IMPROVEMENTS ON CCA MODEL WITH APPLICATION TO FACE RECOGNITION

Quan-Sen Sun[a, b], Mao-Long Yang[a], Pheng-Ann Heng[c] and De-Sen Xia[a]
[a]Department of Computer Science, Nanjing University of Science &Technology, Nanjing 210094,People' Republic of China. [b]Department of Mathematics, Jinan University, Jinan 250022, People' Republic of China. [c]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong

Abstract: Two new methods for combination feature extraction are proposed in this paper. The methods are based on the framework of CCA in image recognition by improving the correlation criterion functions. Comparing with CCA methods, which can solve the classification of high-dimensional small size samples directly, being independent of the total scatter matrix singularity of the training simples, and the algorithms' complexity can be lowered. We prove that the essence of two improved criterion functions is partial least squares analysis (PLS) and multivariate linear regression (MLR). Experimental results based on ORL standard face database show that the algorithms are efficient and robust.

Key words: canonical correlation analysis(CCA); feature extraction; feature fusion; partial least squares (PLS); multivariate linear regression(MLR); face recognition

1. INTRODUCTION

In recent years, feature level fusion plays an important role in the process data fusion, which has achieved delightful development [1]. The advantage of feature level fusion is obvious, different feature vectors extracted from the same pattern always reflects different features of patterns. By optimizing and combining these different features, it not only keeps the effective discriminant information of multi-feature, but also eliminates redundant

information to certain degree. This is especially important to the problem of classification and recognition.

There exist two feature fusion methods. One is to group two sets of feature vectors into one union-vector [2], and then to extract features in a higher-dimension real vector space. Another one is to combine two sets of feature vectors by a complex vector[1], and then to extract features in the complex vector space. Both feature fusion methods aim at increasing the recognition rate.

In paper [3], we have proposed a new feature fusion strategy based on the idea of CCA, by creating a framework of CCA in image recognition. We have obtained good results in the application of the framework in the fields of face recognition and handwritten character recognition.

The work in detail is presented in this paper. The relation between the CCA, PLS and MLR was built by improving the correlation criterion function. PLS and MLR have been applied in the field of feature fusion. Comparing with CCA methods, the advantage of PLS and MLR is that they can solve classification problem of high-dimensional space and small sample size, be independent of the total scatter matrix singularity of the training simples, and the algorithms' complexity is lowered greatly. Experimental results based on ORL standard face database show that the algorithms are efficient and robust, which are superior to experimental results of classical Eigenfaces and Fisherfaces.

2. A FRAMEWORK OF CCA IN IMAGE RECOGNITION

2.1 The basic idea of CCA

In Multivariate Statistical Analysis, the correlation problem of two random vectors are often studied, that is to convert the correlation research of two random vectors into that of a few pairs of variables, which are uncorrelated. H. Hotelling developed this idea in 1936 [4].

Considering two zero-mean random vectors X and Y, CCA finds a pair of directions α and β that maximize the correlation between the projections $x^* = \alpha^T x$ and $y^* = \beta^T y$. This correlation is called the canonical correlation.

In general, the projective directions α and β are obtained by maximizing the correlation criterion function as follows:

$$\rho = \frac{E[\alpha^{\mathrm{T}} x y^{\mathrm{T}} \beta]}{\sqrt{E[\alpha^{\mathrm{T}} x x^{\mathrm{T}} \alpha] \cdot E[\beta^{\mathrm{T}} x y^{\mathrm{T}} \beta]}} = \frac{\alpha^{\mathrm{T}} S_{xy} \beta}{\sqrt{\alpha^{\mathrm{T}} S_{xx} \alpha \cdot \beta^{\mathrm{T}} S_{yy} \beta}}.$$

Where S_{xx} and S_{yy} denote the covariance matrixes of x and y respectively, while S_{xy} is their between-set covariance matrix.

2.2 The theory of combine feature extraction

Suppose $\omega_1, \omega_2, \cdots, \omega_c$ are c known pattern classes. Let $\Omega = \{\xi \mid \xi \in \mathrm{R}^n\}$ be a training sample space. Given $A = \{x \mid x \in \mathrm{R}^p\}$, $B = \{y \mid y \in \mathrm{R}^q\}$, where x and y are two feature vectors of the same sample ξ extracted by different means. We will discuss the feature fusion in the transformed training sample feature space A and B.

Our idea is to extract the canonical correlation features between x and y based on the idea of CCA proposed in Section 2.1, we denote them as $\alpha_1^{\mathrm{T}} x$ and $\beta_1^{\mathrm{T}} y$ (the first pair), $\alpha_2^{\mathrm{T}} x$ and $\beta_2^{\mathrm{T}} y$ (the second pair), $\cdots$, $\alpha_d^{\mathrm{T}} x$ and $\beta_d^{\mathrm{T}} y$ (the d th pair). Given the following:

$$X^* = (\alpha_1^{\mathrm{T}} x, \alpha_2^{\mathrm{T}} x, \cdots, \alpha_d^{\mathrm{T}} x)^{\mathrm{T}} = (\alpha_1, \alpha_2, \cdots, \alpha_d)^{\mathrm{T}} x = W_x^{\mathrm{T}} x;$$
$$Y^* = (\beta_1^{\mathrm{T}} y, \beta_2^{\mathrm{T}} y, \cdots, \beta_d^{\mathrm{T}} y)^{\mathrm{T}} = (\beta_1, \beta_2, \cdots, \beta_d)^{\mathrm{T}} y = W_y^{\mathrm{T}} y.$$

In paper [3], we had already given two feature fusion strategies:

$$\text{FFS I}: Z_1 = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_x^{\mathrm{T}} x \\ W_y^{\mathrm{T}} y \end{pmatrix} = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} x \\ y \end{pmatrix} \tag{1}$$

$$\text{FFS II}: Z_2 = X^* + Y^* = W_x^{\mathrm{T}} x + W_y^{\mathrm{T}} = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} x \\ y \end{pmatrix} \tag{2}$$

Two linear transformations (1) and (2) are used for classification by the projected feature vectors correspondingly, while the transformation matrix is: $W_1 = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}$ and $W_2 = \begin{pmatrix} W_x \\ W_y \end{pmatrix}$,where $W_x = (\alpha_1, \alpha_2, \cdots \alpha_d), W_y = (\beta_1, \beta_2, \cdots \beta_d)$.

Define. We call α_i and β_i as the i^{th} pair of canonical projective vectors (CPV) of x and y, and $\alpha_i^{\mathrm{T}} x$ and $\beta_i^{\mathrm{T}} y$ as the i^{th} canonical features of x and y. We also call Z_1 and Z_2 as the canonical discriminant features.

Next, we will discuss how to obtain the value and quality of CPV as follows:

Supposed S_{xx} and S_{yy} are positive definite, and $S_{xy}^{\mathrm{T}} = S_{yx}$, $r = rank(S_{xy})$.

We can give the criterion function as follows:

$$J(\alpha,\beta)=\frac{\alpha^{\mathrm{T}}S_{xy}\beta}{(\alpha^{\mathrm{T}}S_{xx}\alpha\cdot\beta^{\mathrm{T}}S_{yy}\beta)^{1/2}} \tag{3}$$

$$\text{Let}\quad \alpha^{\mathrm{T}}S_{xx}\alpha=\beta^{\mathrm{T}}S_{yy}\beta=1 \tag{4}$$

Then the problem is equivalent to finding the CPV α and β with constraint (4) which maximizes the criterion function (3).

Supposing the first pair of CPV (α_1,β_1) has been computed, after the first $(k\text{-}1)$ CPV $(\alpha_1,\beta_1),(\alpha_2,\beta_2),\cdots(\alpha_{k-1},\beta_{k-1})$ have been chosen, the k^{th} one can be computed by solving the following optimization problem:

$$\text{Model 1}\begin{cases}\max J(\alpha,\beta)\\ \alpha^{\mathrm{T}}S_{xx}\alpha=\beta^{\mathrm{T}}S_{yy}\beta=1\\ \alpha_i^{\mathrm{T}}S_{xx}\alpha=\beta_i^{\mathrm{T}}S_{yy}\beta=0\ \ (i=1,2,\cdots,k-1)\end{cases} \tag{5}$$

According to the method of Lagrange multipliers, the question can be transformed to the solving of two generalized eigenproblem:

$$\begin{cases}S_{xy}S_{yy}^{-1}S_{yx}\alpha=\lambda^2S_{xx}\alpha & (6)\\ S_{yx}S_{xx}^{-1}S_{xy}\beta=\lambda^2S_{yy}\beta & (7)\end{cases}$$

In order to obtain the solution under the restricted condition (5), supposing that $H=S_{xx}^{-1/2}S_{xy}S_{yy}^{-1/2}$. Applying Singular Value Decompose (SVD) theorem on matrix H, we obtain $H=\sum_{i=1}^{r}\lambda_i u_i v_i^{\mathrm{T}}$, where $\lambda_1^2\geq\lambda_2^2\geq\cdots\geq\lambda_r^2$ are entire nonzero eigenvalues of $G_1=H^{\mathrm{T}}H$ and $G_2=HH^{\mathrm{T}}$, u_i and v_i are the orthogonal eigenvectors of G_1 and G_2 corresponding to the nonzero eigenvalue λ_i^2, where $i=1,2,\cdots,r$.

From above, we can refer to the important theorem [3]:

Theorem 1. Given $\beta_i=S_{yy}^{-1/2}v_i$, $\alpha_i=\lambda_i^{-1}S_{xx}^{-1}S_{xy}\beta_i$, $i=1,\cdots,r$. Then

(1) α_i and β_i are the eigenvectors of generalized eigenequation (6)and (7) corresponded to λ_i^2;

$$(2)\begin{cases}\alpha_i^{\mathrm{T}}S_{xx}\alpha_j=\beta_i^{\mathrm{T}}S_{yy}\beta_j=\delta_{ij}\\ \alpha_i^{\mathrm{T}}S_{xy}\beta_j=\lambda_i\delta_{ij}\end{cases}(i,j=1,2,\cdots,r).\ \text{where}\ \delta_{ij}=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases} \tag{8}$$

From the above discussion, we can draw the following conclusion:

Theorem 2. Under criterion (3), the number of the efficient CPV satisfying the restricted condition(4) and (5), is r pair(s) at most($r = rank(S_w)$), and getting $d(\leq r)$ pair CPV are compose of the eigenvectors corresponding to first d maximum eigenvalues of two generalized eigenequation (6) and (7) that satisfy Eq.(8).

3. THE IMPROVEMENTS OF THE MODEL

From the above we can find that the matrices S_{xx} and S_{yy} must be positive definite when solving the projections and the discriminant features based on the criterion function (3). In the field of pattern recognition, especially in face recognition, cases of high-dimensional space and small sample size are common, where the total scatter matrix of the training simples is singular. A solution, where S_{xx} and S_{yy} are singular, was proposed in paper [3], which can obtain the projective vectors in low–dimensional feature space. Two ways are proposed as follow by improving the criterion function (3).

3.1 Improvement 1: PLS

In the denominator of the criterion function (3), replacing both S_{xx} and S_{yy} by the unit matrices, the criterion function can be written as:

$$J_P(\alpha,\beta) = \frac{\alpha^{\mathrm{T}} S_{xy} \beta}{(\alpha^{\mathrm{T}}\alpha \cdot \beta^{\mathrm{T}}\beta)^{1/2}} \tag{9}$$

Then the problem is equivalent to finding the projections α and β with constraint (10) which maximizes the criterion function (9).

$$\alpha^{\mathrm{T}}\alpha = \beta^{\mathrm{T}}\beta = 1 \tag{10}$$

Based on the criterion function (9), the process of obtaining projection and correlation feature with the constraint (10) shares exactly the same idea with partial least squares analysis, namely PLS [5].

The step of feature fusion and image recognition, which employ PLS, is similar to what we have discussed in section 2: to solve the projection; then to extract combination features by FFS I and FFS II. So we can write model 1 as:

$$\text{Model 2} \quad \begin{cases} \max J_p(\alpha,\beta) \\ \alpha^{\mathrm{T}}\alpha = \beta^{\mathrm{T}}\beta = 1 \\ \alpha_i^{\mathrm{T}}\alpha = \beta_i^{\mathrm{T}}\beta = 0 \ \ (i = 1,2,\cdots,k-1) \end{cases} \tag{11}$$

Then the generalized eigenequations (6) and (7) can be written as:

$$\begin{cases} S_{xy}S_{yx}\alpha = \lambda^2\alpha & (12) \\ S_{yx}S_{xy}\beta = \lambda^2\beta & (13) \end{cases}$$

We can obtain the optimal solutions satisfying model 2 from eigen-equations (12) and (13) as follows:

Theorem 3. Under the criterion (9), the number of effective projective vectors, which satisfy restricted constraints (10) and (11), is r ($r = rank(S_{xy})$) pairs at most. $d(\le r)$ pairs of projective vectors are composed of vectors which are selected from the eigenvectors corresponding to the first d maximum eigenvalues of eigenequations (14) and (15) and satisfying:

(a) all the eigenvectors should satisfy: $\alpha_i = \lambda_i^{-1}S_{xy}\beta_i$, $\beta_i = \lambda_i^{-1}S_{yx}\alpha_i$, $i = 1,\cdots,r$;

(b) $\begin{cases} \alpha_i^{\mathrm{T}}\alpha_j = \beta_i^{\mathrm{T}}\beta_j = \delta_{ij} \\ \alpha_i^{\mathrm{T}}S_{xy}\beta_j = \lambda_i\delta_{ij} \end{cases} \quad (i,j = 1,2,\cdots,r)$.

where λ_i^2 is the non-zero eigenvalue of the two eigenequations, where eigenvectors are α_i and β_i ($i = 1,2,\cdots,r$), respectively.

Proof. By the method of Lagrange multipliers, function $L(\alpha,\beta)$ is defined as $L(\alpha,\beta) = \alpha^{\mathrm{T}}S_{xy}\beta - \frac{\lambda_1}{2}(\alpha^{\mathrm{T}}\alpha - 1) - \frac{\lambda_2}{2}(\beta^{\mathrm{T}}\beta - 1)$. Let $\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial \beta} = 0$, then the following can be derived:

$$\begin{cases} S_{xy}\beta - \lambda\alpha = 0 \\ S_{yx}\alpha - \lambda\beta = 0 \end{cases} \Rightarrow \begin{cases} \alpha = \lambda^{-1}S_{xy}\beta \\ \beta = \lambda^{-1}S_{yx}\alpha \end{cases} \tag{14}$$

Then eigenequations (12) and (13) can be derived from Eq.(14). Both $S_{xy}S_{yx}$ and $S_{yx}S_{xy}$ are symmetric matrices, $rank(S_{xy}S_{yx}) = rank(S_{xy}) = r$, so that the two eigenequations have the same non-zero eigenvalue $\lambda_1^2 \ge \lambda_2^2 \ge \cdots \ge \lambda_r^2 > 0$, and the r pairs of eigenvectors corresponding to them are orthonormal, namely $\alpha_i^{\mathrm{T}}\alpha_i = \beta_i^{\mathrm{T}}\beta_i = \delta_{ij}$. Conclusion (a) is true because Eq.(16), and conclusion (b) is true as $\alpha_i^{\mathrm{T}}S_{xy}\beta_j = \alpha_i^{\mathrm{T}}S_{xy}(\lambda_j^{-1}S_{yx}\alpha_j) = \lambda_j^{-1}\alpha_i^{\mathrm{T}}(\lambda_j^2\alpha_j) = \lambda_j\delta_{ij}$, too. □

3.2 Improvement by MLR:

In the denominator of the criterion function (3), replacing S_{yy} by the unit matrices, the criterion function can be written as

$$J_M(\alpha,\beta)=\frac{\alpha^{\mathrm{T}}S_{xy}\beta}{(\alpha^{\mathrm{T}}S_{xx}\alpha\cdot\beta^{\mathrm{T}}\beta)^{1/2}} \tag{15}$$

In this sense, the problem is equivalent to the finding of projections α and β with the constraint (16) which maximizes the criterion function (15):

$$\alpha^{\mathrm{T}}S_{xx}\alpha=\beta^{\mathrm{T}}\beta=1 \tag{16}$$

Based on the criterion function (15), the process of obtaining the projection and correlation feature with the constraint (16) implies the idea of multivariate linear regression, namely MLR.

In fact, under the condition of Section 2.1, the goal of MLR is to minimize following the square error[6]:

$$\varepsilon^2=E[y^{\mathrm{T}}y]-2\theta\alpha^{\mathrm{T}}S_{xy}\beta+\theta^2\alpha^{\mathrm{T}}S_{xx}\beta \tag{17}$$

where θ is a regression coefficient. Let

$$\frac{\partial\varepsilon^2}{\partial\theta}=2(\theta\alpha^{\mathrm{T}}S_{xx}\alpha-\alpha^{\mathrm{T}}S_{xy}\beta)=0\Rightarrow\theta=\frac{\alpha^{\mathrm{T}}S_{xy}\beta}{\alpha^{\mathrm{T}}S_{xx}\alpha}\ .$$

By inserting this expression into Eq. (17), we get

$$\varepsilon^2=E[y^{\mathrm{T}}y]-\frac{(\alpha^{\mathrm{T}}S_{xy}\beta)^2}{\alpha^{\mathrm{T}}S_{xx}\alpha}.$$

In order to minimize ε^2, what should be done is just to maximize the following quotient function:

$$\rho=\frac{\alpha^{\mathrm{T}}S_{xy}\beta}{(\alpha^{\mathrm{T}}S_{xx}\alpha)^{1/2}}=\frac{\alpha^{\mathrm{T}}S_{xy}\beta}{(\alpha^{\mathrm{T}}S_{xx}\alpha\cdot\beta^{\mathrm{T}}\beta)^{1/2}} \tag{18}$$

where $\beta^{\mathrm{T}}\beta=1$.

Comparing function (15) with function (18), one will find that they go all the way if constraint (16) is satisfied.

The discussion on the application of feature fusion and image recognition by MLR is similar to that in section 3.1, which is omitted.

3.3 Comparison of the three methods

In order to depict clearly, we call the three methods as CCA, PLS, MLR respectively for short, and call the projection and the discriminant feature by criterion functions (3), (9), and (14) as correlation projective vector (CPV) and correlation discriminant feature vector (CDV), respectively.

All three methods can solve the problem of compressing the pattern feature dimensions effectively. They all make the two sets of feature having maximal correlation based on maximizing their covariance in order to identify the projection. But their restrictions to projection are different: the projection should be conjugate orthogonal for S_{xx} and S_{yy} in CCA, and should be orthonormal in PLS, while one set projection should be orthonormal, the other should be conjugate orthogonal for S_{xx} or S_{yy} in MLS.

The two methods of PLS and MLR can be seen as the special instances of the first method (CCA) in some sense, so the process in solving the CPV and the CDV is coincident. On the other hand, PLS is superior to CCA in arithmetic complexity, while MLR's is between that of PLS and CCA.

PLS has a better generalization capability than the other two methods because it is independent of the total scatter matrix singularity of the training simples. Moreover, PLS is effective on the classification of both large size simples and small ones. MLR can make up the influence caused by the total scatter matrix when it is singular. We only need to make sure that the total scatter matrix is composed of one of the two set features of the same pattern is nonsingular.

4. EXPERIMENT AND ANALYSIS

Experiment is performed on the ORL face image database. There are 10 different images for 40 individuals. For some people, images were taken at different times. And the facial expression (open/closed eyes, smiling/ nonsmiling) and facial details (glasses/no glasses) are variables. The images were taken against a dark homogeneous background and the people are in upright, frontal position with tolerance for some tilting and rotation of up to 20°. Moreover, there is some variation in scale of up to about 10%. All

images are grayscale and normalized with a resolution of 92×112. Some images in ORL are shown in Fig.1.

Fig.1. Ten images of one person in ORL face database

In this experiment, we use the first five images of each person for training and the remaining five for testing. Thus, the total amount of training samples and testing samples are both 200.

In the experiment, the rank of the total covariance matrix S_t is computed first, and it is equal to 199. Then, we translate the original image vectors into 199 dimensional feature space $\Omega = \{\xi \mid \xi \in \mathrm{R}^{199}\}$ by K-L transform, and then decompose it into 2 parts: 59D and 140D, namely $\xi = \begin{bmatrix} x \\ y \end{bmatrix}$, which composes the feature sub-space $A = \{x \mid x \in \mathrm{R}^{59}\}$ and $B = \{y \mid y \in \mathrm{R}^{140}\}$, respectively.

Then the CPV and the correlation discriminant feature vector are solved by the above three methods, respectively, and the combination features are extracted by the two feature fusion strategies FFSI and FFSII, which are classified finally by the minimum distance classifier and the nearest-neighbor classifier, respectively. The corresponding recognition rates are shown in table 1.

Furthermore, we present the recognition results on ORL face image database of classical Eigenfaces method[7] and Fisherface method[8] in table 1.

Table 1 Recognition rates of different classifier

Classifier	FFS1			FFS2		Eigenface	Fesherface
	CCA	PLS	MLR	PLS	MLR		
M-distance	0.915	0.920	0.940	0.925	0.935	0.895	0.885
N-neighbor	0.905	0.950	0.910	0.945	0.905	0.930	0.885

Table 1 shows that the recognition rates of the three methods by FFSI and FFSII with the minimum distance classifier are higher than those of the Eigenfaces method and the Fisherfaces method, which can be up to 91%. MLR, moreover, is better than CCA and PLS by FFSI and FFSII, of which the optimal recognition correct rates can be up to 94%.

Recognition rates of the PLS by FFSI and FFSII with the nearest-neighbor classifier are higher than that of the other methods which can be up to 95%, while the rates of CCA and MLR are less than the Eigenfaces method, but higher than that of the Fisherfaces method.

Experimental results have shown that the two improved models (PLS and MLR) are superior to previous methods (CCA) in arithmetic complexity and classification capability.

5. CONCLUSION

We have presented two new feature fusion methods — PLS and MLR by improving the combination feature extraction methods — CCA, which broaden the field of combination feature extraction. Experimental results show that the improved methods are superior to the original methods, which arithmetic complexities are reduced significantly.

ACKNOWLEDGEMENTS

We wish to thank the CUHK fund from HKSAR Government under Grant No. 4185 / 00E for supporting.

REFERENCES

1. Yang Jian, Yang Jing-yu, Zhang David, Lu Jian-feng, Feature fusion: parallel strategy vs. serial strategy, Pattern Recognition, 36 (2003) 1961-1971.
2. Liu C J, Wechsler H, A shape-and texture-based enhanced Fisher classifier for face recognition, IEEE Transactions on Image Processing, 10 (4) (2001)598-608.
3. Quan-Sen Sun, Sheng-Gen Zeng, Yan Liu, Pheng-Ann Heng, De-Shen Xia. A new method of feature fusion and its application in image recognition. Pattern Recognition(USA), in review.
5. A. Höskuldsson, PLS regression methods, Journal of Chemometrics, 2(1988)211-228.
6. M. Borga, Learning Multidimensional Signal Processing, Linköping Studies in Science and Technology, issertations, No.531, Department of Electrical Engineering, Linköping University, Linköping , Sweden, 1998.
7. Turk M and Pentland A. Eigenfaces for recognition. J. Cognitive Neuroscience, 1991, 3(1):71-86.
8. Peter N. Belhumeur, et al. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Machine Intell. 19(7) (1997) 711-720.