Investigation of KLIM Algorithm Applied to Face Recognition

Yunfei Jiang, Rukun Hu, Ping Guo*, Xin Zheng
Laboratory of Image Processing and Pattern Recognition, Beijing Normal University, Beijing 100875, China
Email: yunfeifei_2001@163.com, hurukun@mail.bnu.edu.cn, pguo@ieee.org, zhengxin@bnu.edu.cn

***Abstract*—Face recognition often suffers from the Small Sample Size problem. Regularization is one of the solutions to this problem. In this paper, we investigate the Kullback-Leibler information measure (KLIM) based regularization classifiers for face recognition. Two parameter estimation approaches including the cross-validation technique and model selection criterion are chosen to optimize the regularization parameter. In the experiments, the ORL face data is used to evaluate these algorithms. We compared the KLIM algorithms with quadratic discriminant analysis, linear discriminant analysis, regularized discriminant analysis, and leave-one-out covariance matrix estimate. Considering both time cost and classification rate, KLIM classifiers exceed the others and obtain stable results.**

***Index Terms*—Gaussian classifier, Regularization, Cross validation, Principal component analysis, Face recognition.**

I. INTRODUCTION

Face recognition is a hot research topic in the fields of pattern recognition and computer vision, which has been widely used in many applications, such as verification of credit card, security access control, and human computer interface. As a result, numerous face recognition algorithms have been proposed, and surveys in this area can be found in [1], [2], [3]. Two central issues to an automatic face recognition system are: 1) feature selection for face representation and 2) classification of a new face image based on the chosen feature representation. In an actual face recognition system, the results of feature selection can be easily affected by some variations in the face images, such as lighting, expression and pose.

Generally, face recognition is a high-dimensional data set classification problem, it often suffers from the Small Sample Size problem (SSS). In SSS problem, usually the class sample size n_j is approximately equal to or smaller than the variable dimension d, the covariance estimation in discriminant analysis will become highly variable, in which case it becomes ill- or poorly-posed classification problems. A typical representative of approaches suffering this problem is Quadratic Discriminant Analysis (QDA) [4]. QDA is widely used if there are sufficient training samples. Unfortunately, sometimes training samples are usually hard to acquire, and the dimensionality of face data is extremely high, thus the estimated covariance matrix will become singular.

There exist two main solutions for SSS problem. One is to classify them directly in high-dimensional space with regularization methods, and the other is to reduce data dimension first, then classify them in feature space. Dimension reduction methods can be divided into two main categories [5]. One is based on local feature, which typically extracted a set of facial features from the image, such as eyes, nose etc, and used it to classify the face; the other is global or "holistic" approaches, which takes a holistic view of the recognition problem, and holistic feature extraction of face images is adopted in this approach. Fisher Linear Discriminant Analysis (F-LDA) [6] is one of the most popular feature extraction techniques in the second approach. F-LDA finds a set of the most discriminant projection vectors by maximizing the between-class scatter matrix (S_b) while minimizing the within-class scatter matrix (S_w) in the projective feature space. The major drawback of applying F-LDA is that when the number of training samples is smaller than that of their dimensionality, it can't solve the SSS problem. Under these circumstances, S_w becomes singular, and it results in the difficulty to calculate the F-LDA vectors. Now, many new approaches has been developed based on F-LDA technique, like Direct Linear Discriminant Analysis (D-LDA) [7], Regularized Linear Discriminant Analysis (R-LDA) [8], and Kernel Direct Linear Discriminant Analysis (KDDA) [9]. Like LDA, Principal Component Analysis (PCA) [10], [11], [12] is also a powerful tool in data dimension reduction as well as an effective feature extraction method in pattern recognition field. Its goal is to describe the pattern with the less quantities of feature, and to reduce dimensionality of the feature space without losing the most important, for discrimination purposes, information. Kernel Principal Component Analysis (KPCA) [13] is the kernel edition of PCA in nonlinear subspace.

Regularization is another solution to solve SSS problem. There are many regularized classification techniques [14] in this research field. Linear discriminant analysis (LDA) [15] could be used as one kind of regularization if the total number of samples is larger than the dimension of variables. The covariance matrix, in LDA, is substituted by common covariance matrix. However, in the case of small sample sizes, the common covariance matrix is also singular. Regularized discriminant analysis [16] (RDA) adds the identity matrix as a regularization term to solve the problem in matrix estimation, and leave-one-out covariance [17], [18] (LOOC) brings the diagonal matrix in, too. The regularization parameters (or called Model) in both methods are optimized by leave-one-out cross-validation [19] method, which cost a mass of computing time. Kullback-Leibler information measure [20] based classifier (KLIM1 and KLIM2) is the one that we derived to estimated the covariance matrix. It only contains one regularization parameter that can be determined with derived model selection criterion. KLIM has been successfully used in stellar spectra data classification [21], [22].

In this paper, we will investigate the performance of the KLIM algorithm in face recognition. To evaluate the performance under different feature dimensionality, we adopt PCA before classification. We compare the KLIM approach with QDA, LDA, RDA, and LOOC approaches on ORL face data. The parameter tuning, computing time and performance are detailed analyzed in our experiments.

II. Discriminant Analysis

Discriminant analysis is to assign an observation $\mathbf{x} \in \mathbb{R}^N$ with unknown class membership to one of k classes $C_1, ..., C_k$ known *a priori*. There is a learning data set $A = \{(\mathbf{x}_1, c_1), ..., (\mathbf{x}_n, c_n) | \mathbf{x}_j \in \mathbb{R}^N \text{ and } c_j \in \{1, ..., k\}\}$, where the vector $\mathbf{x}_j$ contains N explanatory variables and c_j indicates the index of the class of $\mathbf{x}_i$. The data set allows to construct a decision rule which associates a new vector $\mathbf{x} \in \mathbb{R}^N$ to one of the k classes. Bayes decision rule assigns the observation $\mathbf{x}$ to the class C_j^* which has the maximum a posteriori probability. Which is equivalent, in view of the Bayes rule, to minimize a cost function $d_j(\mathbf{x})$,

$$j^* = \arg\min_j d_j(\mathbf{x}), \qquad j = 1, 2, \cdots, k \tag{1}$$

$$d_j(\mathbf{x}) = -2\ln(\widehat{\alpha}_j f_j(\mathbf{x})). \tag{2}$$

Where $\widehat{\alpha}_j$ is the prior probability of class C_j and $f_j(\mathbf{x})$ denotes the class conditional density of $\mathbf{x}$, $\forall j = 1, ..., k$.

Assumes that the class conditional density $f_j(x)$ for the class C_j is Gaussian $\mathcal{N}(\mathbf{x}, \widehat{\mathbf{m}}_j, \widehat{\mathbf{\Sigma}}_j)$, which the mathematical expressions are as the followings,

$$\mathcal{N}(\mathbf{x}, \widehat{\mathbf{m}}_j, \widehat{\mathbf{\Sigma}}_j) = \frac{\exp[-\frac{1}{2}(\mathbf{x} - \widehat{\mathbf{m}}_j)^T \widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{x} - \widehat{\mathbf{m}}_j)]}{(2\pi)^{\frac{d}{2}} |\widehat{\mathbf{\Sigma}}_j|^{\frac{1}{2}}} \tag{3}$$

Substitute this expression in Eq. 2, then leads to the discriminant function,

$$d_j(\mathbf{x}) = (\mathbf{x} - \widehat{\mathbf{m}}_j)^T \widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{x} - \widehat{\mathbf{m}}_j) + \ln|\widehat{\mathbf{\Sigma}}_j| - 2\ln\widehat{\alpha}_j. \tag{4}$$

Where $\widehat{\mathbf{m}}_j$ is the mean vector, and $\widehat{\mathbf{\Sigma}}_j$ is the covariance matrix of the j-th class. If the *prior* probability $\widehat{\alpha}_j$ is the same for all classes, the term $2\ln\widehat{\alpha}_j$ can be omitted and the discriminant function reduces to a more simple form.

III. Classical Discriminant Classifiers

Some classical discriminant analysis methods can be obtained by combining additional assumptions with the Bayes decision rule, such as QDA, LDA, RDA, and LOOC. Among them, LDA, RDA, and LOOC are three regularization methods. The crucial difference of these methods is the diversity of the covariance matrix estimation formula. Before the introduction of KLIM approach, we will give a brief review of these methods.

A. QDA

Quadratic Discriminant Analysis (QDA) [4], [14] is widely used in pattern recognition problem. In QDA, the parameters in Eq. 4 can be estimated with traditional maximum likelihood estimator.

$$\widehat{\mathbf{m}}_j = \frac{1}{n_j}\sum\nolimits_{i=1}^{n_j} \mathbf{x}_i, \tag{5}$$

$$\widehat{\mathbf{\Sigma}}_j = \frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}(\mathbf{x}_i - \widehat{\mathbf{m}}_j)(\mathbf{x}_i - \widehat{\mathbf{m}}_j)^T. \tag{6}$$

where the $\mathbf{x}_i$ is a sample from class j with probability one, and n_j is the training sample number of class j. When using an unbiased estimation,

$$\widehat{\mathbf{\Sigma}}_j = \frac{1}{n_j - 1}\sum\nolimits_{i=1}^{n_j}(\mathbf{x}_i - \widehat{\mathbf{m}}_j)(\mathbf{x}_i - \widehat{\mathbf{m}}_j)^T. \tag{7}$$

this is called sample covariance matrix.

In practice, this method suffers SSS problem in high-dimensional spaces since it requires estimating many parameters. For small sample number case, it will lead to the ill-posed problem. In that case the parameter estimates can be highly unstable, giving rise to high variance in classification accuracy. By employing a method of regularization, one attempts to improve the estimates by biasing them away from their sample based values towards values that are deemed to be more "physically plausible".

B. LDA

Linear discriminant analysis (LDA) [15] is a regularization methods to deal with the poorly-posed problem. The LDA used in regularization is different from F-LDA in dimension reduction. The $\widehat{\mathbf{\Sigma}}_j$ in Eq. 6 is replaced with the following pooled covariance matrix, also called common covariance matrix,

$$\widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum\nolimits_{j=1}^{k} n_j \widehat{\mathbf{\Sigma}}_j. \tag{8}$$

C. RDA

RDA is a regularization method which was proposed by Friedman [16]. RDA is designed for small number samples case, where the covariance matrix in Eq.(4) takes the following form:

$$\mathbf{\Sigma}_j(\lambda, \gamma) = (1 - \gamma)\mathbf{\Sigma}_j(\lambda) + \gamma\frac{\text{Trace}[\mathbf{\Sigma}_j(\lambda)]}{d}\mathbf{I}_d \tag{9}$$

with

$$\mathbf{\Sigma}_j(\lambda) = \frac{(1 - \lambda)n_j\widehat{\mathbf{\Sigma}}_j + \lambda N\widehat{\mathbf{\Sigma}}}{(1 - \lambda)n_j + \lambda N}. \tag{10}$$

The two parameters λ and γ, which are restricted to the range 0 to 1, are regularization parameters to be selected according to maximizing the leave-one-out correct classification rate (CCR). λ controls the amount of the $\widehat{\mathbf{\Sigma}}_j$ that are shrunk towards $\widehat{\mathbf{\Sigma}}$, while γ controls the shrinkage of the eigenvalues towards equality as $\text{Trace}[\mathbf{\Sigma}_j(\lambda)]/d$ is equal to the average of the eigenvalues of $\mathbf{\Sigma}_j(\lambda)$.

Fig. 1. Examples of face images for two persons chosen from the ORL face database.

D. LOOC

There exists another covariance matrix estimation formula which was proposed by Hoffbeck and Landgrebe [17]. They examine the diagonal sample covariance matrix, the diagonal common covariance matrix, and some pair-wise mixtures of those matrices. The proposed estimator has the following form:

$$\mathbf{\Sigma}_j(\xi_j) = \xi_{j1} diag(\widehat{\mathbf{\Sigma}}_j) + \xi_{j2}\widehat{\mathbf{\Sigma}}_j + \xi_{j3}\widehat{\mathbf{\Sigma}} + \xi_{j4} diag(\widehat{\mathbf{\Sigma}}). \quad (11)$$

The elements of the mixing parameter $\xi_j = [\xi_{j1}, \xi_{j2}, \xi_{j3}, \xi_{j4}]^T$ are required to sum up to unity: $\Sigma_{l=1}^{4}\xi_{jl} = 1$. In order to reduce the computation cost, they only considered three cases: $(\xi_{j3}, \xi_{j4}) = 0$, $(\xi_{j1}, \xi_{j4}) = 0$, and $(\xi_{j1}, \xi_{j2}) = 0$. They called the covariance matrix estimator as LOOC because the mixture parameter ξ was optimized by leave-one-out cross-validation method.

IV. KLIM Based Classifiers

Toward the ill-posed problem in discriminant analysis and reduce the computation time in regularization parameter selection, we have developed a KLIM1 estimator [23] based on Kullback-Leibler information measure, and we assume that the class conditional density f_i for the class C_j is Gaussian $\mathcal{N}(\mathbf{x}, \widehat{\mathbf{m}}_j, \widehat{\mathbf{\Sigma}}_j)$. The matrix estimation formula of KLIM1 is shown in the following:

$$\mathbf{\Sigma}_j^{(1)}(h) = h\mathbf{I}_d + \widehat{\mathbf{\Sigma}}_j, \quad (12)$$

where h is a regularization parameter, $\mathbf{I}_d$ is a $d \times d$ dimensional identity matrix.

This class of formula can solve matrix singular problem in high-dimension setting. In fact, as long as h is not too small, $\mathbf{\Sigma}_j^{-1}(h)$ exists with a finite value and the estimated classification rate will be stable.

We also derived a KLIM2 estimator based on KLMI1, which is computed as follows:

$$\begin{aligned}\mathbf{\Sigma}_j^{(2)}(h) \;=\; & (1 + \frac{h}{2}\text{Trace}[\widehat{\mathbf{\Sigma}}_j + \widehat{\mathbf{\Sigma}}])h\mathbf{I}_d + \\ & \frac{h^2}{1+\eta} ev(\widehat{\mathbf{\Sigma}}_j + \widehat{\mathbf{\Sigma}}) + \frac{\widehat{\mathbf{\Sigma}}_Q}{(1+\eta)}, \quad (13)\end{aligned}$$

where $ev(\mathbf{\Sigma})$ stands for a diagonal matrix in which the diagonal elements are the eigenvalues of $\mathbf{\Sigma}$. The following notations are used in the above equations,

$$\eta = \frac{h}{2n_j}\text{Trace}[\mathbf{H}(j)] \quad (14)$$

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_Q \;=\; & \widehat{\mathbf{\Sigma}}_j + \\ & \frac{h}{2n_j}\sum_{i=1}^{n_j}[\text{Trace}[\mathbf{H}(j)]](\mathbf{x}_i - \widehat{\mathbf{m}}_j)(\mathbf{x}_i - \widehat{\mathbf{m}}_j)^T (15)\end{aligned}$$

where the Hessian matrix:

$$\mathbf{H}(j) = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{H}(j), \quad (16)$$

and

$$\begin{aligned}\mathbf{H}_i(j) \;=\; & (\mathbf{x}_i - \widehat{\mathbf{m}}_j)(\mathbf{x}_i - \widehat{\mathbf{m}}_j)^T + \\ & \sum_{l=1}^{k}\widehat{\alpha}_l(\mathbf{x}_i - \widehat{\mathbf{m}}_j)(\mathbf{x}_i - \widehat{\mathbf{m}}_j)^T \quad (17)\end{aligned}$$

For estimating the regularization parameter with cross-validation is time consuming, we derived a model selection criterion to compute the parameter h in [22]:

$$h = \frac{1}{dN^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (18)$$

V. Experiments

In our experiments, in order to investigate the performance of KLIM classifiers with different feature dimensionality, we first adopt PCA to reduce data dimensionality. The inputs of the classifiers are the selected features, and we use two different methods (the cross-validation technique and model selection criterion) to optimize the regularization parameters. Also we compare the KLIM algorithm with QDA, LDA, RDA, and LOOC algorithms in the classification of ORL face database. Correct classification rate (CCR) and time cost are used to evaluate the performance for each classifiers.

A. Database

All the experiments were conducted on the ORL face database, which is a popular database in face recognition research. This database consists of 40 persons, with each person's face appearing in 10 images, and comprises 400 images altogether. The images are taken at different time instances, with different lighting conditions, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the persons in an upright, frontal position, with tolerance for some movement. All of

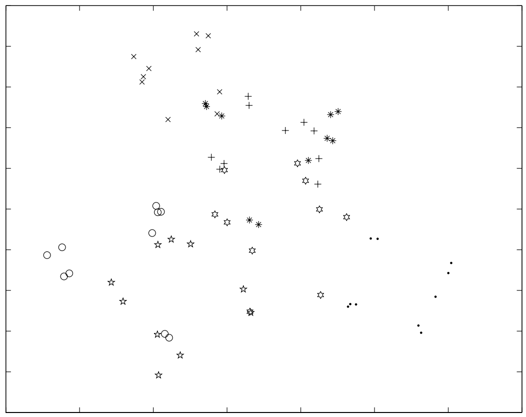

Fig. 2. Two dimensional projection of the first group data by PCA.

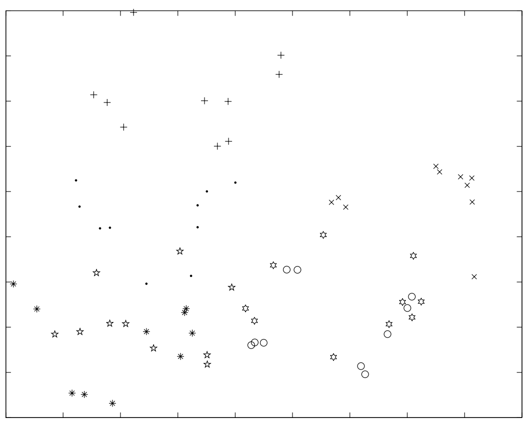

Fig. 3. Two dimensional projection of the second group data by PCA.

the images are 112×92 in size. Fig. 1 shows examples of the face images for two persons taken at different time instances.

To implement the face recognition experiment, we design two group of experiments. In every group, we randomly select 7 persons from the ORL database, and the ORL database is randomly partitioned into a training set and a testing set with no overlap between them. In each experiment, four images per person are randomly drawn for test to measure the classification accuracy, the remain six images are used to consist the training set and used to estimate the mean and covariance matrices. Thus, a training set of 42 images and a test set of 28 images are created. To show the separability of the two group data, we reduce the dimensionality to $d = 2$ with PCA, as shown in Fig. 2 and Fig. 3, from which we can see that some data overlaps from different classes after dimension reduction with PCA. The ill-posed problem will occur without feature dimension, and separability of the face data is depressed since dimension reduction loses discriminant information, so we can introduce regularization approaches to deal with above problem.

B. Regularization Parameter Estimation

The regularization parameter in KLIM is determined by statistical cross-validation technique (CV) or calculation by model selection criterion (MSC). For the parameter estimation with cross-validation, 20 h values are chosen geometry proportionately between 0 to 1. Cross-validation is also used in RDA and LOOC for parameter estimation. In RDA, the values of both λ and γ are sampled over a very coarse grid, (0.0, 0.2, 0.4, 0.6, 0.8, 1.0), resulting in 36 data points. In LOOC, the four parameters are picked according to the table in [17].

C. Experiment with different feature dimensions

We perform PCA primarily to obtain different feature dimensionality. For the convenience of comparison, we reduce the dimensionality of face data to seven different levels (40, 30, 25, 20, 10, 5, and 2). In experiments, we noted the correct classification rate (CCR) and time cost for each classifiers. Twenty five runs of each experiment were performed, and all the results reported below are the average over the twenty five runs.

Table I and Table III is the results of the first group. Table II and Table IV are the results of the second group. In Table I and Table II, the CCR is reported in percentage, and the value in parentheses represents the standard deviation. Furthermore, the notation N/A represent that the covariance matrix is singular, in which case reliable results can not be obtained. As it can be seen from the Table I and II, the feature dimension is critical in discriminant analysis in face recognition. The ill-posed problem will degrade the performance since classifiers are trained with higher feature dimension and fewer samples, while discriminant information will degrade if the feature dimension is reduced to very low. Although some samples can be separated in high dimension space, they cannot be separated anymore when projected into a reduced dimension space.

In both group experiments, QDA usually produces good results when training sample number is sufficient $d < 6$, but it faces ill-posed problems when the dimensionality is high $d > 6$ and the results will be unreliable. LDA achieves a best performance 95.29% with $d = 20$ in the first group experiment, and 99.71% with $d = 25$ in the second group experiment. RDA performs stable in the experiments with the best performance 97.86% with $d = 40$ and 100% with $d = 40$. LOOC is better in some experiments with an average performance 97.605% and 98.29% with $d > 10$, but it gets worse when we reduce the dimensionality to $d = 2$ with PCA. Compared with the above discriminant classifiers, KLIM's performance is stable with $d > 2$ both in KLIM1 and KLIM2 approaches. KLIM1 achieves the highest average performance 95.1014%, while RDA is 94.8171% in the first group experiment, and equivalent result 95.4186% in the second group experiment.

D. Experiment comparison with time cost

The time cost of the six kind of classifiers are compared in Table III and Table IV for the two group data, from which we can see that the training and testing time increase with the increasing of the feature dimensions. In the experiments,

TABLE I

COMPARISON RESULTS OF KLIM1, KLIM2, QDA, LDA, RDA, AND LOOC APPROACHES WITH DIFFERENT FEATURE DIMENSIONS FOR THE FIRST GROUP EXPERIMENT DATA.

Classifier	d = 40	d = 30	d = 25	d = 20	d = 10	d = 5	d = 2
QDA	N/A	N/A	N/A	N/A	N/A	72.86(0.0155)	88.00(0.0032)
LDA	N/A	94.86(0.0033)	95.29(0.0018)	95.71(0.0012)	92.14(0.0019)	92.14(0.0017)	85.43(0.0020)
RDA	97.86(0.0009)	97.29(0.0010)	96.86(0.0013)	97.14(0.0014)	95.14(0.0015)	92.14(0.0010)	87.29(0.0042)
LOOC	97.71(0.0008)	97.57(0.0009)	97.57(0.0009)	97.57(0.0006)	91.43(0.0023)	89.29(0.0019)	86.86(0.0033)
KLIM1(MSC)	96.43(0.0008)	96.14(0.0007)	96.71(0.0008)	96.29(0.0008)	95.14(0.0011)	94.29(0.0011)	90.71(0.0011)
KLIM2(MSC)	93.00(0.0011)	91.43(0.0010)	91.29(0.0010)	91.14(0.0010)	87.00(0.0038)	70.29(0.0142)	N/A
KLIM1(CV)	96.57(0.0009)	96.29(0.0007)	96.29(0.0008)	96.14(0.0008)	94.86(0.001)	93.57(0022)	88.14(0035)
KLIM2(CV)	94.57(0.001)	94.14(0.0014)	93(0.0022)	0.9343(0.0027)	89(0.0059)	82.29(0.0114)	90.14(0.004)

TABLE II

COMPARISON RESULTS OF KLIM1, KLIM2, QDA, LDA, RDA, AND LOOC APPROACHES WITH DIFFERENT FEATURE DIMENSIONS FOR THE SECOND GROUP EXPERIMENT DATA.

Classifier	d = 40	d = 30	d = 25	d = 20	d = 10	d = 5	d = 2
QDA	N/A	N/A	N/A	N/A	N/A	80.86(0.014)	74.29(0.0061)
LDA	N/A	99.439(0.0003)	99.71(0.0002)	99.43(0.0002)	99.57(0.0001)	96(0.0018)	80.43(0.0033)
RDA	100(0)	99.86(0)	99.57(0.0002)	99.86(0)	99.71(0.0002)	97.57(0.0019)	80(0.0035)
LOOC	98.14(0.0011)	98.29(0.001)	98.29(0.001)	98.29(0.001)	98(0.0011)	93(0.0026)	79.86(0.003)
KLIM1(MSC)	99(0.0003)	99.29(0.0002)	99.29(0.0002)	99.43(0.0002)	100(0)	97.43(0.0013)	75.71(0.0041)
KLIM2(MSC)	98.71(0.0004)	98.57(0.0005)	98.43(0.0005)	98.57(0.0004)	98.57(0.0004)	93.14(0.0024)	N/A
KLIM1(CV)	99(0.0003)	99.29(0.0002)	99.29(0.0002)	99.43(0.0002)	100(0)	97.43(0.0031)	75.71(0.0053)
KLIM2(CV)	98.71(0.0005)	98.57(0.0004)	98.43(0.0004)	98.57(0.0004)	98.57(0.0020)	93.14(0.0033)	N/A

TABLE III

COMPARISON TIME COST OF KLM1, KLIM2, QDA, LDA, RDA, AND LOOC APPROACHES FOR THE FIRST GROUP EXPERIMENT DATA.

Classifier		d = 40	d = 30	d = 25	d = 20	d = 10	d = 5	d = 2
QDA	Train Time	0.0031	0.0018	0.0019	0.0012	0.0006	0.0019	0.0056
	Test Time	1.0563	0.985	1.0438	1.2307	1.0555	0.05	0.038
LDA	Train Time	0.0031	0.0018	0.0019	0.0012	0.0006	0.0019	0.0056
	Test Time	1.3825	0.315	0.2407	0.185	0.0882	0.0495	0.0344
RDA	Train Time	19.7425	10.9068	9.225	7.7138	5.45	3.4806	0.9382
	Time of Para Estimation	19.7418	10.9056	9.2238	7.7132	5.4494	3.4806	0.9369
	Test Time	0.5812	0.3888	0.2912	0.2288	0.1263	0.0844	0.0331
LOOC	Train Time	0.507	0.3888	0.2937	0.223	0.1056	0.1169	0.0569
	Time of Para Estimation	0.507	0.3888	0.2931	0.2224	0.105	0.1162	0.0569
	Test Time	0.3711	0.2375	0.182	0.147	0.0719	0.0482	0.0325
KLIM1(MSC)	Train Time	0.0031	0.0018	0.0019	0.0012	0.0006	0.0019	0.0056
	Time of Para Estimation	0.0031	0.0018	0.0019	0.0012	0.0006	0.0019	0.0056
	Test Time	0.5219	0.3257	0.2351	0.1726	0.0806	0.0498	0.035
KLIM2(MSC)	Train Time	0.0219	0.014	0.0136	0.0106	0.0076	0.0088	0.01
	Time of Para Estimation	0.0031	0.0018	0.0019	0.0012	0.0006	0.0019	0.0056
	Test Time	0.4994	0.2986	0.2269	0.1644	0.0787	0.0514	0.0387
KLIM1(CV)	Train Time	38.6295	31.7069	39.305	37.1488	32.3725	10.0123	1.5923
	Time of Para Estimation	38.6295	31.7069	39.305	37.1488	32.3725	10.0123	1.5923
	Test Time	0.6143	0.4036	0.8868	0.5881	0.17	0.0913	0.0344
KLIM2(CV)	Train Time	40.2832	32.7619	41.8126	39.9131	35.6219	22.1	2.1444
	Time of Para Estimation	40.2638	32.7495	41.8	39.905	35.6187	22.0951	2.1389
	Test Time	0.5557	0.3356	0.5786	0.4481	0.2188	0.0526	0.0338

we also give the time cost of the parameters estimation with cross-validation and estimation criterion. The computation complexity of LDA and QDA is fixed as the baseline. RDA is the slowest classifiers for both its parameter tuning and training which involves two-dimensional optimization, while LOOC and KLIM only require one-dimension optimization computation complexity. The covariance matrix estimation in KLIM1 is simple while that of KLIM2 is a little complex. Also for time complexity, the training time of KLIM1 is the same with RDA, which illustrates that the optimal h in the computation of covariance matrix adds the recognition rate and almost without increasing training time. In the experiments with KLIM based approaches, two parameters calculation ways are compared in Table III and Table IV. The cross-validation takes about 40 seconds to parameters estimation for $d = 40$, while the estimation criterion only takes about 0.003 seconds. The speed is improved above 3000 times for KLIM based approaches.

From the experiments with ORL face data, we can find QDA is poor for face recognition problem. LDA is fast but the performance will degrade with the increasing of the feature dimensionality. RDA always gives satisfactory results with some more computations in optimizing regularization parameters, and LOOC could be better under some specific situations. Considering both the time cost and classification rates, the performance of KLIM classifiers is better than other classifiers in the high dimension-setting case.

TABLE IV
COMPARISON TIME COST OF KLM1, KLIM2, QDA, LDA, RDA, AND LOOC APPROACHES FOR THE SECOND GROUP EXPERIMENT DATA.

Classifier		d = 40	d = 30	d = 25	d = 20	d = 10	d = 5	d = 2
QDA	Train Time	0.0031	0.0013	0.0038	0.0013	0.0018	0.0019	0.0044
	Test Time	1.1995	1.2843	1.2393	1.4333	1.3813	0.0506	0.0356
LDA	Train Time	0.0031	0.0013	0.0044	0.0013	0.0018	0.0019	0.0044
	Test Time	1.643	0.4002	0.2857	0.2074	0.1032	0.0526	0.0376
RDA	Train Time	20.5137	11.025	10.3994	7.8644	5.6481	3.4326	0.9619
	Time of Para Estimation	20.5131	11.0226	10.3981	7.8638	5.6481	3.4326	0.96
	Test Time	0.588	0.3507	0.2831	0.2162	0.122	0.0857	0.0368
LOOC	Train Time	0.53	0.3961	0.31	0.2401	0.113	0.1144	0.0557
	Time of Para Estimation	0.5294	0.3955	0.31	0.2401	0.113	0.1144	0.0557
	Test Time	0.3901	0.2457	0.1888	0.1499	0.0732	0.0519	0.0337
KLIM1(MSC)	Train Time	0.0037	0.0019	0.0063	0.0013	0.0024	0.0019	0.0044
	Time of Para Estimation	0.0037	0.0019	0.005	0.0013	0.0024	0.0019	0.0044
	Test Time	0.5438	0.3261	0.2387	0.1769	0.085	0.0536	0.0355
KLIM2(MSC)	Train Time	0.0211	0.0139	0.0156	0.0106	0.0054	0.0074	0.0094
	Time of Para Estimation	0.0031	0.0013	0.0044	0.0013	0.0018	0.0019	0.0044
	Test Time	0.55	0.3313	0.2438	0.1844	0.0864	0.0558	0.0287
KLIM1(CV)	Train Time	46.3682	41.2437	43.7668	39.5006	39.5073	9.5357	1.6012
	Time of Para Estimation	46.3682	41.2431	43.7668	39.5006	39.5073	9.5357	1.6006
	Test Time	0.6037	0.3886	1.0444	1.0556	0.572	0.09	0.0344
KLIM2(CV)	Train Time	56.3464	49.2188	46.8118	43.6068	39.0231	23.438	2.1225
	Time of Para Estimation	56.3275	49.2039	46.8013	43.5976	39.0168	23.433	2.1187
	Test Time	0.5649	0.3507	1.075	1.0276	0.7219	0.0514	0.0338

VI. CONCLUSION

In this paper, we investigated the performance of KLIM algorithm with ORL face data sets, and compared it with QDA, LDA, RDA, and LOOC approaches. The performance of KLIM classifiers is better than other classifiers in the high dimension-setting case, and has a stable performance with the change of feature dimensionality.

The classification accuracy of KLIM is directly associated with the regularization parameter. We compared optimized regularization parameter by our derived model selection criterion and cross-validation method. Experiment shows that the model selection criterion obtains a faster and more accuracy results compared with cross-validation method.

ACKNOWLEDGMENT

The research work described in this paper was fully supported by the grants from the National Natural Science Foundation of China (Project No. 60675011, 60703070).

Authors would like to thanks Ms. Jing Shao for her discussion and part of the experiment work.

REFERENCES

[1] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips, "Face recognition: a literature survey," ACM Computing Surveys, vol. 35, No. 4, pp. 399-458, 2003.

[2] A. Samal and P. A.Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey," Pattern Recognition, vol. 25, pp. 65-77, 1992.

[3] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," Proceedings of the IEEE, vol. 83, pp. 705-740, 1995.

[4] A.R. Webb, Statistical pattern recognition, In: Oxford University Press, London, 1994.

[5] S. Ranganath and K. Arun, "Face Recognition Using Transform Features and Neural Networks," Pattern Recognition, vol. 30, No. 10, pp. 1615-1622, 1997.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," IEEE Trans. Pattern Anal. and Machine Intell., vol. 9, pp. 711-720, 1997.

[7] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional dataCwith application to face recognition," Pattern Recognition, vol. 34, pp. 2067-2070, 2001.

[8] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," Pattern Recognition Letter, vol. 26, No. 2, pp. 181-191, 2005.

[9] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," IEEE Trans. Neural Networks, vol. 14, No. 1, pp. 117-126, 2003.

[10] I. T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1996.

[11] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object recognition," IEEE Trans. on Pattern Anal. and Machine Intell., vol. 19, No. 7, pp. 696-710, 1997.

[12] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol. 3, No. 1, pp. 71-86, 1991.

[13] B. Scholkopf, A. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, vol. 10, No. 5, pp. 1299-1319, 1998.

[14] S. Aeberhard, D. Coomans, and O. de Vel, "Comparative analysis of statistical pattern recognition methods in high dimensional settings," Pattern Recognition, vol. 27, 1994.

[15] R.A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, vol. 7, pp. 179-188, 1936.

[16] J.H. Friedman, "Regularized discriminant analysis," Jounal of American Statistial Analysis, vol. 84, No. 405, pp. 1437-1642, 1989.

[17] J.P. Hoffbeck and D.A. Landgrebe, "Covariance matrix estimation and classification with limited training data," IEEE Trans. on Pattern Anal. and Machine Intell., vol.18, pp. 763-767, 1996.

[18] S. Tadjudin and D.A. Landgrebe, "Covariance estimation with limited training samples," IEEE Trans. Geosci. Remote Sensing, vol. 37, No. 4, pp. 2113-2118, 1999.

[19] I. Rivals and L. Personnaz, "On cross validation for model selection," Neural Computation, vol. 11, pp. 863-870, 1999.

[20] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.

[21] P. Guo, Y. Jia, and M. R. Lyu, "A study of regularized Gaussian classifier in high-dimension small sample set case based on MDL principle with application to spectrum recognition," Pattern Recognition, in press.

[22] P. Guo, "Spectral Pattern Recognition with Regularized Gaussian Classifier," IEEE Int. Conf. Neural Networks & Signal Processing Nanjing, China, December 2003.

[23] P. Guo and M.R. Lyu, "Classification for high-dimension small-sample data sets based on kullback-leibler information measure," Proceedings of the 2000 Int. Conf. on Artificial Intelligence, pp. 1187-1193, 2000.