $C_{xy}$ is the cross-covariance matrix of $x$ and $y$, while $C_{xx}$ and $C_{yy}$ are covariance matrices of $x$ and $y$ respectively; we assume $C_{xx}$ and $C_{yy}$ are full rank. The correlation coefficient $\rho$ is given as

$$\rho = \frac{E[\alpha^T x \beta^T y]}{\sqrt{E[\alpha^T \alpha] \cdot E[\beta^T \beta]}} = \frac{E[\alpha^T x y^T \beta]}{\sqrt{E[\alpha^T x x^T \beta] \cdot E[\beta^T y y^T \beta]}} = \frac{\alpha^T C_{xy} \beta}{\sqrt{\alpha^T C_{xx} \alpha \cdot \beta^T C_{yy} \beta}}. \quad (8)$$
The sign of $\rho$ changes with sign of $a$ or $b$, so it is sufficient to consider only positive values of $\rho$.

We find $a$ and $b$ that maximize $\rho$ by taking the partial derivative of $\rho$ with respect to $a$ and $b$ [9]. Setting the partial derivatives to zero gives us the following set of equations.

\begin{equation}
C_{xy}b = \left( \frac{a^T C_{xy} b}{a^T C_{xx} a} \right) C_{xx} a
\end{equation}

\begin{equation}
C_{xy}a = \left( \frac{a^T C_{xy} b}{b^T C_{yy} b} \right) C_{yy} b.
\end{equation}

Equations(9) and (10) can be further simplified to obtain

\[
C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{xy}^T a = \left( \frac{a^T C_{xy} b}{\sqrt{a^T C_{xx} a \cdot b^T C_{yy} b}} \right)^2 a
\]

\begin{equation}
\rho^2 a.
\end{equation}

\[
C_{yy}^{-1} C_{xy}^T C_{xx}^{-1} C_{xy} b = \left( \frac{a^T C_{xy} b}{\sqrt{a^T C_{xx} a \cdot b^T C_{yy} b}} \right)^2 b
\]

\begin{equation}
\rho^2 b.
\end{equation}

Thus $a$ and $b$ are the eigenvectors of $C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{xy}^T$ and $C_{yy}^{-1} C_{xy}^T C_{xx}^{-1} C_{xy}$ respectively, with squared correlation coefficients $\rho^2$ being the eigenvalues. Solving Equations(11) and (12) we get $n$ solutions for $a$, $b$ and $\rho$, where $n = \min(n_1,n_2)$.

The linear combinations $x_i = a_i^T X$ and $y_i = b_i^T Y$ are called as canonical variates and correlation coefficients $\rho$ are called as canonical correlations [9]. The canonical variates corresponding to different roots of Equations(9) and Equations(10) are uncorrelated, i.e

\[
a^T C_{xx} a_j = 0
\]

\[
b^T C_{yy} b_j = 0
\]

\[
a^T C_{xy} b_j = 0
\]

for $i \neq j$ (13)

3.1. CCA based Feature Fusion for Face Recognition

In [8] two methods were proposed to combine the outputs of two arbitrary feature extractors (for use in character recognition) i.e Feature Fusion [8]. Here we apply the same methods to combine two features obtained from any of the extractors discussed in Section 2. The goal is to get the maximum information out of the two feature extractors. We apply CCA to the feature extractors and fuse the two features as discussed below.

Let the two feature extractors be trained by $L$ training images. Let $X = [x_1, x_2, \ldots, x_L]$ and $Y = [y_1, y_2, \ldots, y_L]$ be the corresponding outputs of the two extractors, and $n_1$ and $n_2$ be the dimensions of the two outputs, where $n_1, n_2 \leq L - 1$.

Let $\tilde{x}_i = \frac{1}{L} \sum_{j=1}^L x_j$ and $\tilde{y}_i = \frac{1}{L} \sum_{j=1}^L y_j$.

The covariance matrices for $X$ and $Y$ are given as $C_{xx}$ and $C_{yy}$ respectively, and are calculated using Equation(1); $C_{xx}$ and $C_{yy}$ are full rank, and $C_{xy}$ is the between-set covariance matrix calculated as

\[
C_{xy} = \frac{1}{L} \sum_{i=1}^L \tilde{x}_i \tilde{y}_i^T = \frac{1}{L} \tilde{X} \tilde{Y}^T.
\]

Canonical correlations and basis vectors are computed for these covariance matrices by using the Equations(11)(12). Let $A$ and $B$ be the set of canonical basis vectors obtained for the features $X$ and $Y$ respectively.

We now consider the fusion of the two feature for an arbitrary image $i$ with $\tilde{x}_i$ and $\tilde{y}_i$ being the outputs of the two feature extractors. We project these features onto the canonical basis vectors and are combined as

\[
z_{i1} = [A^T \tilde{x}_i \quad B^T \tilde{y}_i] = [A \quad 0] \tilde{x}_i
\]

\[
z_{i2} = (A^T \tilde{x}_i + B^T \tilde{y}_i) = [A^T \quad B^T] \begin{bmatrix} \tilde{x}_i \\ \tilde{y}_i \end{bmatrix}.
\]

We refer to Equation(14) as Feature Fusion Method - 1 (FFM-1), and Equation(15) as Feature Fusion method - 2 (FFM-2). $z_{i1}$ and $z_{i2}$ are the combined features for an image $i$. Note that we are combining two different features independent of their coordinates and their length.

We use both Euclidean distance [3] and Cosine similarity measure [5] to find a match. The test image is represented by $z_{ik}$, where $k$ corresponds to the Fusion method used. To find a match using Euclidean distance, we find the matching training image that satisfies

\[
\arg \min_{j \in \{1,2,\ldots,L\}} \|z_{ik} - z_{jk}\|^2
\]

for $k = 1, 2$. (17)

In Cosine similarity measure the cosine of the angle between the probe vector and training vectors is computed, and the match for the test vector $z_{ik}$ is found by performing

\[
\arg \max_{j \in \{1,2,\ldots,L\}} \left\{ \frac{z_{ik}^T z_{jk}}{|z_{ik}| \cdot |z_{jk}|} \right\}
\]

for $k = 1, 2$. (18)

4. Performance Comparison on various Face Databases

We have tested CCA-based fusion on Yale Face Database, AR Face Database [10] and FERET Database [11],[12]. Intensity normalization was done on the AR and FERET Databases using Block Histogram Modification (BHM)[13]. This simple technique is effective in reducing the illumination variations to some extent [13].

4.1. Yale Face Database

The Yale Face Database has a total of 165 images with 15 subjects, with 11 images per subject. These images contain various facial expressions and illumination changes.
We have cropped the face images to retain a size $64 \times 64$, while maintaining the same coordinates for the eyes in all training and test images. We use 4 images per subject in the training set and the remaining for testing. All 15 subjects are included in the training set. PCA, LDA and DCT-based PCA analysis is done with the training set to obtain the feature sets necessary for CCA-based Feature Fusion. For the DCT-based PCA, the first 1500 DCT coefficients are retained for the PCA.

The results obtained with PCA-LDA fusion and LDA-DCT-based PCA fusion are presented in Tables 1 and 2. It is observed CCA based Feature Fusion methods give better recognition rates with respect to individual methods. The combination of LDA and DCT gives the best results compared to other combinations of PCA and LDA. The Cosine similarity measure gives better results compared to the Euclidean distance. FFM-1 and FFM-2 work equally effectively.

4.2. AR Face Database

The AR Face Database [10] consists of over 4000 images corresponding to 126 subjects (76 males, 50 females). All images are of frontal view with different expression, occlusions and illumination variations. The images have been cropped to retain a size $64 \times 64$. We train the system with 4 images per subject. There are total of 272 training and 504 testing images. We perform PCA, LDA and DCT-based PCA on the training set. Combination of PCA and LDA, PCA and DCT-based PCA, LDA and DCT-based PCA are used for CCA-based fusion.

The results obtained with PCA and LDA fusion and LDA and DCT-based PCA fusion are presented in Tables 3 and 4. From these results it is seen that LDA and DCT combination give a good boost in the recognition rates, an increase of over 14% is observed. The combination of PCA and LDA is also effective, giving an increase of rates above 10%, although PCA and DCT-based PCA fusion dont give a substantial amount of improvement. It is again observed that the Cosine similarity measure works better than the Euclidean distance.

4.3. FERET Database

The FERET database [11],[12] consists of 14,051, gray scale images of around 1199 individuals. The database contains significant variation of pose, illumination and facial expression. In our tests we have considered 359 subjects. Two images per subject are used for training the system 1166 images are used for testing. As to the other databases, images are cropped to retain a size of $56 \times 56$. As discussed earlier, for DCT-based PCA we retain the top 1500 DCT coefficients, Principal Components are computed for these DCT feature vectors.

The results obtained in fusion of PCA and LDA and fusion of LDA and DCT-based PCA are presented in Tables 5 and 6. It is again observed that fusion of PCA and DCT-based PCA does not give substantial improvement in the rates. An increase of at least 11% recognition rate is observed in the performance of Cosine similarity measure, with both PCA and LDA fusion as well as LDA and DCT-based PCA fusion. Thus, from the results it is seen that CCA-based fusion with Cosine similarity measure gives better recognition rates compared to individual methods.

Table 3: Performance of PCA, LDA and FFM using PCA and LDA on AR Database

<table>
<thead>
<tr>
<th>Dist. Measure</th>
<th>PCA</th>
<th>LDA</th>
<th>FFM-1</th>
<th>FFM-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>67.23%</td>
<td>69.09%</td>
<td>73.99%</td>
<td>75.00%</td>
</tr>
<tr>
<td>Cosine</td>
<td>72.64%</td>
<td>72.13%</td>
<td>82.26%</td>
<td>83.11%</td>
</tr>
</tbody>
</table>

Table 4: Performance of LDA, DCT and FFM using LDA and DCT-based PCA on AR Database

<table>
<thead>
<tr>
<th>Dist. Measure</th>
<th>LDA</th>
<th>DCT</th>
<th>FFM-1</th>
<th>FFM-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>69.09%</td>
<td>67.74%</td>
<td>81.59%</td>
<td>81.25%</td>
</tr>
<tr>
<td>Cosine</td>
<td>72.13%</td>
<td>70.10%</td>
<td>86.49%</td>
<td>86.49%</td>
</tr>
</tbody>
</table>

Table 5: Performance of PCA, LDA and FFM using PCA and LDA on FERET Database

<table>
<thead>
<tr>
<th>Dist. Measure</th>
<th>PCA</th>
<th>LDA</th>
<th>FFM-1</th>
<th>FFM-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>54.63%</td>
<td>47.34%</td>
<td>49.14%</td>
<td>51.02%</td>
</tr>
<tr>
<td>Cosine</td>
<td>59.00%</td>
<td>59.52%</td>
<td>73.41%</td>
<td>75.04%</td>
</tr>
</tbody>
</table>

Table 6: Performance of LDA, DCT-based PCA and FFM using LDA and DCT-based PCA on FERET Database

<table>
<thead>
<tr>
<th>Dist. Measure</th>
<th>LDA</th>
<th>DCT</th>
<th>FFM-1</th>
<th>FFM-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>47.34%</td>
<td>53.77%</td>
<td>55.40%</td>
<td>55.57%</td>
</tr>
<tr>
<td>Cosine</td>
<td>59.52%</td>
<td>56.18%</td>
<td>71.01%</td>
<td>71.36%</td>
</tr>
</tbody>
</table>
5. Conclusions
In this paper we have discussed methods of combining two features extractors using CCA. We have observed that fusion of DCT-based PCA and LDA with Cosine similarity measure works the best, and gave better recognition rates than the individual methods. FFM-1 and FFM-2 are equally effective and both are simple to implement, although due to concatenation the vector length of FFM-1 is twice that of FFM-2 with FFM-1 computationally complex. Another important observation is that the Cosine similarity measure consistently outperforms the Euclidean distance.

6. Acknowledgements
Portion of this paper uses FERET Database [11],[12] collected under the FERET program.

7. References