L1-Norm Kernel Discriminant Analysis via Bayes Error Bound Optimization for Robust Feature Extraction

Wenming Zheng, *Member, IEEE*, Zhouchen Lin, *Senior Member, IEEE*, and Haixian Wang, *Member, IEEE*

***Abstract*— A novel discriminant analysis criterion is derived in this paper under the theoretical framework of Bayes optimality. In contrast to the conventional Fisher's discriminant criterion, the major novelty of the proposed one is the use of L1 norm rather than L2 norm, which makes it less sensitive to the outliers. With the L1-norm discriminant criterion, we propose a new linear discriminant analysis (L1-LDA) method for linear feature extraction problem. To solve the L1-LDA optimization problem, we propose an efficient iterative algorithm, in which a novel surrogate convex function is introduced such that the optimization problem in each iteration is to simply solve a convex programming problem and a close-form solution is guaranteed to this problem. Moreover, we also generalize the L1-LDA method to deal with the nonlinear robust feature extraction problems via the use of kernel trick, and hereafter proposed the L1-norm kernel discriminant analysis (L1-KDA) method. Extensive experiments on simulated and real data sets are conducted to evaluate the effectiveness of the proposed method in comparing with the state-of-the-art methods.**

***Index Terms*— Linear discriminant analysis (LDA), L1-norm linear discriminant analysis (L1-LDA), L1-norm kernel discriminant analysis (L1-KDA), robust feature extraction.**

I. Introduction

FEATURE extraction plays a very important role in pattern classification [1]. A major goal of feature extraction is to reduce the dimensionality of data points for the purpose of data visualization or discrimination. During the last several decades, many feature extraction methods have been developed in the literatures [2]. Among the various methods, principal component analysis (PCA) and linear discriminant analysis (LDA) are two of the most popular ones [3]. PCA is an unsupervised feature extraction method aiming to find a set of optimal representative projection vectors such that the projections of the data points can best preserve the structure of the data distribution, whereas LDA is a supervised method aiming to find an optimal set of discriminative projection vectors that simultaneously maximize the between-class distance and minimize the within-class distance such that they have better discrimination ability in the reduced feature space. A common property of PCA and LDA is that both methods are derived based on the L2 norm.

Although the L2-norm-based feature extraction methods had been successfully applied to many pattern recognition applications, they are prone to suffering from the outliers compared with the L1-norm-based ones since the effect of large norm outliers may be more exaggerated by the L2 norm than the L1 one [7]. Consequently, to overcome this drawback of L2 norm, many researchers turned to use L1 norm instead of L2 norm in developing the robust feature extraction method, e.g., robust PCA, in recent years [4]–[9]. In contrast to the L2-norm PCA method, the major advantage of the L1-norm PCA (L1-PCA) method is that it may be less sensitive to the effect of outliers [7]. Despite of the potential advantages of L1-PCA, however, it is notable that the optimization of the L1-PCA method is more difficult than the conventional L2-norm PCA method due to the absolute value operator in L1-PCA, where the L2-norm PCA can be simply solved via the singular value decomposition (SVD) of the covariance matrix. To resolve the optimal solution of L1-PCA, Kwak [7] proposed a greedy iteration algorithm to find its local optimizer, and experimentally demonstrated the superiority over the conventional PCA method in the facial image reconstruction experiments in the cases of contaminated face image data [7].

Noting that L1-PCA is an unsupervised feature extraction method whose goal is to find the optimal representative projection vectors rather than the optimal discriminative ones, it may not deliver good result for the pattern classification problems. In such cases, supervised feature extraction methods such as LDA would be a better choice. To deal with the robust discriminative feature extraction problem, Kim *et al*. [10] proposed a robust fisher discriminant analysis (RFDA) by optimizing the class means and class covariance matrices under a model of data uncertainty in a classification problem. Huang *et al*. [11] proposed a robust regression approach based on the low-rank representation of the

Manuscript received August 23, 2012; revised August 28, 2013; accepted September 2, 2013. Date of publication October 3, 2013; date of current version March 10, 2014. This work was supported in part by the National Basic Research Program of China under Grant 2011CB302202, in part by the National Natural Science Foundation of China under Grant 61231002 and Grant 61073137, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20130020, in part by the Ph.D. Program Foundation of the Ministry Education of China under Grant 20120092110054, and in part by the Program for Distinguished Talents of Six Domains in Jiangsu Province of China under Grant 2010-DZ088.

W. Zheng and H. Wang are with the Key Laboratory of Child Development and Learning Science, Research Center for Learning Science, Southeast University, Nanjing 210096, China (e-mail: wenming_zheng@seu.edu.cn; hxwang@seu.edu.cn).

Z. Lin is with the Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: zlin@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2013.2281428

training data matrix, in which the training data matrix is factorized into a low-rank data matrix corresponding to the clean data and a spare data matrix corresponding to noise data. Tzimiropoulos *et al*. [12] proposed a gradient orientation method for image description, and then carried out the feature extraction on the image gradient orientation to achieve the robust feature extraction. Zafeiriou *et al*. [13] proposed a regularized kernel discriminant analysis (RKDA) method with a cosine kernel based on the image gradient orientation method for robust feature extraction. He *et al*. [14] proposed to use the maximum correntropy criterion for robust feature extraction. Recently, Wang *et al*. [15] proposed an L1-norm based common spatial patterns (L1-CSP) method for robust electroencephalography (EEG) feature extraction, where the L2 norm appeared in the conventional CSP method is replaced by the L1 norm and the better performance had been visualized in the EEG classification experiments.

In this paper, we investigate the robust feature extraction problem using discriminant analysis method. However, different from the previous robust discriminant analysis methods, we focus our attention on the L1 norm to deal with the feature extraction problems in the cases that the data samples are contaminated by outliers. Although the similar idea recently appeared in [16] and [17], both papers lack of rigorous theoretical derivations of the L1-norm discriminant criteria, in which the authors just simply replace the L2 norm in the traditional LDA formulation with the L1 one. Different from both [16] and [17], in this paper we first derive a novel L1-norm discriminant criterion under a rigorous theoretical framework of Bayes optimality instead of simply replacing the L2 norm with L1 norm in the conventional Fisher's discriminant criterion. Then we propose an L1-norm linear discriminant analysis (L1-LDA) based on this new discriminant criterion for linear feature extraction. Due to the absolute value operation, however, the conventional optimization approach of solving a generalized eigensystem for L2-norm LDA will not be applicable for the L1-LDA method. Hence, a new optimization method for L1-LDA is required in this paper.

In our preliminary work on L1-CSP [15], we proposed a gradient ascending-based algorithm to iteratively update the optimal spatial filters of L1-CSP, where a nonconvex surrogate function was introduced for this purpose. However, this method needs to choose an appropriate stepsize in updating the new spatial filters. Since the surrogate function is nonconvex, the inappropriate choice of the stepsize will affect the optimality of the solution. To obtain the local optimal solution of L1-LDA, we introduce a new surrogate function. Compared with the previous one used in [15], the new surrogate function is convex such that the original L1-LDA optimization problem can be solved via solving a series of convex programming problems in which a close-form solution can be obtained in each convex programming problem.

To deal with the nonlinear robust feature extraction problem, in this paper we also generalize the L1-LDA method by mapping the input data points from the input space to a high-dimensional reproducing kernel Hilbert space (RKHS) via a nonlinear mapping, and then perform the linear feature extraction in RKHS using the L1-LDA method. This method is referred to as the L1-norm kernel discriminant analysis (L1-KDA) method in the paper. By utilizing the kernel trick as well as the representation theory [18], we show that the L1-KDA method can be solved using the same optimization approach of L1-LDA. In addition, similar to what Yang *et al*. [19] had found about the equivalence between the kernel discriminant analysis (KDA) [20] method and the kernel principal component analysis (KPCA) [21] plus LDA, it is interesting to see that the L1-KDA method can also be expressed as L1-KDA = KPCA + L1-LDA. Finally, to evaluate the robustness as well as the better discriminative ability of the proposed method compared with several state of the art methods, we will conduct extensive experiments on both simulated and real data sets in this paper.

This paper is organized as follows. In Section II, we present the L1-norm discriminant criterion under the theoretical framework of Bayes optimality. The L1-LDA method is presented in Section III. In Section IV, we propose the L1-KDA method. In Section V, we present the experiments of evaluating the proposed method. In Section VI, we conclude the paper.

II. L1-Norm Discriminant Criterion via Bayes Error Bound Estimation

Suppose that $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^d | i = 1, \ldots, N\}$ is the data set with N data samples corresponding to c classes, and let $\mathbf{l}_i = (l_{ij})_{N\times 1}$ denote the corresponding label vector of $\mathbf{x}_i$, where each element $l_{ij} \in \{1, \ldots, c\}$ indicates the class membership associated with the data sample $\mathbf{x}_i$. Let $\mathbf{x} \in \mathbb{R}^d$ be a sample vector, and $p_i(\mathbf{x})$ and P_i be the probability density function (PDF) and the prior probability of the ith class, respectively. Then, the multiclass Bayes error can be expressed as [22]:

$$\varepsilon = 1 - \int \max_i \{P_i\, p_i(\mathbf{x})\}\, d\mathbf{x} \tag{1}$$

which satisfies the following inequality [22]:

$$\varepsilon \leq \sum_{i<j} \int \sqrt{P_i\, p_i(\mathbf{x}) P_j\, p_j(\mathbf{x})} d\mathbf{x}. \tag{2}$$

Assume that the c classes of data sets are homocedastic and the PDF of the ith class is a Gaussian function, i.e., $p_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}_i, \Sigma)$, where $\mathbf{m}_i$ and Σ denote the class mean and the class covariance matrix, respectively. If we project the samples onto a projection vector $\omega \in \mathbb{R}^d$, then the projected data samples become $\tilde{\mathbf{X}} = \{\tilde{\boldsymbol{x}}_i = \omega^T \mathbf{x}_i \in \mathbb{R}^d | i = 1, \ldots, N\}$, and the PDF of the projected samples will become $p_i(\tilde{x}) = \mathcal{N}(\tilde{x}|\tilde{m}_i, \sigma^2)$, where ω^T denotes the transpose operation of ω, $\tilde{m}_i = \omega^T \mathbf{m}_i$ is the ith class mean and σ is the standard variance of the data samples $\tilde{x}_i$ $(i = 1, \ldots, N)$ and can be calculated as follows:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\tilde{x}_i - \tilde{m}_{l_i})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\omega^T \mathbf{x}_i - \omega^T \mathbf{m}_{l_i})^2}. \tag{3}$$

The standard variance σ measures the deviation of the data samples away from their corresponding class means. However, it is notable that the value of (3) may be prone to suffering from outliers since the effect of the outliers with a large distance $|\tilde{x}_i - \tilde{m}_{l_i}|$ would be largely exaggerated by the

square operation. For this reason, we use the average distance $1/N\sum_{i=1}^{N}|\tilde{x}_i-\tilde{m}_{l_i}|$ to represent the standard variance σ in order to alleviate the effect of the square operation, i.e.,

$$\sigma=\frac{1}{N}\sum_{i=1}^{N}\left|\tilde{x}_i-\tilde{m}_{l_i}\right|=\frac{1}{N}\sum_{i=1}^{N}\left|\omega^T\mathbf{x}_i-\omega^T\mathbf{m}_{l_i}\right|. \quad (4)$$

Then, by substituting the expressions of $p_i(\tilde{x})=\mathcal{N}(\tilde{x}|\tilde{m}_i,\sigma^2)$ into (2), we obtain the multiclass Bayes error upper bound can be expressed as [22] follows:

$$\varepsilon(\omega)\le\sum_{i<j}\sqrt{P_iP_j}\exp\left\{-\frac{1}{8}\left[\frac{\omega^T(\mathbf{m}_i-\mathbf{m}_j)}{\frac{1}{N}\sum_{k=1}^{N}|\omega^T\mathbf{x}_k-\omega^T\mathbf{m}_{l_k}|}\right]^2\right\}. \quad (5)$$

The detailed derivations of (5) is also given in Appendix A.

It is still difficult to compute the optimal projection vector ω such that the right hand side of (5) is minimized. So we have to simplify it. For this purpose, we introduce the following lemma:

Lemma 1: Let $h(x)=\exp(-x^2)$ $(0\le x\le a)$. Then $\hat{h}(x)=1-(1-\exp(-a^2))/ax$ $(0\le x\le a)$ is the tightest linear upper bound of $h(x)$.

Proof: $h(x)$ is a convex function on the interval $[0,a]$. So the linear function passing through its two ends, $(0,h(0))$ and $(a,h(a))$, is the tightest linear upper bound of $h(x)$. This function is $\hat{h}(x)$.

By applying Lemma 1 to the expression of the right-hand side of (5), we obtain that

$$\varepsilon(\omega)\le\sum_{i<j}\sqrt{P_iP_j}\left\{1-\frac{B_{ij}|\omega^T(\mathbf{m}_i-\mathbf{m}_j)|}{\frac{1}{N}\sum_{k=1}^{N}|\omega^T\mathbf{x}_k-\omega^T\mathbf{m}_{l_k}|}\right\} \quad (6)$$

where B_{ij} are tradeoff coefficients.

Without loss of generality, we simply set the value of each tradeoff coefficient B_{ij} $(i<j)$ equals to the same fixed one, i.e., $B_{ij}=B$, and then further assume that the c prior probabilities P_i are equal to P, i.e., $P=P_1=\frac{N_1}{N}=\cdots=P_c=\frac{N_c}{N}$, where N_i $(i=1,\cdots,c)$ is the number of ith class data samples. Let $\mathbf{m}=1/c\sum_{j=1}^{c}\mathbf{m}_j$ be the mean of the c classes. Then, by applying the following inequality:

$$\begin{aligned}&\sum_{i=1}^{c}\sum_{j=1}^{c}\left|(\omega^T\mathbf{m}_i-\omega^T\mathbf{m}_j)\right|\\&\ge\sum_{i=1}^{c}\left|\sum_{j=1}^{c}(\omega^T\mathbf{m}_i-\omega^T\mathbf{m}_j)\right|\\&=\sum_{i=1}^{c}\left|c\omega^T\mathbf{m}_i-\sum_{j=1}^{c}\omega^T\mathbf{m}_j\right|=c\sum_{i=1}^{c}\left|\omega^T\mathbf{m}_i-\omega^T\mathbf{m}\right|\end{aligned} \quad (7)$$

to the Bayes error upper bound of (6), we obtain the following new Bayes error upper bound with simpler expression:

$$\begin{aligned}\varepsilon(\omega)&\le\sum_{i<j}\sqrt{P_iP_j}\left\{1-\frac{B_{ij}|\omega^T(\mathbf{m}_i-\mathbf{m}_j)|}{\frac{1}{N}\sum_{k=1}^{N}|\omega^T\mathbf{x}_k-\omega^T\mathbf{m}_{l_k}|}\right\}\\&\le\sum_{i<j}P-\frac{cPB\sum_{i=1}^{c}|\omega^T(\mathbf{m}_i-\mathbf{m})|}{\frac{2}{N}\sum_{k=1}^{N}|\omega^T\mathbf{x}_k-\omega^T\mathbf{m}_{l_k}|}\end{aligned} \quad (8)$$

In addition, by applying the following inequality:

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}|\omega^T\mathbf{x}_i-\omega^T\mathbf{m}_{l_i}|\\&=\frac{1}{N}\sum_{i=1}^{N}|\omega^T\mathbf{x}_i-\omega^T\mathbf{m}+\omega^T\mathbf{m}-\omega^T\mathbf{m}_{l_i}|\\&\le\frac{1}{N}\sum_{i=1}^{N}|\omega^T\mathbf{x}_i-\omega^T\mathbf{m}|+\frac{1}{N}\sum_{i=1}^{N}|\omega^T\mathbf{m}-\omega^T\mathbf{m}_{l_i}|\\&=\frac{1}{N}\sum_{i=1}^{N}|\omega^T\mathbf{x}_i-\omega^T\mathbf{m}|+\sum_{i=1}^{c}P|\omega^T\mathbf{m}_i-\omega^T\mathbf{m}|\end{aligned} \quad (9)$$

to the Bayes error upper bound of (8), we finally obtain the following new Bayes error upper bound:

$$\varepsilon(\omega)\le\sum_{i<j}P-\frac{cPB\sum_{i=1}^{c}|\omega^T(\mathbf{m}_i-\mathbf{m})|}{\frac{2}{N}\sum_{k=1}^{N}|\omega^T\mathbf{x}_k-\omega^T\mathbf{m}|+2P\sum_{k=1}^{c}|\omega^T\mathbf{m}_k-\omega^T\mathbf{m}|}. \quad (10)$$

Consequently, to minimize the Bayes error, we should minimize the upper bound of (10). Equivalently, we have to maximize the following objective function:

$$J_1(\omega)=\frac{\sum_{i=1}^{c}|\omega^T(\mathbf{m}_i-\mathbf{m})|}{\frac{1}{PN}\sum_{i=1}^{N}|\omega^T\mathbf{x}_i-\omega^T\mathbf{m}|+\sum_{i=1}^{c}|\omega^T\mathbf{m}_i-\omega^T\mathbf{m}|}$$

which is also equivalent to maximizing the following L1-norm objective function:

$$J_2(\omega)=\frac{\sum_{i=1}^{c}|\omega^T(\mathbf{m}_i-\mathbf{m})|}{\sum_{i=1}^{N}|\omega^T\mathbf{x}_i-\omega^T\mathbf{m}|}=\frac{\|\omega^T\mathbf{B}\|_1}{\|\omega^T\mathbf{T}\|_1} \quad (11)$$

where $\mathbf{B}=[(\mathbf{m}_1-\mathbf{m}),\ldots,(\mathbf{m}_c-\mathbf{m})]$, $\mathbf{T}=[(\mathbf{x}_1-\mathbf{m}),\ldots,(\mathbf{x}_N-\mathbf{m})]$, and $\|\cdot\|_1$ denotes the L1-norm operation of vector.

III. L1-LDA for Linear Feature Extraction

In the aforementioned Section II, we developed a new discriminant criterion with L1 norm. With this new discriminant criterion, in this section we will propose an L1-LDA method for linear feature extraction, in which the optimal discriminant vectors are defined as the ones that maximize the L1-norm discriminant criterion $J_2(\omega)$. In other words, the optimal discriminant vectors of L1-LDA are defined as the solution of the following optimization problem:

$$\arg\max_{\omega}\frac{\|\omega^T\mathbf{B}\|_1}{\|\omega^T\mathbf{T}\|_1}. \quad (12)$$

The optimization problem of (12) is nonconvex and hence there may exist several local minimizers. Moreover, due to the absolute value operation, the optimization problem of (12) is much more difficult than that of the traditional LDA method [1]. To solve the L1-LDA problem, we firstly give the definition about the optimality of two vectors.

Definition 1: Let ϕ_1 and ϕ_2 denote two d-dimensional projection vectors. Let $\mathbf{b}_j$ and $\mathbf{t}_j$ be the jth column of $\mathbf{B}$ and $\mathbf{T}$, respectively. Suppose that

$$Q(\phi_2) = \frac{\|\phi_2^T\mathbf{B}\|_1}{\|\phi_2^T\mathbf{T}\|_1} = \frac{\sum_{j=1}^{c}|\phi_2{}^T\mathbf{b}_j|}{\sum_{j=1}^{N}|\phi_2{}^T\mathbf{t}_j|} \geq \frac{\sum_{j=1}^{c}|\phi_1{}^T\mathbf{b}_j|}{\sum_{j=1}^{N}|\phi_1{}^T\mathbf{t}_j|} = \frac{\|\phi_1^T\mathbf{B}\|_1}{\|\phi_1^T\mathbf{T}\|_1} = Q(\phi_1). \quad (13)$$

Then, we say ϕ_2 is a better vector than ϕ_1 in terms of $Q(\phi)$.

A. Solving the First Discriminant Vector of L1-LDA

Noting that the objective function of (12) is invariant to the magnitude of ω, we can scale ω such that the nominator of (12) equal to 1. Then, the optimization problem of (12) can be rewritten as

$$\arg\min_{\omega}\sum_{j=1}^{N}|\omega^T\mathbf{t}_j| = \omega^T\left(\sum_{j=1}^{N}\frac{\mathbf{t}_j\mathbf{t}_j^T}{|\omega^T\mathbf{t}_j|}\right)\omega \quad (14)$$

$$\text{s.t.}\|\omega^T\mathbf{B}\|_1 = \sum_{j=1}^{c}\text{sgn}(\omega^T\mathbf{b}_j)\omega^T\mathbf{b}_j = 1$$

where $\text{sgn}(a)$ denotes the positive or negative sign of a

$$\text{sgn}(a) = \begin{cases} +1, & a \geq 0; \\ -1, & a < 0. \end{cases} \quad (15)$$

To solve the optimal projection vector of (14), we propose an iterative algorithm in what follows. The basic idea of the proposed algorithm is to iteratively update the projection vector ω until it converges to a local optimizer. Specifically, suppose that $\omega^{(p)}$ is the optimal projection vector solved in the pth iteration, and $\omega^{(p+1)}$ is the one of the $(p+1)$th iteration, where $\omega^{(p+1)}$ is defined by

$$\omega^{(p+1)} = \arg\min_{\omega}\omega^T\mathbf{V}_t(\omega^{(p)})\omega \quad (16)$$

$$\text{s.t.}\sum_{j=1}^{c}s_j^{(p)}\omega^T\mathbf{b}_j = 1$$

where $\mathbf{V}_t(\omega^{(p)}) = \sum_{i=1}^{N}\mathbf{t}_i\mathbf{t}_i^T/|\omega^{(p)}{}^T\mathbf{t}_i|$ and $s_j^{(p)} = \text{sgn}(\omega^{(p)}{}^T\mathbf{b}_j)$ Then we can prove that $\omega^{(p+1)}$ is better than $\omega^{(p)}$. This observation is summarized in Theorem 1. To prove it, we first introduce the following Lemma 2.

Lemma 2: For any vector $\mathbf{a} = [a_1, \ldots, a_N]^T \in \mathbb{R}^N$, the following variational equality holds [23]:

$$\|\mathbf{a}\|_1 = \min_{\mathbf{z}\in\mathbb{R}_+^N}\frac{1}{2}\sum_{j=1}^{N}\frac{a_j^2}{z_j} + \frac{1}{2}\|\mathbf{z}\|_1 \quad (17)$$

and the minimum is uniquely reached at $z_j = |a_j|$ for $j = 1, \ldots, N$, where $\mathbf{z} = [z_1, \ldots, z_N]^T$.

Theorem 1: Suppose that $\omega^{(p)}$ is a d-dimensional vector such that $\sum_{j=1}^{c}|\omega^{(p)}{}^T\mathbf{b}_j| = 1$. Let $\mathbf{s}^{(p)} = (s_j^{(p)})_{c\times 1}$ be a c-dimensional vector and $s_j^{(p)} = \text{sgn}(\omega^{(p)}{}^T\mathbf{b}_j)$. Suppose that $\omega^{(p+1)}$ is the solution of (16), then $\omega^{(p+1)}$ is better than $\omega^{(p)}$. The proof of Theorem 1 is given in Appendix B.

Theorem 1 guarantees the convergence of the aforementioned L1-LDA algorithm. Noting that the optimization problem of (16) is a linear constraint quadratic programming problem and has a close-form solution, we can get the solution via Lagrangian multiplier approach [24]. The Lagrangian can be expressed as follows:

$$L(\omega) = \frac{1}{2}\omega^T\mathbf{V}_t(\omega^{(p)})\omega - \lambda\left(\sum_{j=1}^{c}s_j^{(p)}\omega^T\mathbf{b}_j - 1\right). \quad (18)$$

Taking the partial derivative of L with respect to ω and setting it to be a zero value, we have

$$\frac{\partial L}{\partial\omega} = \mathbf{V}_t(\omega^{(p)})\omega - \lambda\sum_{j=1}^{c}s_j^{(p)}\mathbf{b}_j = 0. \quad (19)$$

From (19), we obtain that

$$\omega^{(p+1)} = \lambda\left[\mathbf{V}_t(\omega^{(p)})\right]^{-1}\mathbf{B}\mathbf{s}^{(p)}. \quad (20)$$

Substituting (20) into the equality $\sum_{j=1}^{c}s_j^{(p)}\omega^T\mathbf{b}_j = 1$, we obtain that

$$\lambda = \frac{1}{\left(\mathbf{B}\mathbf{s}^{(p)}\right)^T\left[\mathbf{V}_t(\omega^{(p)})\right]^{-1}\left(\mathbf{B}\mathbf{s}^{(p)}\right)}. \quad (21)$$

Combining (20) and (21), we have the following close-form solution of $\omega^{(p+1)}$:

$$\omega^{(p+1)} = \frac{\left[\mathbf{V}_t(\omega^{(p)})\right]^{-1}\mathbf{B}\mathbf{s}^{(p)}}{\left(\mathbf{B}\mathbf{s}^{(p)}\right)^T\left[\mathbf{V}_t(\omega^{(p)})\right]^{-1}\left(\mathbf{B}\mathbf{s}^{(p)}\right)}. \quad (22)$$

By increasing the iteration number p until $\omega^{(p)}$ converges to a fixed value, we obtain the local optimizer of ω as

$$\omega = \lim_{p\to\infty}\omega^{(p)}.$$

B. Solving Multiple Discriminant Vectors of L1-LDA

Assume that we have obtained the first $r-1$ $(r > 1)$ discriminant vectors $\omega_1, \ldots, \omega_{r-1}$. Then, the rth discriminant vector ω_r is defined as the solution of the following optimization problem:

$$\arg\max_{\omega}\frac{\|\omega^T\mathbf{B}\|_1}{\|\omega^T\mathbf{T}\|_1}, \quad \text{s.t. } \omega^T\mathbf{S}_t\omega_j = 0, (j = 1, \ldots, r-1) \quad (23)$$

where $\mathbf{S}_t = 1/N\mathbf{T}\mathbf{T}^T$ is the covariance matrix of the data samples, and $\omega^T\mathbf{S}_t\omega_j = 0, (j = 1, \ldots, r-1)$ is served as a statistically uncorrelated restriction such that the discriminant vectors are statistically uncorrelated [29]. To solve the optimal discriminant vectors of (23), we introduce the following Lemma 3, whose proof can be obtained by following the method in [25]. We give the proof in Appendix C.

Lemma 3: Let $\mathbf{U}_{r-1} = [\mathbf{S}_t\omega_1, \ldots, \mathbf{S}_t\omega_{r-1}]$. Suppose that $\mathbf{U}_{r-1} = \mathbf{Q}_{r-1}\mathbf{R}_{r-1}$ is the QR decomposition of $\mathbf{U}_{r-1}$, where the columns of $\mathbf{Q}_{r-1}$ are orthonormal and $\mathbf{R}_{r-1}$ is an upper triangular matrix. Then, there exists a (nonunique) α such that the discriminant vector ω that satisfies $\omega^T\mathbf{U}_{r-1} = \mathbf{0}$ can be expressed as

$$\omega = (\mathbf{I}_d - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)\alpha,$$

where $\mathbf{I}_d$ is a $d \times d$ identity matrix.

From Lemma 3, we obtain that solving the optimal discriminant vector ω_r in (23) is equivalent to solving the following optimization problem:

$$\alpha_r = \arg\max_{\omega} \frac{\|\alpha^T(\mathbf{I}_d - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)\mathbf{B}\|_1}{\|\alpha^T(\mathbf{I}_d - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)\mathbf{T}\|_1}. \tag{24}$$

This optimization problem can be solved using the solution method addressed in Section III-A. In this case, we can obtain the optimal discriminant vector $\omega_r = (\mathbf{I}_d - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)\alpha_r$. By repeating the above procedures, we can solve r (>1) discriminant vectors $\omega_1, \ldots, \omega_r$ of L1-LDA. Table I summarizes the complete algorithm of solving the multiple optimal discriminant vectors of L1-LDA.

Assume that we obtain r discriminant vectors of L1-LDA and denote the transform matrix of L1-LDA by $\mathbf{W}_r = [\omega_1, \ldots, \omega_r]$. Then the projection of a test sample $\mathbf{x}_t$ onto $\mathbf{W}_r$ can be expressed as

$$\mathbf{y}_t = \mathbf{W}_r^T \mathbf{x}_t. \tag{25}$$

IV. L1-KDA for Nonlinear Feature Extraction

In this section, we will generalize the L1-LDA method via the kernel trick [18] such that it is able to deal with the nonlinear robust feature extraction problem. Let Φ be a nonlinear mapping that maps the data points $\mathbf{x}_i$ from the input space $\mathbb{R}^d$ to a high-dimensional reproducing kernel Hilbert space (RKHS) $\mathcal{F}$,

$$\Phi : \mathbb{R}^d \mapsto \mathcal{F},\ \mathbf{x}_i \mapsto \Phi(\mathbf{x}_i) \tag{26}$$

in which the inner product of two data points in $\mathcal{F}$, say $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$, can be calculated via a kernel function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j).$$

Based on the derivation of L1-LDA in Section II, we can obtain that, in the kernel feature space $\mathcal{F}$, the L1-KDA method can be expressed as solving the following optimization problem:

$$\arg\max_{\omega} \frac{\left\|\omega^T \mathbf{B}^\Phi\right\|_1}{\left\|\omega^T \mathbf{T}^\Phi\right\|_1} \tag{27}$$

where the matrices $\mathbf{B}^\Phi$ and $\mathbf{T}^\Phi$ can be, respectively, defined as

$$\mathbf{B}^\Phi = \left[(\mathbf{m}_1^\Phi - \mathbf{m}^\Phi), \ldots, (\mathbf{m}_c^\Phi - \mathbf{m}^\Phi)\right]$$

and

$$\mathbf{T}^\Phi = [(\Phi(\mathbf{x}_1) - \mathbf{m}^\Phi), \ldots, (\Phi(\mathbf{x}_N) - \mathbf{m}^\Phi)]$$

in which $\mathbf{m}_i^\Phi$ and $\mathbf{m}^\Phi$ are defined as

$$\mathbf{m}_i^\Phi = \frac{1}{N_i} \sum_{j:l_j=i} \Phi(\mathbf{x}_j) \text{ and } \mathbf{m}^\Phi = \frac{1}{c} \sum_{j=1}^{c} \mathbf{m}_j^\Phi$$

and N_i is the number of the ith class data samples.

Now denote the data matrix in the feature space $\mathcal{F}$ by $\Phi(\mathbf{X}) = [\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_N)]$. Then, the mean vector of the ith class data set can be expressed as

$$\mathbf{m}_i^\Phi = \Phi(\mathbf{X})\mathbf{n}_i \tag{28}$$

TABLE I

Efficient Algorithm for Solving the Multiple Optimal Discriminant Vectors of L1-LDA

Input: The N data samples $\mathbf{x}_i$, class label vector l_i $(i = 1, \cdots, N)$ and the number of projection vectors r, and a small threshold ε;

1) Calculate the matrix $\mathbf{B}_0 = [(\mathbf{m}_1 - \mathbf{m}), \cdots, (\mathbf{m}_c - \mathbf{m})]$, and the matrix $\mathbf{T}_0 = [(\mathbf{x}_1 - \mathbf{m}), \cdots, (\mathbf{x}_N - \mathbf{m})]$, where $\mathbf{m}_i$ is the i-th class mean and $\mathbf{m} = \frac{1}{c}\sum_{j=1}^{c} \mathbf{m}_j$;
2) Calculate the covariance matrix $\mathbf{S}_t = \frac{1}{N}\mathbf{T}\mathbf{T}^T$;
3) Set $\mathbf{B} \leftarrow \mathbf{B}_0$ and $\mathbf{T} \leftarrow \mathbf{T}_0$;

For i=1 to r **Do**

1) Set $p \leftarrow 0$, $\omega^{(p)} = \mathbf{0}$;
2) Choose an initial vector $\omega^{(p+1)}$ such that $\|\omega^{(p+1)} - \omega^{(p)}\| > \varepsilon$;
3) **While** $\|\omega^{(p+1)} - \omega^{(p)}\| > \varepsilon$ **Do**
   a) Set $p \leftarrow p + 1$;
   b) Set $s_j^{(p)} \leftarrow \text{sgn}({\omega^{(p)}}^T \mathbf{b}_j)$, $(j = 1, \cdots, c)$;
   c) Calculate $\mathbf{V}_t(\omega^{(p)}) = \sum_{i=1}^{N} \frac{\mathbf{t}_i \mathbf{t}_i^T}{|{\omega^{(p)}}^T \mathbf{t}_i|}$;
   d) Calculate $\omega^{(p+1)}$:

$$\omega^{(p+1)} \leftarrow \frac{\left[\mathbf{V}_t(\omega^{(p)})\right]^{-1} \mathbf{B}\mathbf{s}^{(p)}}{\left(\mathbf{B}\mathbf{s}^{(p)}\right)^T \left[\mathbf{V}_t(\omega^{(p)})\right]^{-1} \left(\mathbf{B}\mathbf{s}^{(p)}\right)};$$

4) Set $\omega_i \leftarrow \omega^{(p)}$;
5) Calculate the QR decomposition of $\mathbf{Q}_i\mathbf{R}_i = [\mathbf{S}_t\omega_1, \cdots, \mathbf{S}_t\omega_i]$, and calculate $\mathbf{B}$ and $\mathbf{T}$:

$$\mathbf{B} \leftarrow \mathbf{B}_0 - \mathbf{Q}_i\mathbf{Q}_i^T\mathbf{B}_0,\ \mathbf{T} \leftarrow \mathbf{T}_0 - \mathbf{Q}_i\mathbf{Q}_i^T\mathbf{T}_0;$$

Output: The r projection vectors ω_i $(i = 1, \cdots, r)$.

where $\mathbf{n}_i = [n_{i1}, \ldots, n_{iN}]^T$ is an $N \times 1$ vector whose jth entry is

$$n_{ij} = \begin{cases} \frac{1}{N_i}, & l_j = i; \\ 0, & \text{otherwise.} \end{cases}$$

From (28) we obtain that $\mathbf{B}^\Phi$ and $\mathbf{T}^\Phi$ can be expressed as

$$\mathbf{B}^\Phi = \Phi(\mathbf{X})\mathbf{N}_b \quad \text{and} \quad \mathbf{T}^\Phi = \Phi(\mathbf{X})\mathbf{N}_t$$

where $\mathbf{N}_b = [(\mathbf{n}_1 - \mathbf{n}), \ldots, (\mathbf{n}_c - \mathbf{n})]$, $\mathbf{n} = \frac{1}{c}\sum_{j=1}^{c} \mathbf{n}_j$, $\mathbf{N}_t = [\mathbf{e}_1 - \mathbf{n}, \ldots, \mathbf{e}_N - \mathbf{n}]$, and $\mathbf{e}_j$ is an $N \times 1$ unit vector with the jth entry equal to 1.

Substituting the expressions of $\mathbf{B}^\Phi$ and $\mathbf{T}^\Phi$ into (27), we obtain that the optimization problem of L1-KDA is equivalent to the following one:

$$\arg\max_{\omega} \frac{\left\|\omega^T \Phi(\mathbf{X})\mathbf{N}_b\right\|_1}{\left\|\omega^T \Phi(\mathbf{X})\mathbf{N}_t\right\|_1}. \tag{29}$$

According to the representation theory [18], we obtain that the optimal solution of ω lies in the span of $\Phi(\mathbf{x}_i)$ $(i = 1, \ldots, N)$, i.e., there exists a coefficient vector α, such that

$$\omega = \Phi(\mathbf{X})\alpha. \tag{30}$$

Consequently, solving the optimal vectors $\omega_1, \ldots, \omega_r$ boils down to solving the coefficient vectors $\alpha_1, \ldots, \alpha_r$ given by

$$\arg\max_{\alpha} \frac{\left\|\alpha^T \mathbf{K}\mathbf{N}_b\right\|_1}{\left\|\alpha^T \mathbf{K}\mathbf{N}_t\right\|_1} \tag{31}$$

where $\mathbf{K} = [k_{ij}]_{N\times N}$ is the $N \times N$ Gram matrix and $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. The optimal solutions of (31) can be solved using the L1-LDA algorithm presented in Table I.

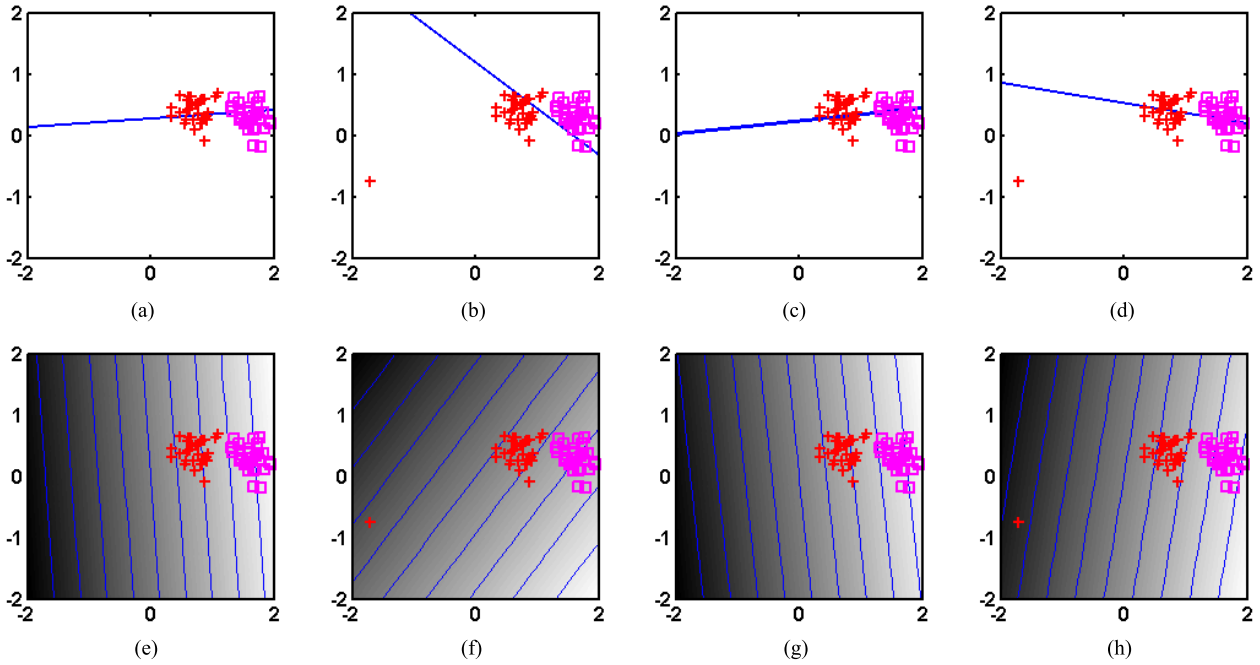


Fig. 1. Comparison of discriminant vectors between LDA and L1-LDA as well as the projections of data points around the regions of $(-2.0, 2.0)$ to $(2.0, -2.0)$ onto the discriminant vector of LDA and L1-LDA, respectively. (a) LDA without noise data. (b) LDA with noise data. (c) L1-LDA without noise data. (d) L1-LDA with noise data. (e) LDA without noise data. (f) LDA with noise data. (g) L1-LDA without noise data. (h) L1-LDA with noise data.

*Remarks:*Yang *et al.* [19] showed that kernel discriminant analysis (KDA) [20] is equivalent to KPCA [21] plus LDA, i.e.,

$$\text{KDA} = \text{KPCA} + \text{LDA}. \tag{32}$$

The similar result can also be observed in the L1-KDA method, in which the L1-KDA method can be expressed as KPCA plus L1-LDA, i.e.,

$$\text{L1-KDA} = \text{KPCA} + \text{L1-LDA}. \tag{33}$$

The derivation of this result is given in Appendix D.

Let $\mathbf{W}_r^{\Phi} = [\omega_1, \ldots, \omega_r]$ be the transform matrix of L1-KDA. Then the projection of a test point $\Phi(\mathbf{x}_t)$ onto $\mathbf{W}_r^{\Phi}$ can be expressed as

$$\mathbf{y}_t = \mathbf{W}_r^{\Phi^T} \Phi(\mathbf{x}_t) = \mathbf{A}_r^T \kappa_t \tag{34}$$

where $\mathbf{A}_r = [\alpha_1, \ldots, \alpha_r]$, $\kappa_t = [\kappa_{t,1}, \ldots, \kappa_{t,N}]^T$, where $\kappa_{t,i} = k(\mathbf{x}_i, \mathbf{x}_t)$.

V. Experiments

In this section, we will conduct extensive experiments on both simulated and real data sets to evaluate the discriminant performance of the proposed L1-LDA and L1-KDA methods. For comparisons, we also conduct the same experiments using several state-of-the-art feature extraction methods, including KPCA [21], KPCA + L1-PCA (L1-KPCA) [7], MCCKPCA [14], KPCA + LDA [20], and RKDA [13]. Throughout the experiments, the nearest neighbor classifier is adopted to evaluate the discriminant ability of the extracted features of the various methods.

A. Experiment on Simulated Data Set

In this experiment, two data sets with 2-D data points are generated to evaluate the robustness of L1-LDA against outliers. The two data sets are centered at $(1.6, 0.3)$ and $(0.7, 0.4)$, respectively. For each data set, a number of 30 data points are randomly generated under the gaussian distribution with zero mean and standard variance 0.2.

To visualize the differences of LDA and L1-LDA in the feature extraction, we calculate the discriminant vectors of the two methods, and then project a set of testing data points sampled from the region of $(-2.0, 2.0)$ to $(2.0, -2.0)$ onto the first discriminant vector of the two methods, respectively. Moreover, to evaluate the robustness of the two methods against outliers, we inserted one data point into the second data set (indicated by '+' with green color), which is far away from most of the data points of the second data set. Hence, this data point can be seen as an outlier of the data set. We recalculate the discriminant vectors as well as the projections of the testing data points onto the first discriminant vector of LDA and L1-LDA, respectively. Fig. 1(a) and (c) shows the optimal projection vectors of LDA and L1-LDA when no outlier points are inserted, whereas Fig. 1(b) and (d) shows the optimal projection vectors of LDA and L1-LDA when an outlier point is inserted into the training data set. The corresponding projection features extracted by LDA and L1-LDA are also shown in Fig. 1(e)–(h), in which the depicted are the feature values (indicated by gray levels) and contuor lines of identical feature values.

By comparing the feature extraction results of LDA and L1-LDA in Fig. 1, we can clearly see that the projection direction of the LDA method [see Fig. 1(a) and (b)] is

Fig. 2. 11 face images of one subject in Yale face data set. (a) Face images that do not suffer from outliers. (b) Face images occluded by a baboon image with the size of 10×10 pixels. (c) The 11 face images occluded by a baboon image with the size of 20×20 pixels. (d) The 11 face images occluded by a baboon image with the size of 30×30 pixels.

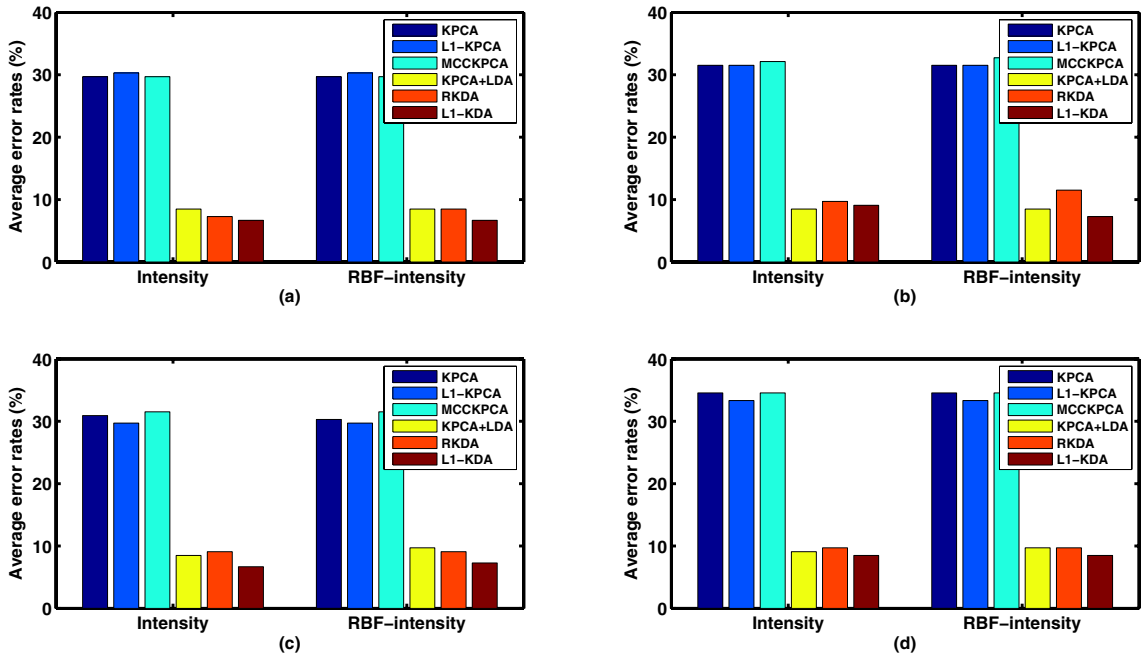


Fig. 3. Comparisons of the average test error rates among the various feature extraction methods on the Yale face data set. (a) The results correspond to the clean training data samples. (b) The results correspond to the training data samples with occlusions of size 10×10 pixels. (c) The results correspond to the training data samples with occlusions of size 20×20 pixels. (d) The results correspond to the training data samples with occlusions of size 30×30 pixels.

significantly changed when the training data set suffers from an outlier. In contrast to LDA, however, the change of the projection vector of L1-LDA is [see Fig. 1(c) and (d)] less. The experimental results indicate that the L1-LDA method would be less sensitive to the outliers than LDA due to the use of L1 norm.

B. Experiments on Yale Face Database

In this experiment, the Yale face database [26] is used to evaluate the recognition performance of L1-KDA compared with the state-of-the-art methods. The Yale face data set consists of 165 face images from 15 subjects. Each subject contains 11 images taken under the variations of different facial expressions and lighting conditions. The original face images have the size of 243×320 pixels with a 256-level gray scale. In this experiment, we aligned the face images such that their eyes are in the similar positions, and then cropped the face images and down-sampled them into the size of 64×64 pixels. Moreover, to alleviate the influences of the lighting condition on the face images, the histogram equalization operation is applied in advance to each face image. Fig. 2(a) shows the 11 face images of one subject. We use fivefold cross-validation strategy [1] to evaluate the recognition performance of the methods. In this method, the whole data set is partitioned into five subsets with approximately equal size of samples. Then, one of the subset is chosen as the testing data set and the other four ones as the training data set. This procedure is repeated until each subset has been used once as the testing data. To evaluate the

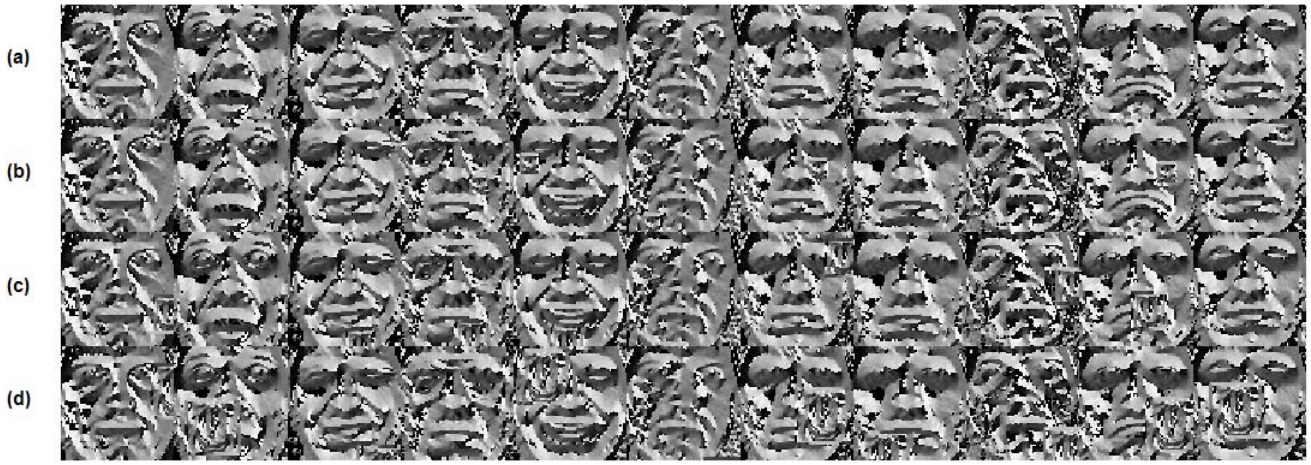


Fig. 4. 44 gradient orientation images corresponding to the 44 face images shown in Fig. 2. (a) The 11 gradient orientation images corresponding to the 11 clean face images in Fig. 2(a). (b) The 11 gradient orientation images corresponding to the 11 face images with 10×10 occlusions in Fig. 2(b). (c) The 11 gradient orientation images corresponding to the 11 face images with 20×20 occlusions in Fig. 2(c). (d) The 11 gradient orientation images corresponding to the 11 face images with 30×30 occlusions in Fig. 2(d).

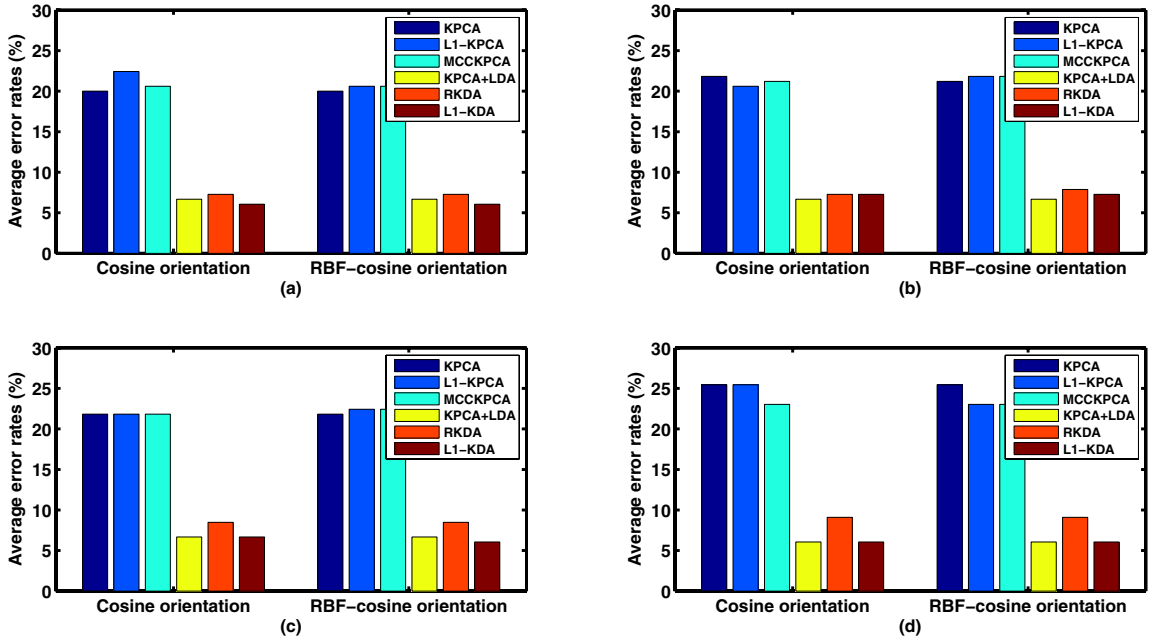


Fig. 5. Comparisons of the average test error rates among the various feature extraction methods on the gradient orientation images of Yale face data set, where (a)–(d) are the experimental results correspond to different face occlusions shown in Fig. 2.

robustness of the proposed method against outliers, in each trial of experiment we use outliers to contaminate the training face images and then use the clean testing face images to evaluate the recognition performance. In the experiments, the outliers are simulated by using three baboon images with the image size of 10×10 pixels, 20×20 pixels, and 30×30 pixels to, respectively, occlude the training face images, where the positions of the occlusions are randomly selected in each training face image. Fig. 2(b), (c), and (d) shows several examples of face images occluded by the baboon images with the size of 10×10 pixels, 20×20 pixels, and 30×30 pixels, respectively.

To conduct the experiments, each intensity face image is concatenated into a 4096-dimensional vector. Then, the monomial kernel with degree one, denoted by $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$, and the RBF kernel with parameter σ, denoted by $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma\right\}$, are, respectively, used to calculate the Gram matrix of the various methods. For RBF kernel, the parameter σ is empirically set to be $\sigma = 1e9$. For KPCA and L1-KPCA, the number of the projection vectors is set to be the one where the sum of variances exceeds 90% of the total variance. For MCCKPCA, KPCA + LDA, RKDA, and L1-KDA, the number of projection vectors is empirically set to be 45, 14, 14, and 14, respectively. In addition, the initial vector of solving each discriminant vector of L1-LDA was set to be the one of L2-norm LDA, i.e., $\omega^{(1)} = \arg\max_\omega \|\omega^T \mathbf{B}\|_2/\|\omega^T \mathbf{T}\|_2$, where $\|\cdot\|_2$ denotes the L2 norm operation of vector.

Fig. 3 summarizes the average test error rates (%) of the methods corresponding to the different kinds of image occlusions and the different kernel functions, where the results

TABLE II
UCI BENCHMARK DATA SETS AND THE AVERAGE ERROR RATES (%) OF SEVERAL METHODS

Data set	WBC	BUPA	PID	WDBC	CHD	SPECTFH	IRIS	TG	VC	MD
# samples	682	345	768	569	297	349	150	215	990	2000
# class	2	2	2	2	2	2	3	3	11	10
dimensionality	9	6	8	30	13	44	4	5	10	47
# PC	9	6	8	7	13	44	4	5	10	33
KPCA (D=1)	4.49	**37.43**	31.69	13.33	42.33	27.41	6.00	5.91	2.53	21.55
L1-KPCA (D=1)	4.35	37.71	30.26	14.39	42.33	25.56	6.00	4.09	2.73	21.75
MCCKPCA (D=1)	4.64	38.00	**28.05**	12.98	43.67	26.67	6.00	**3.64**	2.73	22.30
KPCA+LDA (D=1)	4.49	44.86	32.08	13.33	39.00	27.41	6.00	5.91	3.13	**20.75**
RKDA (D=1)	**4.35**	45.43	36.49	13.51	37.67	**22.22**	4.67	10.45	**1.01**	20.80
L1-KDA (D=1)	4.49	41.71	31.17	**5.79**	**23.33**	28.15	**2.67**	5.91	1.21	20.85
KPCA (D=2)	**3.62**	42.00	33.38	14.56	44.00	25.56	5.33	5.91	1.31	21.20
L1-KPCA (D=2)	4.06	40.00	31.82	14.39	48.33	27.41	5.33	6.82	**1.21**	21.25
MCCKPCA (D=2)	4.35	41.14	35.84	13.16	43.00	27.78	5.33	**5.00**	1.72	21.15
KPCA+LDA (D=2)	4.78	49.71	33.38	14.56	38.33	27.41	3.33	6.82	2.02	21.20
RKDA (D=2)	4.49	40.00	34.03	14.56	39.33	19.63	**5.33**	8.18	2.02	20.45
L1-KDA (D=2)	4.64	**34.00**	**29.35**	**8.42**	**24.67**	34.81	6.00	8.18	2.32	**17.95**
KPCA (D=3)	4.35	43.71	36.36	14.39	46.67	**25.19**	6.67	5.91	1.31	20.05
L1-KPCA (D=3)	4.64	45.43	34.81	14.21	46.67	25.93	5.33	5.00	1.72	20.10
MCCKPCA (D=3)	4.35	42.29	39.22	13.68	48.67	25.56	4.67	**5.00**	**0.09**	20.50
KPCA+LDA D=3)	**3.19**	46.00	36.36	14.39	48.00	23.33	6.00	7.27	2.22	21.40
RKDA (D=3)	4.20	40.29	37.79	13.16	39.33	20.74	7.33	9.55	2.63	20.00
L1-KDA (D=3)	5.22	**36.00**	**30.65**	**8.25**	**25.00**	40.37	**2.00**	10.00	3.13	**17.40**
KPCA (D=4)	4.35	38.29	37.14	13.51	53.00	24.81	6.67	6.36	1.82	20.55
L1-KPCA (D=4)	5.07	45.43	36.49	15.09	44.67	25.19	5.33	5.45	1.82	20.65
MCCKPCA (D=4)	4.93	49.71	41.04	13.68	48.67	24.07	7.33	**4.55**	**1.21**	21.60
KPCA+LDA D=4)	3.77	45.71	37.14	13.51	51.33	25.19	8.00	6.36	2.83	21.55
RKDA (D=4)	**3.48**	39.14	38.44	**11.58**	36.33	**22.96**	8.00	9.55	2.93	19.00
L1-KDA (D=4)	5.51	**35.14**	**35.06**	12.81	**27.67**	39.63	**4.00**	7.27	2.02	**17.60**

labeled with intensity and RBF-intensity are associated with the monomial kernel function and RBF kernel function, respectively. Fig. 3(a) shows the experimental results of the various methods on the face images without occlusions added into the training face images, whereas Fig. 3(b)–(d) shows the experimental results of the various methods in which the training face images with occlusions corresponding to Fig. 2(b)–(d), respectively. From Fig. 3, we can see that the supervised feature extraction methods (KPCA + LDA, RKDA, and L1-KDA) achieve better recognition results than the unsupervised methods (KPCA, L1-KPCA, and MCCKPCA). On the other hand, among the three supervised methods, the proposed L1-KDA method achieves the lowest error rates in most of experiments.

To further evaluate the recognition performance of the proposed L1-KDA method, we adopt the method of applying the image gradient orientation descriptor [12] to the Yale face data set, and then reconduct the same experiments as those of using the intensity images. Fig. 4 shows the 44 gradient orientation images corresponding to the 44 intensity face images shown in Fig. 2. Similar to [13], we use the cosine orientation kernel and the RBF-cosine orientation kernel, respectively, to calculate the Gram matrix in the experiments. Suppose that θ_i and θ_j are two orientation images corresponding to the intensity image $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Then, the cosine orientation kernel and the RBF-cosine orientation kernel are, respectively, defined as:

Cosine orientation kernel: $k(\theta_i, \theta_j) = \Sigma_{p=1}^{d} \cos(\theta_{ip} - \theta_{jp})/d$,

and RBF-cosine orientation kernel:

$$k(\theta_i, \theta_j) = \exp\left\{-\frac{1 - \sum_{p=1}^{d} \cos(\theta_{ip} - \theta_{jp})/d}{\sigma}\right\}$$

where d is the number of each face image pixels, θ_{ip} is the pth entry of θ_i, and θ_{jp} is the pth entry of θ_j. In the experiments, the RBF-cosine orientation kernel parameter σ is empirically set to be $\sigma = 49$. For KPCA and L1-KPCA, the number of the projection vectors is set to be the one where the sum of variances exceeds 90% of the total variance. For MCCKPCA, KPCA + LDA, RKDA, and L1-KDA, the number of projection vectors is empirically set to be 90, 14, 14, and 15, respectively.

Fig. 5 summarizes the average test error rates (%) of the various methods, in which Fig. 5(a) shows the experimental results without occlusions added into the training face images, and Fig. 5(b)–(d) shows the results with the occlusions corresponding to Fig. 2(b)–(d), respectively, added into the training face images. From Fig. 5, we can see that the experimental results coincide with those in Fig. 3, in which the supervised feature extraction methods achieve better recognition results

than the unsupervised methods, and the proposed L1-KDA method also demonstrates the lowest average error rates in most of experiments.

C. Experiment on UCI Data Sets

In this experiment, we use 10 UCI data sets [27] previously used in [28] to evaluate the discriminant performance of the proposed methods. These data sets are

1) Wisconsin breast cancer (WBC).
2) BUPA liver disorder (BUPA).
3) Pima indians diabetes (PID).
4) Wisconsin diagnostic breast cancer (WDBC).
5) Cleveland heart-disease (CH).
6) SPECTF heart (SPECTFH).
7) Iris plants (IRIS).
8) Thyroid gland (TG).
9) Vowel context (VC).
10) Multifeature digit (Zernike moments) (MD).

To evaluate the recognition performance of the various methods, the twofold cross-validation strategy [1] is adopted in the experiments. Similar to [28], before the experiments, PCA is firstly applied on the training set as a data preprocessing step, and then the various kernel based feature extraction methods are applied to evaluate the recognition performance of these methods. In the experiments, the monomial kernel function with degree D, denoted by

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^D$$

is used to calculate the Gram matrix. For KPCA and L1-KPCA, the number of the projection vectors is set to be the one when the sum of variances exceeds 90% of the total variance. For MCCKPCA, the number of projection vectors is set to be equal to that of KPCA. Table II summarizes the main properties of the 10 UCI data sets and the average error rates (%) of the various methods, where "#PC" in the fourth row shows the number of the principal components we use after the PCA preprocessing. From Table II, we can see that the proposed L1-KDA method achieves lower error rates (%) than the other methods for most UCI data sets, despite of the different choices of the monomial kernel degree D.

VI. Conclusion

In this paper, we have developed a novel L1-norm discriminant criterion under the rigorous theoretical framework of Bayes error bound. With the new discriminant criterion, we proposed the L1-LDA method for linear feature extraction. To efficiently solve the L1-LDA optimization problem, we also proposed an iterative algorithm in which a surrogate quadratic convex function is introduced such that a close-form solution can be obtained in each iteration. Moreover, we also proposed the L1-KDA method via kernel trick as a generalization of L1-LDA to cope with the robust nonlinear feature extraction problems. To evaluate the effectiveness of the proposed method, we conducted extensive experiments on both simulated data set and real data sets. The experimental results on simulated data set show that L1-LDA is superior over LDA in terms of robustness against outlier. For real data sets, we use Yale face database and UCI data sets to, respectively, test the performance of the proposed method. The experimental results on both Yale face database and UCI data sets show that the L1-KDA method achieves better discriminant performances than several state of the art kernel-based feature extraction methods. The experimental results confirm the superiority of using L1 norm over L2 norm in dealing with the feature extraction problem under the environment of outliers. This is mainly due to the more powerful ability of L1 norm in suppressing the effect of outliers than L2 norm.

Although the L1 norm is mainly used to deal with the robust feature extraction problems as for the case of outliers, it is notable that the use of L1 norm can also be used for other applications such as feature selection. In [23], [30], [31], and [32], the authors successfully used an L1-norm penalty to obtain sparse projection vectors for both feature extraction and feature selection. The research on simultaneously dealing with robust feature extraction and feature selection using L1-norm-based discriminant analysis approaches would be our future work.

Appendix A

Derivations of (5): From (2), we obtain that

$$\begin{aligned}\varepsilon(\omega) &\le \sum_{i<j} \int \sqrt{P_i p_i(\tilde{x}) P_j p_j(\tilde{x})}\, d\tilde{x} \\ &= \sum_{i<j} \sqrt{P_i P_j} \int \sqrt{p_i(\tilde{x}) p_j(\tilde{x})}\, d\tilde{x}. \end{aligned} \tag{35}$$

Since $p_i(\tilde{x})$ is the Gaussian function, it can be expressed as

$$p_i(\tilde{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\tilde{x} - \tilde{m}_i)^2\right\}.$$

Hence, we obtain that

$$\begin{aligned}&\sqrt{p_i(\tilde{x}) p_j(\tilde{x})} \\ &\quad= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{4\sigma^2}\left[(\tilde{x} - \tilde{m}_i)^2 + (\tilde{x} - \tilde{m}_j)^2\right]\right\} \\ &\quad= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(x - \frac{\tilde{m}_i + \tilde{m}_j}{2}\right)^2 \right.\right. \\ &\qquad\qquad \left.\left. + \left(\frac{\tilde{m}_i - \tilde{m}_j}{2}\right)^2\right]\right\} \\ &\quad= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left(x - \frac{\tilde{m}_i + \tilde{m}_j}{2}\right)^2\right\} \\ &\qquad\times \exp\left\{-\frac{(\tilde{m}_i - \tilde{m}_j)^2}{8\sigma^2}\right\}.\end{aligned}$$

From the fact that

$$\int \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left(x - \frac{\tilde{m}_i + \tilde{m}_j}{2}\right)^2\right\} d\tilde{x} = 1$$

we obtain that

$$\int \sqrt{p_i(\tilde{x}) p_j(\tilde{x})}\, d\tilde{x} = \exp\left\{-\frac{(\tilde{m}_i - \tilde{m}_j)^2}{8\sigma^2}\right\}. \tag{36}$$

Substituting the expression of (36) into (35), we obtain that

$$\varepsilon(\omega) \le \sum_{i<j} \sqrt{P_i P_j} \exp\left\{-\frac{1}{8}\left(\frac{\tilde{m}_i - \tilde{m}_j}{\sigma}\right)^2\right\}$$
$$= \sum_{i<j} \sqrt{P_i P_j} \exp\left\{-\frac{1}{8}\left[\frac{\omega^T(\mathbf{m}_i - \mathbf{m}_j)}{\frac{1}{N}\sum_{k=1}^{N}|\omega^T\mathbf{x}_k - \omega^T\mathbf{m}_{l_k}|}\right]^2\right\}$$

here we utilize the equalities $\tilde{m}_i = \omega^T\mathbf{m}_i$, $\tilde{m}_j = \omega^T\mathbf{m}_j$, and $\sigma = 1/N\sum_{k=1}^{N}|\omega^T\mathbf{x}_k - \omega^T\mathbf{m}_{l_k}|$.

APPENDIX B

Proof of theorem 1: From the definition of $\omega^{(p+1)}$ in (16), we have that

$$\sum_{j=1}^{c} s_j^{(p)} {\omega^{(p+1)}}^T \mathbf{b}_j = 1. \tag{37}$$

Let

$$J(\omega) = \frac{1}{2}\omega^T\mathbf{V}_t(\omega^{(p)})\omega + \frac{1}{2}\|\omega^{(p)T}\mathbf{T}\|_1.$$

Then, from the physical meaning of $\omega^{(p+1)}$, we have that

$$J(\omega^{(p+1)}) = \frac{1}{2}\sum_{i=1}^{N}\frac{\left[{\omega^{(p+1)}}^T\mathbf{t}_i\right]^2}{|{\omega^{(p)}}^T\mathbf{t}_i|} + \frac{1}{2}\|\omega^{(p)T}\mathbf{T}\|_1$$
$$\le J(\omega^{(p)}) = \sum_{i=1}^{N}|{\omega^{(p)}}^T\mathbf{t}_i| \tag{38}$$

On the other hand, from Lemma 2, we have that

$$J(\omega^{(p+1)}) = \frac{1}{2}\sum_{i=1}^{N}\frac{\left[{\omega^{(p+1)}}^T\mathbf{t}_i\right]^2}{|{\omega^{(p)}}^T\mathbf{t}_i|} + \frac{1}{2}\|\omega^{(p)T}\mathbf{T}\|_1$$
$$\ge \frac{1}{2}\sum_{i=1}^{N}\frac{\left[{\omega^{(p+1)}}^T\mathbf{t}_i\right]^2}{|{\omega^{(p+1)}}^T\mathbf{t}_i|} + \frac{1}{2}\|\omega^{(p+1)T}\mathbf{T}\|_1$$
$$= \sum_{i=1}^{N}|{\omega^{(p+1)}}^T\mathbf{t}_i|. \tag{39}$$

Combining (38) and (39), we have that

$$\sum_{i=1}^{N}|{\omega^{(p)}}^T\mathbf{t}_i| \ge \sum_{i=1}^{N}|{\omega^{(p+1)}}^T\mathbf{t}_i|. \tag{40}$$

From (40) and (37), we obtain that

$$Q(\omega^{(p+1)}) = \frac{\sum_{j=1}^{c}|{\omega^{(p+1)}}^T\mathbf{b}_j|}{\sum_{j=1}^{N}|{\omega^{(p+1)}}^T\mathbf{t}_j|}$$
$$\ge \frac{\sum_{j=1}^{c}s_j^{(p)}{\omega^{(p+1)}}^T\mathbf{b}_j}{\sum_{j=1}^{N}|{\omega^{(p+1)}}^T\mathbf{t}_j|} = \frac{1}{\sum_{j=1}^{N}|{\omega^{(p+1)}}^T\mathbf{t}_j|}$$
$$\ge \frac{1}{\sum_{j=1}^{N}|{\omega^{(p)}}^T\mathbf{t}_j|}. \tag{41}$$

From the equality $\sum_{j=1}^{c}|{\omega^{(p)}}^T\mathbf{b}_j| = 1$, we have that

$$Q(\omega^{(p)}) = \frac{\sum_{j=1}^{c}|{\omega^{(p)}}^T\mathbf{b}_j|}{\sum_{j=1}^{N}|{\omega^{(p)}}^T\mathbf{t}_j|} = \frac{1}{\sum_{j=1}^{N}|{\omega^{(p)}}^T\mathbf{t}_j|}. \tag{42}$$

Combining (41) and (42), we have that

$$Q(\omega^{(p+1)}) \ge Q(\omega^{(p)}). \tag{43}$$

So, $\omega^{(p+1)}$ is better than $\omega^{(p)}$.

APPENDIX C

Proof of lemma 3: Since $\omega^T\mathbf{U}_{r-1} = \mathbf{0}$, $\mathbf{Q}_{r-1}\mathbf{R}_{r-1}$ is the QR decomposition of $\mathbf{U}_{r-1}$, and $\mathbf{R}_{r-1}$ is nonsingular, we obtain that

$$\omega^T\mathbf{Q}_{r-1} = \mathbf{0}. \tag{44}$$

Let $\mathbf{Q}_{r-1}^{\perp}$ be the complement basis of $\mathbf{Q}_{r-1}$ such that the matrix $\mathbf{Q} = (\mathbf{Q}_{r-1} \quad \mathbf{Q}_{r-1}^{\perp})$ is an orthogonal matrix. Then we have $\mathbf{Q}_{r-1}^{\perp}(\mathbf{Q}_{r-1}^{\perp})^T = \mathbf{I}_d - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T$ due to $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}_d$. From $\omega^T\mathbf{Q}_{r-1} = \mathbf{0}$, there exists a $\beta \in \mathbb{R}^{d-r+1}$ such that

$$\omega = \mathbf{Q}_{r-1}^{\perp}\beta. \tag{45}$$

On the other hand, $\text{rank}\{(\mathbf{Q}_{r-1}^{\perp})^T\} = d - r + 1$. Therefore, the columns of $(\mathbf{Q}_{r-1}^{\perp})^T$ form a basis of $\mathbb{R}^{d-r+1}$. So there exists a $\alpha \in \mathbb{R}^d$, such that

$$\beta = (\mathbf{Q}_{r-1}^{\perp})^T\alpha. \tag{46}$$

Combining (45) and (46), we obtain that

$$\omega = \mathbf{Q}_{r-1}^{\perp}(\mathbf{Q}_{r-1}^{\perp})^T\alpha = (\mathbf{I}_d - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)\alpha \tag{47}$$

APPENDIX D

Proof of (33): To show that

$$\text{L1-KDA} = \text{KPCA} + \text{L1-LDA}$$

we have to show the optimal solution of L1-KDA can be expressed as

$$\omega = \mathbf{W}_{\text{KPCA}}^{\Phi}\beta \tag{48}$$

where $\mathbf{W}_{\text{KPCA}}^{\Phi}$ is the transform matrix of KPCA, and β is the optimal solution of L1-LDA in the KPCA projection subspace.

Let $\mathbf{Y}$ be the transformed data of $\Phi(\mathbf{X})$ on the KPCA transform matrix, then we have

$$\mathbf{Y} = (\mathbf{W}_{\text{KPCA}}^{\Phi})^T\Phi(\mathbf{X}). \tag{49}$$

In the transformed subspace, the L1-LDA problem can be formulated by

$$\beta = \arg\max_{\beta}\frac{\left\|\beta^T\mathbf{B}_Y\right\|_1}{\left\|\beta^T\mathbf{T}_Y\right\|_1} \tag{50}$$

where $\mathbf{B}_Y$ and $\mathbf{T}_Y$ can be respectively expressed as

$$\mathbf{B}_Y = \mathbf{Y}\mathbf{N}_b, \quad \mathbf{T}_Y = \mathbf{Y}\mathbf{N}_t. \tag{51}$$

Let ψ be the optimal discriminant vector of KPCA + L1-LDA, then ψ can be expressed as

$$\psi = \mathbf{W}_{\text{KPCA}}^{\Phi}\beta. \tag{52}$$

On the other hand, since $\mathbf{W}_{\text{KPCA}}^{\Phi}$ is the KPCA transform matrix, it can be expressed as [21]

$$\mathbf{W}_{\text{KPCA}}^{\Phi} = \Phi(\mathbf{X})\mathbf{P} \tag{53}$$

where $\mathbf{P}$ is a coefficient matrix. Substituting (53) into (52) and (49), respectively, we obtain that

$$\psi = \Phi(\mathbf{X})\mathbf{P}\beta \tag{54}$$

$$\mathbf{Y} = \mathbf{P}^T(\Phi(\mathbf{X}))^T\Phi(\mathbf{X}) = \mathbf{P}^T\mathbf{K} \tag{55}$$

where $\mathbf{K}$ is the Gram matrix.

Finally, substituting (55) and (51) into (50), we obtain that

$$\beta = \arg\max_{\beta} \frac{\left\|\beta^T\mathbf{P}^T\mathbf{K}\mathbf{N}_b\right\|_1}{\left\|\beta^T\mathbf{P}^T\mathbf{K}\mathbf{N}_t\right\|_1}. \tag{56}$$

Let $\alpha = \mathbf{P}\beta$, then from (54) and (56), we obtain that

$$\psi = \Phi(\mathbf{X})\alpha \tag{57}$$

where

$$\alpha = \arg\max_{\alpha} \frac{\left\|\alpha^T\mathbf{K}\mathbf{N}_b\right\|_1}{\left\|\alpha^T\mathbf{K}\mathbf{N}_t\right\|_1}. \tag{58}$$

By comparing the optimization problems of (57) and (58) with those of (30) and (31), we obtain that ψ is also the optimal discriminant vector of L1-KDA.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York, NY, USA: Academic Press, 1990.

[2] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. PAMI*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[3] C. M. Bishop, *Pattern Recogntion and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[4] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, Jun. 2006, pp. 281–288.

[5] A. Baccini, P. Besse, and A. D. Falguerolles, *A L1-Norm PCA and a Heuristic Approach*, E. Diday, Y. Lechevalier, and P. Opitz, Eds. New York, NY, USA: Springer-Verlag, 1996, pp. 359–368.

[6] Q. Ke and T. Kanade, "Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 592–599.

[7] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[8] X. Li, Y. Pang, and Y. Yuan, "L1-norm-based 2DPCA," *IEEE Trans. Syst., Man, Cybern., Part B, Cybern.*, vol. 40, no. 4, pp. 1170–1175, Aug. 2009.

[9] Y. Pang, X. Li, and Y. Yuan, "Robust tensor analysis with L1-norm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 2, pp. 172–178, Feb. 2010.

[10] S.-J. Kim, A. Magnani, and S. P. Boyd, "Robust Fisher discriminant analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 659–666.

[11] D. Huang, R. S. Cabral, and F. D. Torre, "Robust regression," in *Proc. ECCV*, 2012, pp. 616–630.

[12] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Principal component analysis of image gradient orientations for face recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2011, pp. 553–558.

[13] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.

[14] R. He, B.-G. Hu, W.-S. Zheng, and X.-W. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1485–1494, Jun. 2011.

[15] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 653–662, Mar. 2012.

[16] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with L1-norm," *IEEE Trans. Cybern.*, Jul. 2013, doi: 10.1109/TCYB.2013.2273355.

[17] F. Zhong and J. Zhang, "Linear discriminant analysis based on L1-norm maximization," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3018–3027, Aug. 2013.

[18] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[19] J. Yang, A. F. Frangi, J.-Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.

[20] M. H. Yang, "Kernel eigenfaces vs. Kernel fisherfaces: Face recognition using kernel methods," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2002, pp. 215–220.

[21] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[22] G. Saon and M. Padmanabhan, "Minimum bayes error feature selection for continuous speech recognition," in *Proc. NIPS*, 2002, pp. 800–806.

[23] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proc. 13th Int. Conf. AISTATS*, 2010, pp. 1–8.

[24] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.

[25] W. Zheng, "Heteroscedastic feature extraction for texture classification," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 766–769, Sep. 2009.

[26] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[27] P. M. Murphy and D. W. Aha, (2004). *UCI Repository of Machine Learning Database* [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository

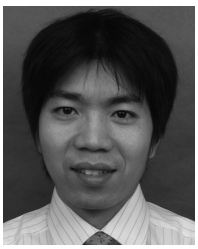
[28] M. Loog and R. P. W. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion," *IEEE Trans. PAMI*, vol. 26, no. 6, pp. 732–739, Jun. 2004.

[29] Z. Jin, J. Y. Yang, Z. S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognit.*, vol. 34, no. 7, pp. 1405–1416, 2001.

[30] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.

[31] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[32] Z. Lai, W. K. Wong, Z. Jin, J. Yang, and Y. Xu, "Sparse approximation to the eigensubspace for discrimination," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1948–1960, Dec. 2012.



Wenming Zheng (M'08) received the B.S. degree in computer science from Fuzhou University, Fuzhou, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, China, in 2004.

He has been with the Research Center for Learning Science, Southeast University, since 2004. Currently, he is a Professor with the Key Laboratory of Child Development and Learning Science of the Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing. His current research interests include neural computation, pattern recognition, machine learning, and computer vision.

Zhouchen Lin (SM'08) received the Ph.D. degree in applied mathematics from Peking University, Beijing, China, in 2000.

He is currently a Professor with the Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Peking University. He is a Chair Professor with Northeast Normal University, Changchun, China. In 2012, he was a Lead Researcher with the Visual Computing Group, Microsoft Research Asia. He was a Guest Professor with Shanghai Jiaotong University, Shanghai, China, Beijing Jiao Tong University, Beijing, and Southeast University, Nanjing, China. He was a Guest Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His current research interests include computer vision, image processing, computer graphics, machine learning, pattern recognition, and numerical computation and optimization.

Dr. Lin is an Associate Editor of the *International Journal of Computer Vision*.

Haixian Wang (M'09) was born in Anhui, China, in 1977. He received the B.S. and M.S. degrees in statistics and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999, 2002, and 2005, respectively.

He was with the Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education of China, Hefei, from 2002 to 2005. Currently, he is with the Key Laboratory of Child Development and Learning Science of the Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing, China. His current research interests include statistical pattern recognition, machine learning, and electroencephalograms signal processing.