

Bagging Based Efficient Kernel Fisher Discriminant Analysis for Face Recognition

Yi Li¹, Baochang Zhang², Shiguang Shan¹, Xilin Chen¹, Wen Gao^{1,2}

¹ICT-ISVISION JDL, Institute of Computing Technology, Chinese Academy of Sciences, China

²College of Computer, Harbin Institute of Technology, Harbin, China

{yli, bczhang, sgshan, xlchen, wgao}@jdl.ac.cn

Abstract

Kernel Fisher Discriminant Analysis (KFDA) has achieved great success in pattern recognition recently. However, the training process of KFDA is too time consuming (even intractable) for a large training set, because, for a training set with n examples, both its between-class and within-class scatter matrices are of $n \times n$ and the time complexity of the KFDA training process is of $O(n^3)$. Aiming at this problem, this paper employs Bagging technique to decrease the time-space cost of KFDA training process. In addition, this paper is more than just a simple application of Bagging. We have made an important adaptation which can further guarantee the performance of KFDA. Our experimental results demonstrate that the proposed method can not only greatly reduce the cost of time of the training process, but also achieve higher recognition accuracy than traditional KFDA and the simple application of Bagging.

1. Introduction

In recent years, the nonlinear feature extraction methods, such as Kernel Principal Component Analysis (KPCA) and Kernel Fisher Discriminant Analysis (KFDA) have been of wide concern. KPCA was originally developed by Scholkopf [1], and KFDA was subsequently proposed by Mika [2]. As a method of nonlinear mapping, kernel-based techniques project the input data into a high dimensional implicit feature space by nonlinear mapping $\phi: x \in R^N \rightarrow \phi(x) \in F$, with a nonlinear dot product kernel function $k(x, y) = (\phi(x) \cdot \phi(y))$. Though KFDA can extract high dimensional features, and performs well for discrimination problems, the training process is generally a computationally expensive task that becomes impractical for large set sizes. The reason is that, with n training examples, the dimensionality of its between-class and within-class scatter matrices is $n \times n$. And the time complexity of the training process is as high as $O(n^3)$. Therefore, unlike the training process of Linear Discriminant Analysis (LDA), a training database with big size will cause serious calculating problems for KFDA [3]. In [3], Liu proposed a method to solve this problem by using kernel trick to select an optimized subset from data and form a subspace of the feature space, however, it is a little complex and not very effective for

reducing time consuming or enhancing recognition accuracy according to the reported results.

Aiming at this problem, this paper employs Bagging technique to decrease the time-space cost of KFDA training process. Since the complexity of KFDA is exponent of the training set size, it is easy to understand that Bagging can decrease the complexity of each component KFDA classifier which is trained on a sub-set of the whole training set. In addition, this paper is more than just a simple application of Bagging. We have made an important adaptation which can further guarantee the performance of KFDA. Different from the traditional bagging method, our idea is that all the samples contribute to the construction of the scatter matrices in the original space R^N , but only a part of them are used to build the optimized discriminant vectors $w^{(k)}$ (k denotes the k^{th} sub-set of the whole training set) of the kernel discriminant feature subspaces $F^{(k)}$. Thus, we try to solve this problem as follows: first construct $w^{(k)}$ of $F^{(k)}$ by using bagging method. Then the whole training set is mapped into $F^{(k)}$ to calculate the scatter matrices, based on which nonlinear discriminant features are extracted.

The performance of the proposed method is evaluated on the FERET and CAS-PEAL databases, and our experimental results demonstrate that the proposed method can not only greatly reduce the cost of time and space of the training process, but also achieve higher recognition accuracy than traditional KFDA and the simple application of Bagging.

2. Related Work

Before describing the proposed method, in this section, we briefly review related work on classifier design by using bagging method and nonlinear discriminant analysis of KFDA.

2.1. Bagging Method

Bagging [4] strategy incorporates the benefits of bootstrapping and aggregation. Multiple classifiers can be generated by training on multiple sets of samples that are produced by bootstrapping, i.e. random sampling with replacement on the training samples. Aggregation of the generated classifiers can then be implemented by majority voting rule (MVR) and Sum Rule [5].

2.2. KFDA

Let x be a vector of the input set with n elements and C classes, and n_i represents the number of samples in the i -th class. The mapping of x_i is noted as $\phi_i = \phi(x_i)$. Performing FLDA in F means to maximize the following Fisher discriminant function:

$$J(w) = \arg \max_w \frac{|w^T S_B^\phi w|}{|w^T S_W^\phi w|} \quad (3)$$

where S_B^ϕ and S_W^ϕ represent the between-class scatter and within-class scatter respectively in F .

$$S_B^\phi = \sum_{i=1}^C (u_i - \bar{u})(u_i - \bar{u})^T, \quad (4)$$

$$S_W^\phi = \sum_{i=1}^C \frac{1}{n_i} \sum_{j=1}^{n_i} (\phi_j - u_i)(\phi_j - u_i)^T$$

here, $u_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \phi_j$, $\bar{u} = \frac{1}{n} \sum_{i=1}^C \phi_i$. Because $w \in F$ must lie in the span of all the samples in F , w can be represented by a linear combination of ϕ_i , i.e.,

$$w = \sum_{i=1}^n \alpha_i \phi_i \quad (5)$$

Thus, equation (3) can be rewritten as:

$$J(\alpha) = \arg \max_\alpha \frac{|\alpha^T K_B \alpha|}{|\alpha^T K_W \alpha|} \quad (6)$$

and the problem of KFDA is converted into finding the leading eigenvectors of $K_w^{-1} K_B$.

$$K_B = \sum_{i=1}^C (m_i - \bar{m})(m_i - \bar{m})^T, \quad (7)$$

$$K_W = \sum_{i=1}^C \frac{1}{n_i} \sum_{j=1}^{n_i} (\zeta_j - m_i)(\zeta_j - m_i)^T$$

where $\zeta_j = (k(x_1, x_j), k(x_2, x_j), \dots, k(x_n, x_j))^T$,

$$m_i = \left(\frac{1}{n_i} \sum_{j=1}^{n_i} k(x_1, x_j), \frac{1}{n_i} \sum_{j=1}^{n_i} k(x_2, x_j), \dots, \frac{1}{n_i} \sum_{j=1}^{n_i} k(x_n, x_j) \right)^T$$

and \bar{m} is the mean of all the ζ_j .

For a new point x , its projection onto w in F can be calculated by

$$(w \cdot \phi(x)) = \sum_{i=1}^n \alpha_i k(x_i, x) \quad (8)$$

3. BKFD

Though KFDA can extract high dimensional features, and performs well for discrimination problems, the training process is generally a computationally expensive task that becomes impractical for large set sizes. In this section, we present a Bagging Based Efficient Kernel Fisher Discriminant Analysis (BKFD) to overcome this problem.

3.1. Bagging Based Efficient Kernel Fisher Discriminant Analysis

Traditional Bagging method divides the training database into several subsets and trains a classifier (or make discriminant) for each subset. Then, two strategies are applied to make the final classification, namely the simple majority voting and the sum rule [6, 7]. However, different from the traditional bagging method, our idea is that all the samples contribute to the construction of the scatter matrices in the original space R^N , but only a part of them are used to build the optimized discriminant vectors $w^{(k)}$ (k denotes the k^{th} sub-set of the whole training set) of the kernel discriminant feature subspaces $F^{(k)}$. Let $x^{(k)}$ be a vector of the input subset made up of randomly selected samples from training database with $n^{(k)}$ elements and $C^{(k)}$ classes, and $n_i^{(k)}$ represents the number of samples in the $i^{(k)}$ -th class. The mapping of $x_i^{(k)}$ is noted as $\phi_i^{(k)} = \phi(x_i^{(k)})$. Thus, we try to solve this problem as follows: first construct $w^{(k)}$ of $F^{(k)}$ by using bagging method to replace the expression (5) as:

$$w^{(k)} = \sum_{i=1}^{n^{(k)}} \alpha_i^{(k)} \phi_i^{(k)} \quad (9)$$

where $w^{(k)} \in F^{(k)}$, $\phi_i^{(k)}$ is the element of $F^{(k)}$, and there

exists $F = \bigcup_{k=1}^L F^{(k)}$, where L is the total number of

$F^{(k)}$. Then the whole training set is mapped into $F^{(k)}$ to calculate the scatter matrices, based on which nonlinear discriminant features are extracted. Thus, equation (4) is not changed, while equation (6) is rewritten as:

$$J^{(k)}(\alpha^{(k)}) = \arg \max_{\alpha^{(k)}} \frac{|\alpha^{(k)T} K_B^{(k)} \alpha^{(k)}|}{|\alpha^{(k)T} K_W^{(k)} \alpha^{(k)}|} \quad (10)$$

where

$$K_B^{(k)} = \sum_{i=1}^{n^{(k)}} (m_i^{(k)} - \bar{m}^{(k)})(m_i^{(k)} - \bar{m}^{(k)})^T, \quad (11)$$

$$K_W^{(k)} = \sum_{i=1}^{n^{(k)}} \frac{1}{n_i^{(k)}} \sum_{j=1}^{n_i^{(k)}} (\zeta_j^{(k)} - m_i^{(k)})(\zeta_j^{(k)} - m_i^{(k)})^T$$

and $\zeta_j^{(k)} = (k(x_1, x_j), k(x_2, x_j), \dots, k(x_{n^{(k)}}, x_j))^T$

$$m_i^{(k)} = \left(\frac{1}{n_i^{(k)}} \sum_{j=1}^{n_i^{(k)}} k(x_1, x_j), \frac{1}{n_i^{(k)}} \sum_{j=1}^{n_i^{(k)}} k(x_2, x_j), \dots, \frac{1}{n_i^{(k)}} \sum_{j=1}^{n_i^{(k)}} k(x_{n^{(k)}}, x_j) \right)^T$$

and $\bar{m}^{(k)}$ is the mean of all the $\zeta_j^{(k)}$.

For a new point x , its projection onto $w^{(k)}$ in $F^{(k)}$ can be calculated by

$$(w^{(k)} \cdot \phi^{(k)}(x)) = \sum_{i=1}^{n^{(k)}} \alpha_i^{(k)} k(x_i^{(k)}, x) \quad (12)$$

3.2. Complexities of BKFD

It is easy for us to know that the complexity of the calculation of scatter matrix is decided by the size of

subset($n^{(k)}$), $O((n^{(k)})^2 \cdot n)$. The complexity of the kernel mapping is about $O(n^{(k)} \cdot n \cdot S^2)$, where S is the size of the input vector. Obviously, the size of the subset greatly affects the complexity of the training procedure of kernel methods.

3.3. Similarity Measure for BKFD

When an image is presented to the proposed method, the augmented Gabor feature vector of the image is first calculated and the lower dimensional feature is derived by using PCA as y . The new feature vector $v^{(k)}$ of the image is defined as follows:

$$v^{(k)} = (w^{(k)} \cdot \phi^{(k)}(y)) = \sum_{i=1}^{p^{(k)}} \alpha_i^{(k)} k(y_i^{(k)}, y) \quad (13)$$

Given that $v_1^{(k)}, v_2^{(k)}$ are the extracted feature vectors corresponding to two face images x_1, x_2 . The similarity rule is based on the cross correlation between corresponding vectors.

$$d^{(k)}(x_1, x_2) = \frac{v_1^{(k)} \cdot v_2^{(k)}}{\|v_1^{(k)}\| \cdot \|v_2^{(k)}\|} \quad (14)$$

If the number of $F^{(k)}$ (also called mixture model) is L , and then the sum rule is used as follows:

$$d(x_1, x_2) = \sum_{k=1}^L d^{(k)}(x_1, x_2) \quad (15)$$

Experiments are performed on two databases, the FERET and the CAS-PEAL databases. In our paper, the kernel function is polynomial style, $k(x, y) = (\frac{x \cdot y}{c})^c$, where c is a constant which is related to the length of the input vector.

4. Experiments

In our experiments, the face image is cropped to the size of 64X64 and is overlapped with a mask to eliminate the background and hair after normalization of Histogram Equalization. Gabor features are down-sampled by factor 0.5, so the final dimension of the Gabor feature is 40960.

4.1. Experiments on FERET Database

We have tested the proposed method on the standard FERET database [8], which has been widely used to evaluate face recognition algorithms. In our experiments, the training set contains about 1002 face images from the standard FERET training CD. The probe database Fb, a set with 1195 images, is used to evaluate the proposed method, because it is the largest test set in the standard FERET database.

G-GKFD is Gabor feature based GKFD method, which is based on all the training samples. Performance of the G-BKFD (single-model, the size of subset is 500) is first

evaluation, though it is a little worse than G-GKFD, the training time is greatly reduced (1/4). From Table.1, we know that G-BKFD (multi-model based) achieves slightly better performance with 2 mixture models (2(500), the size of subset is 500) than G-GKFD with 1 model(1(1002), the size of subset is 1002), but the training time of it is greatly reduced(about half of the G-GKFD). Accordingly, 3(500) or 4(500) means 3 or 4 mixture models with size of the subset as 500. When the size of the subset is reduced by 50%, the consuming time will be saved by about 75% of the training time. In Table.1, the recognition rate of G-BKFD with two models has slightly increased compared to G-GKFD with 1 model; however, the training time has been reduced by 50%. In Fig.1, we find that the proposed application of bagging strategy is much better than the traditional one with the same divided subsets. Partly because each bagging subset didn't contain enough samples to capture the variance among the whole training database.

Table 1. Comparison experiments on fb probe database

	G-BKFD	G-GKFD	G-BKFD		
N	1(500)	1(1002)	2(500)	3(500)	4(500)
R	94.1(Aver.)	95.3	95.5	95.6	95.7
T	17m	70m	2*17m	3*17m	4*17m

N refers to the number of mixture models, R refers to recognition accuracy, and T refers to training time consuming

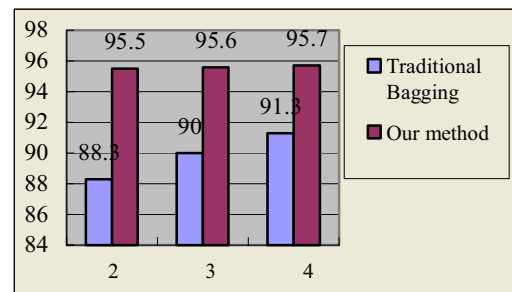


Fig 1. Comparison experiments of traditional bagging and the proposed application of bagging method for fb probe database with standard FERET training database

4.2. Experiments on CAS-PEAL Database

The CAS-PEAL face database was constructed with the sponsoring of the National Hi-Tech Program and ISVISION[9] of China. We use its released version, CAS-PEAL-R1, for evaluation according to its accompanying standard evaluation protocol. To this protocol, the training set contains 1200 images of 300 subjects, and the testing set contain 9060 face images of 1040 subjects, which is partitioned into the Gallery and 7

probe sets. One image per subject is in the Gallery, and all the others are used as probes. Details of the face database are shown at <http://www.jdl.ac.cn/peal/index.html>. In our experiments, we choose two large probe sets, Expression (with 1884 images from the 1040 subjects) and Accessory, (with 2616 images from the 1040 subjects) to test the proposed method.

Table 2. Comparison experiments on the CAS-PEAL-R1 database

N	G-GKFD	G-BKFD		
	1(1200)	2(600)	3(600)	4(600)
Expression	91.7	92.2	92.3	92.8
Accessory	76.3	78.2	78.9	79.1

4.3. Experiments on FERET plus CAS-PEAL-R1

In this part, we combine the FERET and CAS-PEAL databases to test the performance of G-BKFD. Thus, the training database contains 2202(1002+1200) face images. Gallery and probe sets have not been changed. From Table.3, we find that G-BKFD is slightly better than G-GKFD in three large probe databases. Specially, we make further comparison experiments between our method and the traditional bagging method, from Fig.2, we found that the proposed method had also achieve better performance. For the traditional bagging method, we divide the training database into the same subsets as our method, that is, size of which is 1002, 1200, 1101, 1101 accordingly, and 1101 images contain half of the FERET training database images, and the other is from half of the CAS-PEAL-R1 training database.

Table 3. Recognition rate for FERET + CAS-PEAL database

N	G-GKFD	G-BKFD		
	1(2202)	2(1002 /1200)	3(1002/1200 /1101)	4(1002/1200 /1101/1101)
R(F-fb)	95.6	95.8	95.9	96.0
R(C-expression)	92.3	92.7	92.9	93.1
R(C- accessory)	75.2	75.7	75.8	75.9

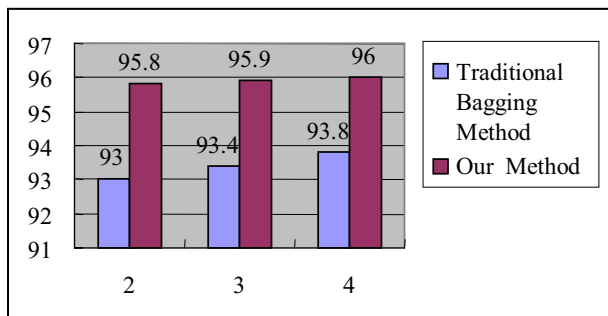


Fig2. Comparison experiments of traditional bagging and the proposed application of bagging method for fb probe database with hybrid training database (2

mixture models are 2(1002,1200), 3 mixture models are 3(1002,1200,1101), 4 mixture models are 4(1002,1200,1101,1101))

5. Conclusions

We have proposed in this paper BKFD method for face recognition, aiming at decrease the time-space cost in the traditional training process of KFDA. By introducing Bagging technique, the training set size for each component KFDA classifier can be controlled to tractable level, and thus the time-space complexity of the whole training process is reduced. At the same time, we also adapt the Bagging technique to the KFDA in order to further improve the performance. The performance of the proposed method is evaluated on two large-scale face database, the FERET and CAS-PEAL-R1, and the experimental results demonstrate that the proposed method can not only greatly reduce the time-space cost of the training process, but also achieve higher recognition accuracy than traditional KFDA and the direct Bagging method

Acknowledgement

This research is partially sponsored by Natural Science Foundation of China under contract No.60332010, "100 Talents Program" of CAS. The first author also would like to gratefully acknowledge support from Natural Science Foundation of China under contract No.60473043.

References

- [1] B.Scholkopf, A.Smola, K.R.Muller, "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation* 10(5) (1998) 1299-1319.
- [2] S.Mika, G.Ratsch, J.Weston, B.Scholkopf, K.R.Muller, "Fisher discriminant analysis with kernels", *IEEE International Workshop on Neural Networks for Signal Processing*, Bol.IX, Madison, USA, August, 1999, pp.41-48.
- [3] Qingshan Liu, RuiHuang, Hanqing Lu and Songde Ma, "Kernel-Based Optimized Feature Vectors Selection and Discriminant Analysis for Face Recognition", *ICPR2002*.
- [4] L. Breiman, "Bagging Predictors," *Int. J. on Machine Learning*, no. 24, pp 123-140, 1996.
- [5] J. Kittler, M. Hatef, P.W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. On PAMI*. Vol. 20, no. 3, pp. 226-239, Mar. 1998.
- [6] Xiaoguang Lu and Anil K.Jain, "Resampling for Face Recognition", *AVBPA2003*.
- [7] S.B.Kotsiantis, P.E.Pintelas, "Combining Bagging and Boosting", *International Journal of Computational Intelligence*, vol.1, 2004.
- [8] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, Patrick J. Rauss, *The FERET Evaluation Methodology for Face-Recognition Algorithms*, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.22, 2000, pp.1090-1104.
- [9] W.Gao, B. Cao, S. Shan, "The CAS-PEAL Large-Scale Face Database and Evaluation Protocols," *Technical Report No. JDL_TR_04_FR_001*, JDL, CAS, 2004.