Efficient Feature Selection for Linear Discriminant Analysis and Its Application to Face Recognition

Zhen Lei   Shengcai Liao   Stan Z. Li

Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun Donglu, Beijing 100190, China.

Abstract

Feature selection is an important issue in pattern recognition. In face recognition, one of the state-of-the-art methods is that some feature selection methods (e.g., AdaBoost) are first utilized to select the most discriminative features and then the subspace learning methods (e.g., LDA) are further applied to learn the discriminant subspace for classification. However, in these methods, the objective of feature selection and subspace learning is not so consistent and the combination is not the optimal. In this paper, we propose a novel and efficient feature selection method that is designed for linear discriminant analysis (LDA). We use the Fisher criterion to select the most discriminative and appropriate features so that the objectives of feature selection and classifier learning are consistent (both follow the Fisher criterion) and the face recognition performance is expected to be improved. Experiments on FRGC v2.0 face database validate the efficacy of the proposed method.

1 Introduction

Feature selection is an important issue in many pattern recognition problems. Due to huge feature dimensions but limited data, the classifier learning in the entire feature space is usually infeasible and inaccurate, which is known as the "curse of dimensionality". Feature selection aims to select the most informative features from the original large feature pool so that the desired information from the original feature set is mostly preserved while the redundancy is reduced.

There have been lots of work on feature selection [8, 6, 4]. This paper mainly focuses on feature selection methods in face recognition area. In face recognition, the framework of selecting informative feature for subspace learning has achieved great success [10, 9, 12]. One of the representative methods of subspace learning is linear discriminant analysis (LDA) [2]. LDA is preferred in face recognition because it is capable of deriving discriminant subspace from large-scale training data for classification. LDA seeks such a subspace where samples from the same class are gathered while samples from different classes are separated, so that the samples are well classified.

As a preprocessing step of LDA, feature selection plays an important role to select the most informative and complementary features for LDA learning. Li et al. [10] utilize AdaBoost learning to select the most discriminative and complementary features for NIR face recognition. The weak classifier in AdaBoost learning is constructed based on a single feature, therefore, the strong classifier learning with AdaBoost is also a feature selection process. After that, LDA is applied to derive the discriminant subspace where the NIR faces are classified. In [14, 9], researchers use conditional mutual information (CMI) to select the effective features which are most relevant to the class labels and non-redundant. In [15], authors propose joint AdaBoost algorithm to select sharing features among different classes for subsequent subspace learning. Both AdaBoost and CMI based feature selection methods achieve promising results in face recognition combined with subspace learning methods like LDA.

There are at least two shortcomings in these existing methods. First, in AdaBoost and CMI, the training samples are usually firstly transformed into a binary class problem by computing the intra- and inter- personal spaces and then the AdaBoost learning or CMI method is applied to select the features. This transformation usually leads to huge increase of training samples. As a result, it is difficult to use all of the training samples in feature selection. Second, although the AdaBoost and CMI feature selection methods are effective ones, they may not be the optimal one for LDA since the objective of these feature selection methods are not generally consistent with LDA. There are also existing work using Fisher score [3] or Laplacian score [7] for feature selection. However, in these methods, the features are selected independently and there is large redundancy among the selected features which affects the efficiency of the consequent classification. In this work, we propose a novel feature selection method that is designed for linear discriminant analysis. The objective of the proposed feature selection method is somewhat consistent with the LDA (both follow the Fisher criterion) and therefore the combination of the proposed feature selection method and LDA is expected to achieve higher classification accuracy. Moreover, in our implementation, we don't need to transform the training samples to a binary classification problem as adopted in AdaBoost or CMI and hence more training samples can be involved in feature selection process to select the informative features.

The remainder of this paper is organized as follows. Section 2 details the principle and procedure of Fisher criterion based feature selection. Section 3 compares the proposed feature selection method with AdaBoost and CMI based ones on FRGC v2.0 face database and in Section 4, we conclude the paper.

2 Fisher Separation Criterion based Feature Selection

The purpose of LDA is to find a subspace which gathers the samples from the same class and meanwhile enlarges the margin of samples from different classes. Mathematically, this objective can be achieved by maximizing the Fisher criterion (the ratio of the between class scatter to the within class scatter). Given the sample set from the k-th class $X^k = \{X_1^k, X_2^k, \cdots, X_{N_k}^k\}$, where N_k is the number of samples in the k-th class, the between class scatter and the within class scatter are computed as

$$\begin{aligned} S_b &= \frac{1}{N}\sum_{i=1}^{C} N_i(M^i - M)(M^i - M)^T \\ S_w &= \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}(X_j^i - M^i)(X_j^i - M^i)^T \end{aligned} \tag{1}$$

where $N = \sum_i^C N_i$ is the total number of samples; C is the number of classes; $M^i = \frac{1}{N_i}\sum_i^{N_i} X_i$ is the mean vector of class i and $M = \frac{1}{N}\sum_i^C \sum_{j=1}^{N_i} X_j^i$ is the mean vector over the whole sample set. The purpose of LDA is to find the projective vector W that maximizes the following Fisher separation criterion J:

$$J = \frac{|W^T S_b W|}{|W^T S_w W|} \tag{2}$$

The solution of W can be obtained by solving the generalized eigen problem of $S_b W = \lambda S_w W$ with its leading eigenvalues.

Considering the feature selection problem, suppose the original feature set to be $\{f_1, f_2, \cdots, f_n\}$, where n is the dimension of the original feature set, the purpose of feature selection is to select d features $F^d = \{f_{v(1)}, f_{v(2)}, \cdots, f_{v(d)}\}$ from n original features which have the largest Fisher separation value. Here $v(i)$ is the i-th feature index in the selected feature subset. Denoting the between class scatter and the within class scatter computed based on the selected feature set F^d as $S_b(F^d), S_w(F^d)$, and the Fisher separation criterion as $J(F^d)$, the objective of selecting d features based on the Fisher separation criterion can be formulated as

$$F^d = \arg\max_{F^d} J(F^d) \tag{3}$$

where $J(F^d) = J(f_{v(1)}, f_{v(2)}, \cdots, f_{v(d)})$ is defined as

$$J(F^d) = \frac{|W^T S_b(F^d) W|}{|W^T S_w(F^d) W|} \tag{4}$$

Directly selecting d features from n feature candidates ($d << n$) is an NP-hard problem. A sub-optimal way is to sequentially select the features in a greedy way. Suppose we have selected k features, at the next step, we select the $(k+1)$-th feature that maximizes the Fisher criterion as

$$f_{k+1} = \arg\max_{f} J(F^k, f) \tag{5}$$

However, this is still probably infeasible in practice because the inverse operation of high-dimensional matrix is computational expensive. Suppose there are in total n features in candidate, and the number of samples is m, at the $(k+1)$-th step, we have to take at least $(n-k)$ matrix inversion operations with the dimension of $(k+1)$ to determine the next optimal feature. According to [5], the complexity of each matrix inversion operation with $(k+1)$ feature dimension is about $O((k+1)^3)$. Therefore, the computational cost at each iteration of feature selection is about $O(n(k+1)^3)$, where n is supposed to be much larger than k. With the increase of k, the computational cost increases cubically and it becomes intractable in practice. In this paper, we further simplify the feature selection process and only involve two variables in Fisher separation computation to reduce the computational cost.

In this method, we use the min-max criterion for feature selection. At each step, given a candidate feature f, we compute a series of Fisher separation improvement value $\Delta J(f|f')$ with one selected feature f' in turn. The Fisher separation improvement value $\Delta J(f|f')$ is defined as follows.

$$\Delta J(f|f') = J(f, f') - J(f') \tag{6}$$

The feature whose minimal Fisher separation improvement value in the selected feature set is the largest among all candidates is preferred as a new one. Besides the Fisher separation criterion, we also hope the new selected feature is a good complementary one to the existing ones. That is, the redundancy among the selected features are expected to be small. In this work, we use the correlation to measure the redundancy between different features. Grouping the feature values of feature i and j over the whole sample set as $f^i = [f_1^i, \cdots, f_n^i]$ and $f^j = [f_1^j, \cdots, f_N^j]$, where N is the number of samples, the correlation between f^i and f^j are computed as

$$\rho(f^i, f^j) = \frac{\sum_{k=1}^{N}(f_k^i - m(f^i))(f_k^j - m(f^j))}{\sqrt{\sum_{k=1}^{N}(f_k^i - m(f^i))^2 \sum_{k=1}^{N}(f_k^j - m(f^j))^2}} \tag{7}$$

where $m(f^i)$ and $m(f^j)$ are the mean values of f^i and f^j, respectively. The larger the correlation value ρ is, more redundancy exists between the features. Therefore, the correlation among the new selected feature and the selected features is required to be small. The final feature selection criterion consists of two terms (Eq. 8). One is the Fisher separation criterion evaluating the discriminant characteristic of features and the other is the correlation term which measures the redundancy of features. In our implementation, by storing the Fisher separation values between candidate features and selected ones, at each step, for every candidate feature, we only need to compute one Fisher separation between the candidate feature and the last selected feature, whose complexity is about $O(2^3 n) = O(8n)$ which is irrelative to k.

$$f_{k+1} = \arg \max_{f \in F \setminus F^k} \{ \min_{f' \in F^k} \Delta J(f|f') - \lambda \max_{f' \in F^k} \rho(f, f') \} \tag{8}$$

3 Experiments

We use FRGC v2.0 face database [13] to compare different feature selection methods on face recognition problem. FRGC ver 2.0 was collected by the University of Notre Dame. The training set consists of 12776 face images from 222 individuals, including 6360 controlled images and 6416 uncontrolled ones. In this experiment, we follow the experiment 4 protocol, which is considered the most difficult case in this database, to evaluate various methods. In the test set, there are 16028 controlled images from 466 persons as the target ones. The query set contains 8014 uncontrolled images. All the images are rotated, scaled and cropped to 142 × 120 according to the provided eye positions.

Two face representations (MBLBP [11] and MLPQ [1]) are utilized in our experiment. After MBLBP or MLPQ filtering, the histogram features are extracted. The feature selection method is applied to select the most discriminant and complementary features from the original feature pool. In classification phase, LDA is adopted based on the selected features and the cosine distance in the reduced subspace is used to measure the dissimilarity between different samples.

For MBLBP and MLPQ, 10 scales are used to extract the original feature set. There are in total $6,842,880$ features for MBLBP or MLPQ. For AdaBoost and CMI feature selection methods, the face images are first converted into intra- and inter-personal pairs respectively. For all methods, we finally select 3000 features from the MBLBP or MLPQ feature pool for the subsequent LDA learning.

Table 1. Verification rates (@FAR=0.001) of different methods on FRGC ver 2.0 database.

Method	Experiment 4		
	ROC I	ROC II	ROC III
BEE Baseline [13]	16.08%	15.18%	14.01%
MBLBP+AdaBoost	74.39%	75.02%	75.61%
MBLBP+CMI	72.21%	72.92%	73.37%
MBLBP+FS	78.71%	79.24%	79.74%
MLPQ+AdaBoost	79.72%	79.78%	79.58%
MLPQ+CMI	80.88%	81.14%	81.32%
MLPQ+FS	83.12%	83.28%	83.22%

Table 1 shows the face recognition results of different methods on FRGC ver 2.0 database following experiment 4 protocol. We also list the BEE baseline performance for comparison. It is easy to see that the proposed Fisher separation based feature selection method improves the face recognition performance by about $5 - 10$ percent compared to AdaBoost and CMI based ones, validating that Fisher separation, which is consistent with the objective of LDA, is an effective and appropriate feature selection criterion for LDA learning. Moreover, comparing the baseline performance with other results, one can find that the "local feature + feature selection + subspace learning" approach is indeed an effective framework for face recognition.

In our method, λ is a trade-off parameter between discrimination and information redundancy. To study the effect of the value of λ, we plot the verification rates (@FAR=0.001) of ROC III with different values of λ in Fig. 1 using MBLBP representation. For comparison, we also plot the results of AdaBoost and CMI based feature selection methods. It shows that the proposed method always outperforms AdaBoost and CMI based methods when the value of λ falls into the range of 0 to 2 and the performance is relatively good and stable when λ is selected between 0.5 and 1.5, indicating that the proposed feature selection method is effective and robust to the value of λ. In our experiment, we finally set λ to 1.0 for its best recognition performance.

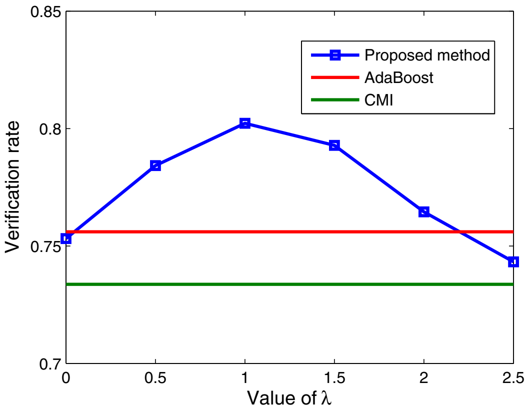

Figure 1. The effect of λ on the face recognition performance.

4 Conclusions

This paper proposes a novel feature selection method based on Fisher criterion for face recognition. Observing that the objective of feature selection is not so consistent with the classifier learning in traditional face recognition methods, we try to explore a feature selection method that is consistent with the classification method to improve the face recognition performance. The simplified feature selection method based on Fisher criterion is proposed and comparison experimental results show that the proposed feature selection method is more appropriate than AdaBoost and CMI based feature selection methods for LDA learning. This work provides an improvement for the state-of-the-art framework of "local feature + feature selection + subspace learning" for face recognition.

Acknowledgement

This work was supported by the Chinese National Natural Science Foundation Project #61070146, #61105023, #61103156, #61105037, National IoT R&D Project #2150510, Chinese Academy of Sciences Project No. KGZD-EW-102-2, European Union FP7 Project #257289 (TABULA RASA http://www.tabularasa-euproject.org), and AuthenMetric R&D Funds.

References

[1] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkilä. Recognition of blurred faces using local phase quantization. In *ICPR*, 2008. 3

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE-TPAMI*, 19(7):711–720, July 1997. 1

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification, Second Edition*. John Wiley and Sons, 2001. 2

[4] F. Fleuret. Fast binary feature selection with conditional mutual information. *JMLR*, 5:1531–1555, 2004. 1

[5] G. H. Golub and C. F. van Van Loan. *Matrix Computations, Third Edition*. The Johns Hopkins University Press, 1996. 2

[6] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3:1157–1182, 2003. 1

[7] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *NIPS*, 2005. 2

[8] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE T-PAMI*, 19:153–158, 1997. 1

[9] Z. Lei, S. Liao, M. Pietikäinen, and S. Z. Li. Face recognition by exploring information jointly in space, scale and orientation. *IEEE T-IP*, 20(1):247–256, Jan. 2011. 1

[10] S. Z. Li, R. Chu, S. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *IEEE T-PAMI*, 29(4):627–639, April 2007. 1

[11] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li. Learning multiscale block local binary patterns for face recognition. In *ICB*, pages 828–837, 2007. 3

[12] C. Liu. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE T-PAMI*, 28(5):725–737, 2006. 1

[13] 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. In *CVPR*, 2005. 3

[14] L. Shen and L. Bai. Information theory for gabor feature selection for face recognition. *EURASIP J. Adv. Sig. Proc.*, 2006. 1

[15] R. Xiao, W.-J. Li, Y. Tian, and X. Tang. Joint boosting feature selection for robust face recognition. In *CVPR*, pages 1415–1422, 2006. 1