A Discriminant Analysis Method for Face Recognition in Heteroscedastic Distributions

Zhen Lei, Shengcai Liao, Dong Yi, Rui Qin, and Stan Z. Li*

Center for Biometrics and Security Research,
Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun Donglu, Beijing 100190, China
{zlei,scliao,dyi,rqin,szli}@cbsr.ia.ac.cn

Abstract. Linear discriminant analysis (LDA) is a popular method in pattern recognition and is equivalent to Bayesian method when the sample distributions of different classes are obey to the Gaussian with the same covariance matrix. However, in real world, the distribution of data is usually far more complex and the assumption of Gaussian density with the same covariance is seldom to be met which greatly affects the performance of LDA. In this paper, we propose an effective and efficient two step LDA, called LSR-LDA, to alleviate the affection of irregular distribution to improve the result of LDA. First, the samples are normalized so that the variances of variables in each class are consistent, and a pre-transformation matrix from the original data to the normalized one is learned using least squares regression (LSR); second, conventional LDA is conducted on the normalized data to find the most discriminant projective directions. The final projection matrix is obtained by multiply the pre-transformation matrix and the projective directions of LDA. Experimental results on FERET and FRGC ver 2.0 face databases show the proposed LSR-LDA method improves the recognition accuracy over the conventional LDA by using the LSR step.

Keywords: Least squares regression (LSR), discriminant analysis, face recognition.

1 Introduction

Subspace learning has attracted much attention and achieved great success in face recognition research area during the last decades. Among various methods, PCA and LDA [1]are the two most representative ones. PCA uses the Karhunen-Loeve transform to produce the most expressive subspace for face representation and recognition by minimizing the residua of the reconstruction. However, it does not utilize any class information and so it may drop some important clues for classification. LDA is then proposed and it seeks subspace of features best separating different face classes by maximizing the ratio of the between-class scatter matrix to the within-class scatter.

In theory, LDA is equivalent to Bayesian method if the distribution of samples in each class is obey to the Gaussian density with the same covariance matrix [2]. However, in practice, the covariance matrices from different classes are always heteroscedastic

* Corresponding author.

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 112–121, 2009.

and it deteriorates the performance of LDA. There are usually two ways to handle this problem: One is to exploit specific heteroscedastic LDA [3,4,5,6] to solve the problem, not only taking into account the discriminatory information between class means, but also the differences of class covariance matrices. The other is to try to make the class covariance matrices consistent and thus to improve the performance of LDA on it.

In this paper, we follow the latter one and propose a least squares regression (LSR) based processing to normalize the distribution of samples before conventional LDA. In [7], researchers take into account the scale properties of each variable of feature on the whole data set and re-scale the variables to enhance the performance of LDA. In fact, each variable in different classes has different scale properties, and it is more reasonable to re-scale each variable in different classes individually. In our method, first, each variable of feature in the same class is normalized to be of unit variance so that the variables in every class have the same variance that makes the class distributions more consistent. Second, for its simplicity and effectiveness, we utilize LSR to learn the transform matrix from original data to the normalized one. If each variable of feature is independent, this normalization process would guarantee the covariance matrices from different classes identical. Even the variables are not independent, we argue this processing will still make the distributions of samples from different classes more consistent and therefore to improve the result of LDA. Besides, this normalization process could also make the sample distribution of each class more compact so that it is able to increase the separability of classes. Fig. 1 shows a toy example. The left one is the original data distribution and the right one is the data distribution after normalization. The line is the LDA projective direction. It explicitly illuminates the effect of the proposed normalization process for improving the separability of LDA result. Regarding the computational cost, compared to the heteroscedastic LDA [6], the proposed method only involves the LSR instead of the complex iterative optimization process, thus it is very efficient and can be applied in large scale data set.

The remainder of the paper is organized as follows. Section 2 reviews conventional LDA and details the LSR-LDA method. Section 3 compares the results of the proposed

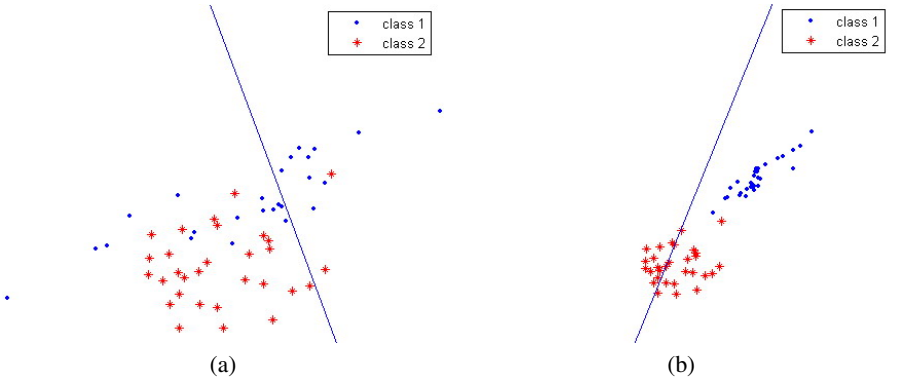


Fig. 1. A toy example that shows the advantage of the proposed normalization step, with data distribution and LDA projection before (left) and after (right) LSR normalization

method with other methods on FERET and FRGC databases and in Section 4, we conclude the paper.

2 LSR-LDA

2.1 Conventional LDA

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ be a d-dimensional data set with n elements from $\{C_i | i = 1, 2, \ldots, L\}$ classes, where L is the number of the total classes. The within class scatter matrix S_w and the between class scatter matrix S_b in LDA are defined as follows:

$$S_w = \frac{1}{n} \sum_{i=1}^{L} \sum_{\mathbf{x}_j \in C_i} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T \tag{1}$$

$$S_b = \frac{1}{n} \sum_{i=1}^{L} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \tag{2}$$

where $\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$ is the mean of data in class C_i, and $\mathbf{m} = \frac{1}{n} \sum_{i=1}^{L} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$ is the global mean vector. LDA searches such optimal projections that after projecting the original data onto these directions, the trace of the resulting between class scatter matrix is maximized while the trace of the within class scatter matrix is minimized. Let $\mathbf{W}$ denote a $d \times d'$ ($d' < d$) projection matrix, LDA then chooses $\mathbf{W}$ so that the following object function is maximized:

$$J = \frac{tr(\tilde{S}_b)}{tr(\tilde{S}_w)} = \frac{tr(\sum_{i=1}^{L} n_i \mathbf{W}^T (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \mathbf{W})}{tr(\sum_{i=1}^{L} \sum_{\mathbf{x}_j \in C_i} \mathbf{W}^T (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T \mathbf{W})} = \frac{tr(\mathbf{W}^T S_b \mathbf{W})}{tr(\mathbf{W}^T S_w \mathbf{W})} \tag{3}$$

The optimal projection matrix $\mathbf{W}_{opt}$ can be obtained by solving the following eigenvalue problem corresponding to the d' largest non-zero eigenvalues.

$$S_w^{-1} S_b \mathbf{W} = \mathbf{W} \mathbf{\Lambda} \tag{4}$$

where $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the eigenvalues of $S_w^{-1} S_b$.

In real application, due to the usually high dimension of data and small size of samples, the within class scatter matrix S_w is often singular and the inverse of S_w does not exist, so the optimal solution of LDA in Eq. 4 cannot be found directly. To deal with this problem, many variants of LDA have been proposed such as PCA+LDA (Fisher-LDA or FLDA), Direct LDA (DLDA), Null space LDA (NLDA) [1,8,9] etc.. FLDA firstly utilizes PCA to select the most expressive subspace and reduces the dimension of feature to make S_w non-singular and then conducts LDA to derive the optimal projective directions. DLDA takes no account of the impact of S_w and finds the most discriminative projections in range space of S_b directly. NLDA is proposed to find the projections that maximize the between class scatter matrix S_b in the null space of S_w so that the singularity problem is settled.

2.2 LSR Normalization

LDA's performance highly depends on the sample distribution. Theoretically, under Gaussian assumption, it achieves its best performance with the same covariance matrix for different classes. Therefore, how to make data distributions from different classes as consistent as possible is a key point to achieve good result for LDA method.

Considering this, before LDA, we normalize the data distribution so as to make each dimension of data in every class to be of the same unit variance. In this way, we can make the distributions of different classes more consistent than original one and thus hope to improve the performance of LDA.

Let the sample set from the k-th class be $\mathbf{X}^k = [\mathbf{x}_1^k, \mathbf{x}_2^k, \cdots, \mathbf{x}_{n_k}^k]$ of d dimension, where n_k is the sample number, and $\mathbf{m}^k = 1/n_k \sum_{i=1}^{n_k} \mathbf{x}_i^k$ is the mean vector of the k-th class. Then we normalize each dimension of data to be of unit variance. That is

$$\sigma_j^k = \sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}(x_{ji}^k - m_j^k)^2}$$

$$x'^k_{ji} = \begin{cases} (x_{ji}^k - m_j^k)/\sigma_j^k + m_j^k & \sigma_j^k \neq 0 \\ x_{ji}^k & \sigma_j^k = 0 \end{cases} \qquad j = 1, 2, \ldots, d, \quad i = 1, 2, \ldots, n_k$$

The above normalization method is feasible in training set. When there comes a new sample, because of its unknown class label, we don't know how to normalize it. To handle this problem, we turn to learn the relationship between the original data and the normalized one so that the learned relationship can then be generalized to normalize the unseen data.

Suppose the normalized sample set be $\mathbf{X}'$. Our purpose is to learn the relationship f that describes the transform from the original data to the normalized one $\mathbf{x}_i' = f(\mathbf{x}_i)$. Under linear assumption, it is simplified to learn the transformation matrix $\mathbf{W}_1$ between the two spaces $\mathbf{x}_i' = \mathbf{W}_1^T \mathbf{x}_i$. Fortunately, this problem could be solved in means of least squares regression (LSR).

$$\mathbf{W}_1 = \arg\min_{\mathbf{W}_1} \sum_{i=1}^{n} \|\mathbf{W}_1^T \mathbf{x}_i - \mathbf{x}_i'\|^2 = \arg\min_{\mathbf{W}_1} tr((\mathbf{W}_1^T \mathbf{X} - \mathbf{X}')(\mathbf{W}_1^T \mathbf{X} - \mathbf{X}')^T) \tag{5}$$

By putting the derivative of above function with respect to $\mathbf{W}_1$ to zero, we obtain

$$2(\mathbf{W}_1^T \mathbf{X} - \mathbf{X}')\mathbf{X}^T = 0 \Rightarrow \mathbf{W}_1 = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{X}'^T \tag{6}$$

In reality, the dimension of data is usually larger than the sample number and hence, the matrix $\mathbf{X}\mathbf{X}^T$ is deficient and its inverse does not exist. On the other hand, even the inverse of $\mathbf{X}\mathbf{X}^T$ exists, the derived result $\mathbf{W}_1$ may over-fit to the training set. In order to avoid the deficient problem and improve the generalization of the result, we impose regularized penalty, also known the prior knowledge [10] onto the objective function in Eq. 5 as

$$\mathbf{W}_1 = \arg\min_{\mathbf{W}_1} tr((\mathbf{W}_1^T \mathbf{X} - \mathbf{X}')(\mathbf{W}_1^T \mathbf{X} - \mathbf{X}')^T + \lambda \mathbf{W}_1 \mathbf{W}_1^T) \tag{7}$$

Input: Let d-dimensional sample set be $\mathbf{X} = \{\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^L\}$ from L classes $\{C_1, C_2, \ldots, C_L\}$, whose corresponding class means are $\{\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^L\}$. $\mathbf{X}^k = [\mathbf{x}_1^k, \mathbf{x}_2^k, \cdots, \mathbf{x}_{n_k}^k]$ denotes the n_k samples from the k-th class.

(a) For every dimension of data, normalize the variance of variable in each class in training set to be unit.

for $k = 1 : L$

  for $j = 1 : d$

    $\sigma_j^k = \sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}(x_{ji}^k - m_j^k)^2}$

    for $i = 1 : n_k$

$$x_{ji}'^k = \begin{cases} (x_{ji}^k - m_j^k)/\sigma_j^k + m_j^k & \sigma_j^k \neq 0 \\ x_{ji}^k & \sigma_j^k = 0 \end{cases}$$

    end

  end

end

(b) Learn the pre-transformation matrix $\mathbf{W}_1$ using least squares regression according to Eq. 8.

(c) Compute the within and between scatter matrices based on the normalized samples according to Eq. 1 and 2.

(d) Compute the conventional LDA projection matrix $\mathbf{W}_2$ on the normalized samples according to Eq. 4.

Output:The two-step LDA projection matrix $\mathbf{W} = \mathbf{W}_1\mathbf{W}_2$.

Fig. 2. LSR based Two Step LDA algorithm

where λ controls the trade-off between the fitting accuracy in the training set and the generalization. We can then obtain the optimal result as

$$\mathbf{W}_1 = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{X}'^T \tag{8}$$

In this way, when there comes a new data $\mathbf{x}_{new}$, we can normalize it by multiplying it with the learned transformation matrix $\mathbf{W}_1$ directly, $\mathbf{x}'_{new} = \mathbf{W}_1^T\mathbf{x}_{new}$.

2.3 LDA after LSR

After the normalization step, various LDA can be conducted on the normalized data to learn the most separable subspace and the final projective directions are obtained by multiplying the normalization transformation matrix with the LDA projections. The whole process of the LSR based two step LDA is illustrated in Fig. 2.

3 Experimental Results and Analysis

3.1 Data Preparation

Two face databases, FERET [11] and FRGC ver 2.0 [12] are tested. All the images are rotated, scaled and cropped to 44×40 according to the provided eye positions

(a) (b)

Fig. 3. Face examples of FERET (a) and FRGC (b) databases

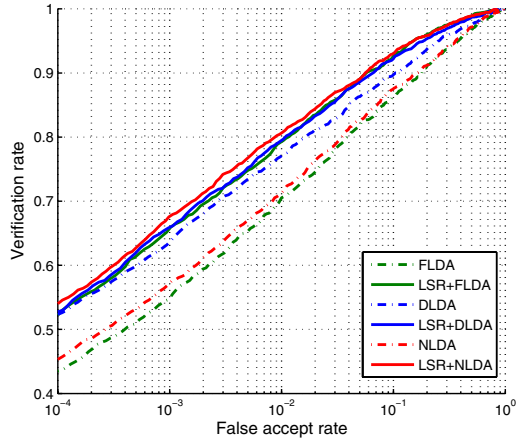
Table 1. The performance of different methods on FERET database

Methods	Rank-1	VR@FAR=0.001	EER
FLDA	0.6684	0.5557	0.1213
NLDA	0.7087	0.5751	0.1158
DLDA	0.7252	0.6390	0.1007
[7]+FLDA	0.6466	0.5012	0.1439
[7]+NLDA	0.6911	0.5396	0.1372
[7]+DLDA	0.7068	0.6087	0.1124
LSR-FLDA	0.6978	0.6594	0.0838
LSR-NLDA	0.7196	0.6793	0.0788
LSR-DLDA	0.7125	0.6618	0.0840

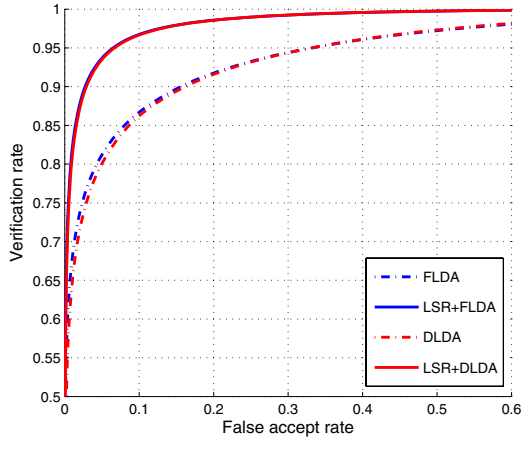
Table 2. The performance of different methods on FRGC v2.0 database

Methods	VR@FAR=0.001	EER
FLDA	0.5082	0.1153
DLDA	0.4366	0.1181
[7]+FLDA	0.5550	0.0875
[7]+DLDA	0.4386	0.1019
LSR-FLDA	0.5856	0.0552
LSR-DLDA	0.5585	0.0568

succeeded by histogram equalization preprocessing. For FERET database, the training set contains 731 images. In test phase, we use the gallery set containing 1196 images from 1196 subjects, and combine four provided probe sets (fb, fc, dupI, dupII) together, totally including 2111 images to compose the probe set. So our test protocol should be more difficult than any of the four original protocols because we consider different factors (expression, illumination, aging etc.) together to evaluate the performance. For FRGC database, the training set consists of 12776 face images from 222 individuals, including 6360 controlled and 6416 uncontrolled images. We choose experimental 4 setting which is considered as the most difficult case in FRGC to test the algorithm. In the test set, there are 16028 controlled images as the target set and 8014 query images which are uncontrolled ones, from 466 persons. The images are captured over several sessions. Fig. 3 illustrates some cropped face examples of FERET and FRGC databases.



(a)



(b)

Fig. 4. Receiver operating characteristic (ROC) curves of different methods on FERET (a) and FRGC (b) databases

3.2 Performance Evaluation

In this experiment, the regularization parameter λ is set to be 1.0 empirically. The cosine distance (Eq. 9) is adopt to measure the dissimilarity of features and the nearest neighbor (NN) classifier is chosen to do the classification task.

$$d_{cos}(\mathbf{x}, \mathbf{y}) = -\frac{\mathbf{x}^T\mathbf{y}}{\sqrt{\mathbf{x}^T\mathbf{x}\mathbf{y}^T\mathbf{y}}} \tag{9}$$

The proposed LSR based normalization method is combined with various versions of LDA such as FLDA, DLDA, NLDA and compared with these LDAs without normalization stage. Moreover, we also compare the results of normalization method in [7]

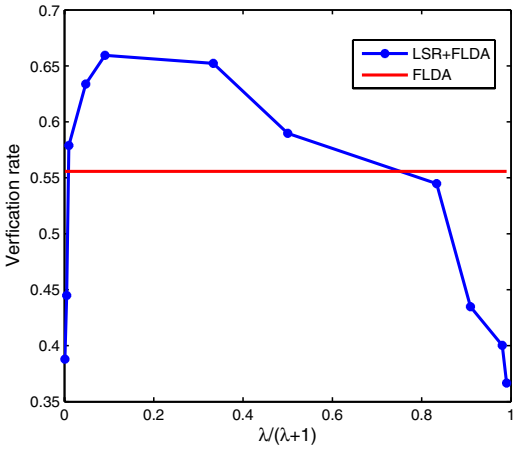


Fig. 5. Face verfication rates with respect to $\lambda/(\lambda + 1)$ on FERET database

Table 3. Computational cost of FLDA and LSR-FLDA on FRGC training set

	FLDA	LSR-FLDA
computational time (s)	23.69	31.71

on FERET and FRGC v2.0 respectively. For FERET, the results are reported as rank-1 recognition rate, verification rate (VR) when the false accept rate (FAR) is 0.001 and equal error rate (EER). For FRGC, we report the result as verification rate (VR) when the false accept rate (FAR) is 0.001 and equal error rate (EER) based on all matchings between query and target images.

Table 1 and 2 illustrate the recognition results of various methods on FERET and FRGC respectively and Fig. 4 shows the corresponding ROC curves. For clarity, we just plot the results of proposed method and conventional LDA in Fig. 4. For FRGC v2.0, because the sample number in training set is larger than the dimension of feature, there is no null space for within scatter matrix S_w, so that we don't have the results of NLDA. From the results, we can see that whether on FERET or FRGC database, whether with FLDA, DLDA or NLDA, the proposed LSR-LDA always obtains better results than conventional LDA, especially for FLDA, NLDA on FERET and FLDA, DLDA on FRGC database. It proves the proposed LSR based normalization step is an effective way to eliminate the affection of sample irregularity and could significantly improve the performance of LDA. It should be noted the normalization method in [7] doesn't always improve the performance of LDA. That may be because the databases in our experiments are larger and more challenge than that in [7].

The regularization coefficient λ is an adjustable parameter in our algorithm. We examine the impact of λ on recognition accuracy with LSR-FLDA on FERET database and plot the results with respect to different values of $\lambda/(\lambda + 1)$ in Fig. 5. The red line is the verification rate of original FLDA without the proposed normalization step. It can

be seen the performance of LSR-FLDA keeps a higher accuracy during a large scale values of λ which indicates the robustness of the proposed algorithm.

Table 3 illustrates the experimental computational cost for FLDA and LSR-FLDA on FRGC training set. It is the average of 10 times running on a Core 2 Duo 2.4GHz and 2GB RAM PC with unoptimized matlab code. It shows that the proposed LSR-FLDA doesn't increase the computational burden too much, about 1.3 times than the cost of original FLDA. Considering the accuracy improvement of LSR-LDA, this additional spend is completely tolerable in practical applications.

4 Conclusions

In this paper, we propose a LSR based two step LDA for face recognition. Before LDA, each dimension of data from the same class are normalized to be of unit variance so that the distributions of different classes are imposed to be consistent. After that, we utilize least squares regression (LSR) to learn the pre-transformation from original data to normalized one. Thus, when there comes new data, it can be transformed to the normalized distribution directly. The final projective directions of LSR-LDA are obtained by multiplying the pre-transformation and the projections of conventional LDA on normalized data. Experimental results on FERET and FRGC databases show the proposed LSR based two step LDA significantly improves the performance compared to the conventional one.

Acknowledgements. This work was supported by the following funding resources: National Natural Science Foundation Project #60518002, National Science and Technology Support Program Project #2006BAK08B06, National Hi-Tech (863) Program Projects #2006AA01Z192, #2006AA01Z193, and #2008AA01Z124, Chinese Academy of Sciences 100 people project, and AuthenMetric R&D Funds.

References

1. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
2. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. John Wiley and Sons, Chichester (2001)
3. Loog, M., Duin, R.P.: Linear dimensionality reduction via a heteroscedastic extension of lda: The chernoff criterion. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(6), 732–739 (2004)
4. Das, K., Nenadic, Z.: Approximate information discriminant analysis: A computationally simple heteroscedastic feature extraction technique. Pattern Recognition 41(5), 1548–1557 (2008)
5. Hsieh, P.F., Wang, D.S., Hsu, C.W.: A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(2), 223–235 (2006)
6. Kumar, N., Andreou, A.G.: Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. Speech Commun. 26(4), 283–297 (1998)

7. An, G., Ruan, Q.: Novel mathematical model for enhanced fisher's linear discriminant and its application to face recognition. In: Proceedings of International Conference on Pattern Recognition, pp. 524–527 (2006)
8. Yu, H., Yang, J.: A direct lda algorithm for high-dimensional data with application to face recognition. Pattern Recognition 34(10), 2067–2070 (2001)
9. Chen, L., Liao, H., Ko, M., Lin, J., Yu, G.: A new lda-based face recognition system which can solve the small sample size problem. Pattern Recognition 33(10), 1713–1726 (2000)
10. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, Heidelberg (2001)
11. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10), 1090–1104 (2000)
12. Phillips, P.J., Flynn, P.J., Scruggs, W.T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.J.: Overview of the face recognition grand challenge. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 947–954 (2005)