Fisher Discriminative Least Square Regression with Self-Adaptive Weighting for Face Recognition

Zhe Chen, Xiao-Jun Wu, and Josef Kittler, Life Fellow, IEEE

Abstract—As a supervised classification method, least square regression (LSR) has shown promising performance in multiclass face recognition tasks. However, the latest LSR based classification methods mainly focus on learning a relaxed regression target to replace traditional zero-one label matrix while ignoring the discriminability of transformed features. Based on the assumption that the transformed features of samples from the same class have similar structure while those of samples from different classes are uncorrelated, in this paper we propose a novel discriminative LSR method based on the Fisher discrimination criterion (FDLSR), where the projected features have small within-class scatter and large inter-class scatter simultaneously. Moreover, different from other methods, we explore relative regression from the view of transformed features rather than the regression targets. Specifically, we impose a dynamic non-negative weight matrix on the transformed features to enlarge the margin between the true and the false classes by self-adaptively assigning appropriate weights to different features. Above two factors can encourage the learned transformation for regression to be more discriminative and thus achieving better classification performance. Extensive experiments on various databases demonstrate that the proposed FDLSR method achieves superior performance to other state-of-the-art LSR based classification methods.

Index Terms—Face recognition, Least square regression, Fisher discrimination criterion, Self-adaptively weighting, Multiclass classification.

1 INTRODUCTION

Due to its effectiveness and simplicity for classification, least square regression (LSR) technique has been widely used in the field of pattern recognition and analysis, such as face recognition [1], feature selection [2] and image retrieval [3], etc. In the past few years, researchers proposed various improvements in LSR, including weighted LSR [4], ridge regression [5], LASSO regression [6], logistic regression [7], SVM [8], least-square SVM [9], local LSR [10], partial LSR [11], kernel LSR [12], and so on. In addition, linear regression based classification (LRC) [13], sparse representation based classification (SRC) [14] and collaborative representation based classification (CRC) [15] are also LSR model based methods, the difference lies in their representations used for classification.

Let $X = [X_1, X_2,...,X_C] = [x_1, x_2,...,x_n] \in R^{d \times n}$ be the training set from $C$ classes and the corresponding label matrix $H = [h_1, h_2,...,h_n] \in R^{C \times n}$, where $d$ and $n$ denote the sample dimensionality and the number of samples, respectively. $X_i$ denotes the sample matrix of the $i$th class. Suppose sample $x_i \in R^d$ belongs to class $j$, then the $j$th value in its corresponding label vector $h_i \in R^C$ is 1, while the others in $h_i$ are all 0s. The LSR model can be formulated as

$$\min_Q \|QX - H\|_F^2 + \lambda \|Q\|_F^2$$  (1)

where $Q \in R^{c \times d}$ is the transformation matrix, $\lambda$ is the positive regularization parameter. $QX$ denotes the transformed features of original training samples $X$. LSR aims to minimize the least square loss between the transformed features $QX$ and the predefined regression target $H$. Problem (1) has a closed-form solution

$$Q = HXT(XX^T + \lambda I)^{-1}$$  (2)

where $XT$ is the transposed matrix of $X$. When given a test sample $y \in R^d$, LSR predicts its label by calculating $max_i(Qy)_i$, where $(Qy)_i$ is the $i$th element of $Qy$. Besides, for high-dimensional data, LRLR [16] uses the low-rank constraint to replace the $l_F$ norm constraint in LSR as follows

$$\min_Q \|QX - H\|_F^2 + \lambda \|Q\|_*$$  (3)

where $\|Q\|_*$ is the nuclear norm (the sum of matrix singular values) of $Q$. LRLR can achieve better performance by excavating low-rank structural information of data.

However, LSR model has many drawbacks in view of face recognition. The main one is that the strict zero-one regression target matrix $H$ is not appropriate for classification. We can find the distances between regression targets of any pair of the samples from different classes are equal to $\sqrt{2}$ fixedly. This is obviously inconsistent with our expectation that the transformed features of inter-class samples should be as uncorrelated as possible. To solve this problem, DLSR
proposes a technique called \( \varepsilon \)-dragging to encourage the inter-class regression targets changing along opposite directions. The objective function of DLSR is

\[
\min_{Q, M, b} \|QX - (H + B \circ M) - be_n\|_F^2 + \lambda \|Q\|_F^2 \quad \text{s.t. } M \succeq 0 \tag{4}
\]

where \( b \in \mathbb{R}^{n \times 1} \) is an offset vector which need to be learned, \( e_n = [1, 1, \ldots, 1] \in \mathbb{R}^{1 \times n} \) is an all 1s row vector. \( M \geq 0 \) forces all elements in \( M \) to be non-negative. \( \circ \) denotes the Hadamard-product operator. \( B \) is a luxury matrix, which is defined as

\[
B_{ij} = \begin{cases} 
+1, & \text{if } H_{ij} = 1 \\
-1, & \text{if } H_{ij} = 0 
\end{cases} \tag{5}
\]

By adding the relaxation term \( B \odot M \) on the original label matrix \( H \), the new regression target becomes \( H' = H + B \odot M \) and its distances between inter-class regression targets are larger than \( \sqrt{2} \).

Based on the manifold learning methods, regularized label relaxation RLRLR [18] method introduces a class compactness graph structure on DLSR as follows

\[
\min_{Q, M} \|QX - (H + B \odot M)\|_F^2 + \lambda \text{tr}(QXLXTQ^T) \quad \text{s.t. } M \succeq 0 \tag{6}
\]

where \( L = Z - W \) is the Laplacian matrix. \( Z \) is a diagonal matrix and its diagonal elements \( Z_{ii} = \sum_j W_{ij} \). \( W \) is the weight matrix of compactness graph which is defined as

\[
W_{ij} = \begin{cases} 
e^{-\frac{|x_i - x_j|^2}{\sigma}}, & \text{if sample } x_i \text{ and } x_j \text{ are from the same class;} \\
0, & \text{otherwise} \tag{7}
\end{cases}
\]

where \( \sigma \) is the heat kernel parameter. The use of class compactness graph can ensure the transformed features of within-class samples to be as close as possible. Thus RLRLR can not only relax strict zero-one regression target into a slack space but also avoid the overfitting problem.

In order to learn more flexible targets, retargeted LSR (ReLSR) [19] method proposes to directly learn slack regression targets from original data. The ReLSR model is defined as

\[
\min_{Q, M, a} \|T - QX - be_n\|_F^2 + \lambda \|Q\|_F^2 \quad \text{s.t. } r_{ij} - \max_{i \neq j} T_{i,j} \geq 1 \tag{8}
\]

where \( T \in \mathbb{R}^{C \times n} \) is the relaxed regression target matrix which need to be learned from data \( X \) in the optimization process. \( r_j \) indicates the true label of sample \( x_j \), i.e., if \( x_j \) belongs to class \( l \), then \( r_j = l \). ReLSR can guarantee samples being correctly classified with large margins (should be larger than 1) and thus the class separability of each data is enhanced.

Recently, groupwise ReLSR (GReLSR) [20] method proves that DLSR is essentially a special case of ReLSR and a new model is developed as follows

\[
\min_{Q, b, M, a} \|QX - (Y + Y \odot M) + be_n - c_e^T a\|_F^2 + \lambda \|Q\|_F^2 + \gamma R(a) \quad \text{s.t. } M \succeq 0, \{M_{ri,:} = 0\}_{i=1}^n \tag{9}
\]

where \( Y = 2H - 1, 1 \) is the all 1s matrix. \( U \in \mathbb{R}^{C \times n} \) is the non-negative slack matrix. \( e_C = [1, 1, \ldots, 1] \in \mathbb{R}^{1 \times C} \) is an all 1s row vector. \( a \) and \( b \) are two offset vectors. In GReLSR, the translation values \( \{a_i\}_{i=1}^n \) of all samples (each column corresponds to a sample) are considered independently. In order to avoid the regression targets of within-class samples are markedly different, the values in \( a \) for the within-class samples should be similar. Thus GReLSR introduces a groupwise constraint term \( R(a) \) which is defined as

\[
R(a) = \sum_{j=1}^C \sum_{i \in S_j} (a_j - \mu_j)^2 \tag{10}
\]

where \( S_j \) consists of the indexes of the samples belonging to the \( j \)th class. By combining the groupwise regularization term \( R(a) \) with ReLSR, GReLSR model will be more suitable for multiclassification. More recently, similar to GReLSR, CLSR [21] proposes to force the within-class samples to pursue the similar soft targets by constructing a discrete label matrix and an auxiliary matrix.

From [22] we know DLSR can be regarded as a special case of ReLSR and GReLSR, which means that all of them (including RLRLR) use the \( \varepsilon \)-dragging technique to learn relaxed regression targets. However, the \( \varepsilon \)-dragging technique can also enlarge the distances of within-class regression targets which is bad for classification. Moreover, they ignore that the discriminability of transformed features is also important for improving the classification performance. In this paper, a new discriminative LSR method named Fisher discrimination criterion based LSR (FDLSR) is proposed for face recognition problem. Specifically, the contributions of FDLSR are presented as follows.

1. By adding a Fisher discrimination term on original LSR model, FDLSR can enhance the discriminability of the learned transformation. The within-class similarity and the inter-class independence of transformed features can be guaranteed by simultaneously minimizing within-class scatter and maximizing inter-class scatter.

2. A non-negative weight matrix is imposed on the transformed features to self-adaptively enlarge the margin between the true and the false classes which can further improve the robustness and discriminability of transformation.

3. A new classification method is proposed. Since the learned features are discriminative adequately, we utilize these features to perform classification. After obtaining the transformation matrix, the nearest neighbor classifier is performed between the mean feature of each class and the transformed feature of test sample, which can effectively avoid misclassification caused by those outliers.

2 The proposed FDLSR framework

In this section, we first present a novel Fisher discrimination LSR (FDLSR) model, in which the Fisher discrimination criterion and the non-negative self-adaptive weighting technique are used to improve the discriminability of transformed features. Then we will present the optimization method of the FDLSR model. At last, an efficient classification approach will be introduced.

2.1 Motivations and framework of FDLSR

All of aforementioned LSR based classification methods try to improve the performance by relaxing the regression pattern. However, the relaxation technique increases the possibility of non-classification which can lead to worse performance. To solve this problem, we introduce a Fisher discrimination term into the objective function of LSR.

\[
R(a) = \sum_{j=1}^C \sum_{i \in S_j} (a_j - \mu_j)^2 \tag{10}
\]
targets. However, the used \( \varepsilon \)-dragging technique can also result in large distances of within-class regression targets, and excessively fitting the largest margins along opposite directions will no doubt lead to the problem of over-fitting. As mentioned previously, exploiting the discriminability among transformed features is helpful to learn a compact and discriminative transformation matrix. In fact, strict zero-one regression targets usually include enough discriminative information. Thus in this paper we focus on how to enhance the discriminability from the view of transformed features instead of learning slack regression targets. Inspired by [22] and [23], we propose the following Fisher discriminative least square regression (FDLSR) framework

\[
\begin{equation} \min_{Q,W} \frac{1}{2} \left\{ \|QX - H\|_F^2 + \alpha \|Q\|_F^2 + \beta \|W^{1/2} \odot (QX)\|_F^2 + \gamma \|W\|_F^2 + \lambda f(QX) \right\} \quad \text{s.t. } W \geq 0, \ W^T1 = 1 \tag{11} \end{equation}
\]

where \( \alpha, \beta, \gamma \) and \( \lambda \) are the regularization parameters. \( Q \) is the transformation matrix and \( QX \) is the transformed features. \( W \) is the weighting matrix with non-negative values of all elements. \( W^{1/2} \) denotes the element-wise square root of \( W \). By adding the non-negative weighted matrix on the transformed features \( QX \), the method will self-adaptively assign smaller weight to the feature with larger value (true class) and assign larger weight to the feature with smaller value (false class) which naturally enlarges the margin between the true and the false classes, thus improving the classification performance. The constraint term \( W^T1 = 1 \) can guarantee all features to be treated equally. What’s more, restraining the value of \( W \) in an appropriate range by minimizing term \( \|W\|_F^2 \) and constraining \( W \geq 0, W^T1 = 1 \) can prevent \( W \) obtaining trivial solution [24].

\( f(QX) \) is the Fisher-based discriminative features regularization function, which is defined as

\[
f(QX) = \sum_{k=1}^{C} (\|QX\|_k^2 - \|M_k - M\|_k^2 + \|QX\|_k^2) + \|QX\|_F^2 \tag{12}\]

where \( (QX)_k \in \mathbb{R}^{n_k} \) is the transformed features of the samples belong to the \( k \)th class. Supposing \( m_k \) and \( m \) are the mean vectors of \( (QX)_k \) and \( QX \), then \( M_k \) and \( M \) are the mean matrices by taking \( n_k \) mean vectors \( m_k \) and \( m \) as its column vectors, respectively. \( \|QX\|_F^2 \) is an elastic term used to make function \( f(QX) \) stable and convex. By minimizing \( f(QX) \), the transformed features will have small within-class scatter and large between-class scatter simultaneously. Both of the similarity of extracted within-class features and the dissimilarity of between-class features can be guaranteed, thus encouraging the learned transformation matrix to be more discriminative. Fig. 1 shows the structure of the proposed FDLSR model.

Actually, both of the non-negative dynamic weighting term and the Fisher discrimination function can improve the discriminability of learned transformation. The former aims to enlarge the class margins between true and false classes, and the latter aims to learn similar within-class features and uncorrelated between-class features simultaneously. All the pursuant factors are beneficial to subsequent classification tasks.

### 2.2 Solution to FDLSR

There are two variables \( Q \) and \( W \) need to be solved in Eq. (11). Obviously, it is impossible to directly obtain their closed-form solutions. In this section, we use the alternating
direction multipliers method (ADMM) to solve the optimization problem of FDLSR. We first introduce two auxiliary variables \( P \) and \( T \) to make problem (11) separable as follows

\[
\min_{Q,W,P,T} \frac{1}{2} \left( \| P - H \|_F^2 + \alpha \| Q \|_F^2 + \frac{\beta}{2} \| W^{1/2} \odot J \|_F^2 + \gamma \| W \|_F^2 + \lambda f(P) \right) \quad \text{s.t. } QX = P, \; QX = J, \; W \preceq 0, \; W^T 1 = 1
\]  

Then we reformulate Eq. (13) into the following augmented Lagrangian function

\[
L(Q, W, P, J, Y_1, Y_2) = \frac{1}{2} \left( \| P - H \|_F^2 + \alpha \| Q \|_F^2 + \frac{\beta}{2} \| W^{1/2} \odot J \|_F^2 + \gamma \| W \|_F^2 + \lambda f(P) \right) + \frac{\mu}{2} \left( \| QX - P + \frac{Y_1}{\mu} \|_F^2 + \| QX - J + \frac{Y_2}{\mu} \|_F^2 \right)
\]  

where \( Y_1 \) and \( Y_2 \) are the Lagrangian multipliers, \( \mu > 0 \) is the penalty parameter. Then we can alternately update each variable with the others fixed.

**Step 1. Update Q:** By fixing variables \( P, J \) and \( W, Q \), can be obtained by minimizing the following problem

\[
L(Q) = \frac{\alpha}{2} \| Q \|_F^2 + \frac{\mu}{2} \left( \| QX - P + \frac{Y_1}{\mu} \|_F^2 + \| QX - J + \frac{Y_2}{\mu} \|_F^2 \right)
\]

We set the derivation of \( L(Q) \) with respect to \( Q \) to zero

\[
\frac{\partial L(Q)}{\partial Q} = \alpha Q + 2\mu QXX^T - \mu PX^T + Y_1 X^T - \mu JX^T + Y_2 X^T = 0
\]

\( Q \) has a closed-form solution as

\[
Q = (\alpha P - Y_1 + \mu J - Y_2) X^T (\alpha I + 2\mu XX^T)^{-1}
\]

**Step 2. Update P:** By fixing variables \( Q, J \) and \( W, P \), can be obtained by minimizing the following problem

\[
L(P) = \frac{1}{2} \| P - H \|_F^2 + \frac{\lambda}{2} f(P) + \frac{\mu}{2} \| QX - P + \frac{Y_1}{\mu} \|_F^2
\]

Referring to literature \([26]\), the derivation of \( \frac{1}{2} f(P) \) with respect to \( P \) is

\[
\frac{\partial \frac{1}{2} f(P)}{\partial P} = 2P + M - 2\tilde{M}
\]

where \( \tilde{M} = [M_1, M_2, ..., M_C] \). Then we set the derivation of \( L(P) \) with respect to \( P \) to zero

\[
\frac{\partial L(P)}{\partial P} = P - H + 2\lambda P + \lambda M - 2\tilde{M} - \mu QX + \mu P - Y_1 = 0
\]

Likewise, \( P \) has a closed solution as

\[
P = \frac{H - \lambda M + 2\lambda \tilde{M} + \mu QX + Y_1}{1 + 2\lambda + \mu}
\]

**Step 3. Update J:** By fixing variables \( Q, P \) and \( W, J \), can be obtained by minimizing the following problem

\[
L(J) = \frac{\beta}{2} \| W^{1/2} \odot J \|_F^2 + \frac{\mu}{2} \| QX - J + \frac{Y_2}{\mu} \|_F^2
\]

Define \( G = QX + \frac{Y_2}{\mu} \), we can reformulate Eq. (22) as

\[
L(J) = \| W^{1/2} \odot J \|_F^2 + \frac{\mu}{\beta} \| J - G \|_F^2
\]

Similar to \([23]\), we can element-wisely update variable \( J \) as follows

\[
\min_{j} \sum_{i=1}^{n} \min_{j} \left( \frac{\beta W_{ij} + \mu}{\beta} J_{ij}^2 - \frac{2\mu}{\beta} J_{ij} G_{ij} \right)
\]

From Eq. (24), the optimal solution to each element of \( J \) is

\[
J_{ij} = \frac{\mu G_{ij}}{\beta W_{ij} + \mu}
\]

**Step 4. Update W:** By fixing variables \( Q, P \) and \( J \), under the condition that \( W \geq 0 \), \( W^T 1 = 1 \), \( W \) can be obtained by minimizing the following problem

\[
L(W) = \frac{\beta}{2} \| W^{1/2} \odot J \|_F^2 + \frac{\gamma}{2} \| W \|_F^2
\]

We first rewrite Eq. (26) as follows

\[
L(W) = \| W^{1/2} \odot J \|_F^2 + \frac{\gamma}{2} \| W \|_F^2
\]

To optimize Eq. (27) is equivalent to the following minimization problem

\[
\min_{W \geq 0, W^T 1 = 1} \sum_{i=1}^{n} \sum_{j=1}^{n} (W_{ij} J_{ij}^2 + \frac{\gamma}{2} W_{ij}^2)
\]

As the sum of the weights with respect to one sample is forced to 1, we can column-wisely solve minimization problem (28) as follows

\[
\min_{W_{j} \geq 0, W_j^T 1 = 1} \sum_{j=1}^{n} \| W_{j} + \frac{\beta}{\gamma} B_{j} \|_F^2
\]

where \( B_{j} \) is the \( j \)th column of \( B = J \odot J \). Then we give the Lagrangian function of Eq. (29) as

\[
L(W_j, \eta_j, \tau_j) = \frac{1}{2} \| W_{j} + \frac{\beta}{\gamma} B_{j} \|_F^2 - \eta_j (W_{j}^T 1 - 1) - \tau_j W_j
\]
where $\eta_j$ and $\tau_j > 0$ are the Lagrangian multipliers. Setting the derivation of Eq. (30) with respect to $W_j$ to zero, we have

$$
\frac{\partial L(W_j, \eta_j, \tau_j)}{\partial W_j} = W_j + \frac{\beta}{2\gamma} B_j - 1\eta_j - \tau_j = 0 \quad (31)
$$

According to [27], we know $\tau_j \odot W_j = 0$ (KKT condition). So we have

$$
W_j = \max(1\eta_j - \frac{\beta}{2\gamma} B_j, 0) \quad (32)
$$

And because $W_j^T 1 = 1$, we have

$$
\sum_{i=1}^{C} (\eta_j - \frac{\beta}{2\gamma} B_{ij}) = 1
$$

$$
\Rightarrow \eta_j = \frac{1}{C} (1 + \frac{\beta}{2\gamma} \sum_{i=1}^{C} B_{ij}) \quad (33)
$$

Once we obtain $\eta_j$, we use Eq. (32) to update $W_j$. Finally, optimal $W$ is obtained.

**Step 5. Update $Y_1$, $Y_2$ and $\mu$:** After updating variables $Q$, $P$, $J$ and $W$, we update Lagrangian multipliers and penalty parameter as follows

$$
Y_1 = Y_1 + \mu (QX - P) \quad (34)
$$

$$
Y_2 = Y_2 + \mu (QX - J) \quad (35)
$$

$$
\mu = \min(\mu_{\text{max}}, \rho \mu) \quad (36)
$$

The detail optimization steps of FDLSR are summarized in Algorithm 1.

---

**Algorithm 1. Solving FDLSR by ADMM**

**Input:** Normalized training samples $X$ and its label matrix $H$; Parameters $\alpha$, $\beta$, $\gamma$, $\lambda$.

**Initialization:** $Q = P = J = W = 0$, $Y_1 = Y_2 = 0$, $\mu_{\text{max}} = 10^8$, $tol = 10^{-6}$, $\mu = 10^{-5}$, $\rho = 1.1$.

**While** not converged do:

1. Update $Q$ by using Eq. (17).
2. Update $P$ by using Eq. (21).
3. Update $J$ by using Eq. (25).
4. Update $W$ by using Eq. (32).
5. Update Lagrange multipliers $Y_1$ and $Y_2$ by using Eq. (34) and Eq. (35).
6. Update penalty parameter $\mu$ by using Eq. (36).
7. Check convergence:

   $$
   \text{if max} \{\|QX - P\|_{\infty}, \|QX - J\|_{\infty}\} \leq tol,
   $$

**End While**

**Output:** $Q, W$

---

### 2.3 Convergence Analysis of FDLSR

In section 2.2, we used an ADMM to solve the FDLSR model. In order to analyze the convergence of our optimization algorithm, we first denote the objective function Eq. (13) as $\Psi(Q, W, P, J)$.

**Lemma 1:** The Algorithm 1 monotonically decreases the value of $\Psi(Q, W, P, J)$.

**Proof:** We denote the value of the objective function Eq. (13) at the $t$th iteration as $\Psi(Q^t, W^t, P^t, J^t)$. During the $(t+1)$th iteration, we first fix variables $W^t$, $P^t$, and $J^t$, and solve the following subproblem

$$
\min_{Q} \Psi(Q, W^t, P^t, J^t) \quad (37)
$$

The optimal solution is $Q^{t+1}$. Since above subproblem is quadratic and convex, we have

$$
\Psi(Q^{t+1}, W^t, P^t, J^t) \leq \Psi(Q^t, W^t, P^t, J^t) \quad (38)
$$

Then we fix variables $Q^{t+1}$, $W^t$, and $J^t$, and solve the following subproblem

$$
\min_{P} \Psi(Q^{t+1}, W^t, P, J^t) \quad (39)
$$

We can find problem (39) is still convex with respect to variable $P$. With the optimal solution $P^{t+1}$, we have

$$
\Psi(Q^{t+1}, W^t, P^{t+1}, J^t) \leq \Psi(Q^{t+1}, W^t, P^t, J^t) \quad (40)
$$

Likewise, we fix variables $Q^{t+1}$, $P^{t+1}$, and $J^t$, and solve the following optimization problem

$$
\min_{J} \Psi(Q^{t+1}, W^t, P^{t+1}, J) \quad (41)
$$

Due to the convexity of this subproblem, it follows that

$$
\Psi(Q^{t+1}, W^t, P^{t+1}, J^{t+1}) \leq \Psi(Q^{t+1}, W^t, P^{t+1}, J^t) \quad (42)
$$

Finally, we fix variables $Q^{t+1}$, $P^{t+1}$, and $J^{t+1}$, and solve the following optimization subproblem

$$
\min_{W} \Psi(Q^{t+1}, W, P^{t+1}, J^{t+1}) \quad (43)
$$

Above subproblem is also convex, so we have

$$
\Psi(Q^{t+1}, W^{t+1}, P^{t+1}, J^{t+1}) \leq \Psi(Q^{t+1}, W^t, P^{t+1}, J^{t+1}) \quad (44)
$$

By combining (38), (40), (42) and (44), we obtain

$$
\Psi(Q^{t+1}, W^{t+1}, P^{t+1}, J^{t+1}) \leq \Psi(Q^t, W^t, P^t, J^t) \quad (45)
$$

In summary, we conclude that Algorithm 1 can monotonically decrease the value of $\Psi(Q, W, P, J)$. Because Eq. (13) is a holistic convex framework, the ADMM used in Algorithm 1 can find the unique optimal solution of FDLSR.

### 2.4 Classification approach

After obtaining the optimal transformation matrix $Q$, traditional LSR model predicts the label of a new test sample $y \in \mathbb{R}^d$ by calculating

$$
\text{Label}(y) = \max_i (Qy)_i, \; i = 1, 2, ..., C
$$

where $Qy \in \mathbb{R}^C$ is the transformed features of test sample $y$. Recently, [28] takes advantage of the promising class sparsity of transformed features, i.e., $QX$, to perform the nearest neighbor classifier. Specifically, they first calculate the Euclidean distance between the transformed features of test sample, i.e., $Qy$, and the transformed features of each training sample, i.e., $Qx_i$, then the test sample will be classified into the class to which the nearest training sample belongs.

However, there often exists various noise in face samples that are bad for feature extraction. Therefore, the transformed features may include some outliers which can lead to misclassification. To solve this problem, we first
class-wisely calculate the mean vector of the transformed features, and then we class-wisely calculate the Euclidean distance between the mean vector and the transformed features of test sample. Finally, we classify the test sample to the class which has the smallest distance. Algorithm 2 shows the detail classification procedures.

Algorithm 2. Classification based on FDLSR

Input: Normalized training samples \( X \), learned transformation matrix \( Q \), test sample \( y \).

Output: Predicted label of \( y \).

Step 1. Calculate transformed features of training and test samples: \( \hat{X} = QX, \hat{y} = Qy \), and normalize them into a unit vector.

Step 2. Calculate the mean vector of \( X_i, i = 1, 2, \ldots, C \), where \( X_i \) is the transformed features of \( i \)th class.

Step 3. Calculate the Euclidean distance between the transformed features of \( X_i \) and \( \hat{y} \), and mean vectors of each class using transformed training features, then classify the test sample to the class which has the smallest distance.

3 EXPERIMENTS

In this section, we perform some experiments to verify the effectiveness of the proposed FDLSR method. We compare the proposed FDLSR method with some competing LSR based classification methods, including LSR, DLSR\[17\], ReLSR\[19\], GReLSR\[20\] and CLSR\[21\], which are fairly evaluated on a range of different databases.

3.1 Experimental results on five face databases

Five popular face databases, i.e., the ORL database\[29\], the Extended Yale B database\[30\], the AR database\[31\], the Labeled Faces in the Wild (LFW)\[32\] database, and the CMU PIE\[33\] are used to evaluate the classification performance of different methods. Examples of images from the five face databases are shown in Fig. 2. The detail introduction to these databases and the corresponding parameters setting are presented as follows

(1) The ORL database includes 400 face images of 40 persons, each person has 10 images. Each image was resized to the size of 28 x 23 in our experiments. The parameters \( \alpha, \beta, \gamma \) and \( \lambda \) in our method were set to 1e-1, 1e-2, 1e-1, and 1e1, respectively. We randomly select 3, 4, 5, and 6 images of each person as training samples, and the remaining images were treated as test samples.

(2) The Extended Yale B database consists of 2414 front-face images of 38 persons. Each person has about 59-64 images. Each image is resized to the size of 32 x 32 in our experiments. The parameters \( \alpha, \beta, \gamma \) and \( \lambda \) in our method are set to 1e-1, 1e-2, 1e-1, and 5e0, respectively. We randomly select 10, 15, 20, and 25 images of each person as training samples, and the remaining images are treated as test samples.

(3) The AR face database consists of over 4000 images of 126 persons. In our experiments, we choose a subset with 3120 images of 120 persons to conduct performance evaluation. Each person has 26 frontal face images. All the images are resized to the size of 50 x 40 in advance. The parameters \( \alpha, \beta, \gamma \) and \( \lambda \) in our method are set to 1e-1, 1e-2, 1e-1, and 1e1, respectively. We randomly select 3, 4, 5, and 6 images from each person as training samples and treat the remaining images as test samples.

(4) The LFW database has more than 13000 face images collected from the Internet, and all images are labeled with the name of the persons. Following\[28\], in this paper we use a subset of LFW which consists of 1251 images of 86 persons to conduct experiments. Each person has about 11-20 images. In our experiments, each image is resized to the size of 32 x 32. The parameters \( \alpha, \beta, \gamma \) and \( \lambda \) in our method are set to 1e-1, 1e-2, 1e-1, and 1e-1, respectively. We randomly select 5, 6, 7, and 8 images from each person as training samples and treat the remaining images as test samples.

(5) The CMU PIE face database\[33\] consists of 41386 front-face images of 68 persons. In our experiments, we use its five subsets (C05, C07,C09, C27 and C29) to conduct evaluation. Each person has about 170 face images and each image is manually resized to the size of 32 x 32. The parameters \( \alpha, \beta, \gamma \) and \( \lambda \) in our method are set to 1e-1, 1e-2, 1e-1, and 1e-1, respectively. We randomly select 5, 8, 10, and 15 images from each person as training samples and the remaining images are set as test samples.

For the comparison of LSR-based methods, i.e., DLSR, ReLSR, GReLSR, CLSR, and FDLSR, the optimal parameters are selected from the candidate set \{1e4, 1e-3, 1e-2, 1e-1, 1\}. To avoid any bias, for each group of experiments, all methods are repeated 10 times with the random selection of training and test samples, then we reported the mean value as the final classification result. The mean recognition rate of these methods are shown in Tables 1-5. Besides, in the ORL database we conduct some experiments to illustrate the difference between various values of the parameter \( \lambda \) (as shown in Fig. 3). According to the results, we can summarize as follows

(1) Overall, DLSR, ReLSR, and GReLSR obtain the competitive recognition rates, which indicates that the \( \varepsilon \)-dragging technique is a significant contribution to improve the classification performance.

(2) For all methods, the average recognition rates increase with the increase of the number of labeled training samples, which states that the more labeled training samples, the more useful information is provided.

(3) Comparing with the other four label-relaxed based regression methods, i.e., DLSR, ReLSR, GReLSR and CLSR, the proposed FDLSR method consistently provides superior classification performance on the above five databases. This is mainly due to the improvement of discriminability with respect to the transformed features. As discussed previously, both the Fisher discrimination criterion and the self-adaptive weighting technique contribute to obtaining discriminative features.

(4) From Fig. 3, we can find the recognition rate improves with the increase of the value of the parameter \( \lambda \), which demonstrates that the introduction of Fisher criterion is indeed beneficial for classification. Besides, we see the transformed features are heterogeneous when \( \lambda = 0 \). However, with the increase of \( \lambda \), the true class structure (the structure of block-diagonal features) becomes more distinct and clear. Moreover, we observe that the similarity of the within-class features and the incoherence of the inter-class features are simultaneously enhanced.
Fig. 2. Examples images from five face database: (1)-(5) respectively correspond to the ORL database, the Extended Yale B database, the AR database, the LFW database and the CMU PIE database.

| Train No. | DLSR  | ReLSR | GReLSR | CLSR  | FDLSR(ours) |
|-----------|-------|-------|--------|-------|-------------|
| 3         | 90.79 | 92.07 | 91.71  | 92.54 | 93.21       |
| 4         | 94.92 | 95.29 | 94.58  | 94.88 | 95.58       |
| 5         | 96.45 | 96.60 | 95.85  | 96.45 | 96.95       |
| 6         | 97.56 | 98.31 | 96.50  | 97.19 | 98.50       |

| Train No. | DLSR  | ReLSR | GReLSR | CLSR  | FDLSR(ours) |
|-----------|-------|-------|--------|-------|-------------|
| 10        | 88.32 | 88.64 | 88.14  | 89.64 | 90.80       |
| 15        | 93.23 | 93.56 | 93.16  | 93.17 | 94.69       |
| 20        | 95.66 | 96.51 | 96.00  | 95.79 | 96.83       |
| 25        | 96.82 | 97.13 | 96.83  | 97.45 | 97.97       |
TABLE 3
Mean classification accuracies (%) of different methods on the AR face database.

| Train No. | DLSR  | ReLSR | GRelSR | CLSR  | FDLSR (ours) |
|-----------|-------|-------|--------|-------|--------------|
| 3         | 83.84 | 84.54 | 85.58  | 86.62 | 88.66        |
| 4         | 89.02 | 89.78 | 90.17  | 90.65 | 92.19        |
| 5         | 92.60 | 92.52 | 93.04  | 93.30 | 93.65        |
| 6         | 94.28 | 94.88 | 94.90  | 95.05 | 95.11        |

TABLE 4
Mean classification accuracies (%) of different methods on the LFW face database.

| Train No. | DLSR  | ReLSR | GRelSR | CLSR  | FDLSR (ours) |
|-----------|-------|-------|--------|-------|--------------|
| 5         | 29.35 | 31.46 | 36.22  | 36.87 | 38.92        |
| 6         | 32.42 | 35.12 | 38.79  | 39.52 | 41.20        |
| 7         | 35.50 | 38.01 | 42.53  | 41.71 | 44.14        |
| 8         | 37.18 | 38.65 | 44.09  | 44.12 | 46.15        |

TABLE 5
Mean classification accuracies (%) of different methods on the CMU PIE database.

| Train No. | DLSR  | ReLSR | GRelSR | CLSR  | FDLSR (ours) |
|-----------|-------|-------|--------|-------|--------------|
| 5         | 73.22 | 74.24 | 74.83  | 75.77 | 80.66        |
| 8         | 83.96 | 84.33 | 83.67  | 84.61 | 87.67        |
| 10        | 87.93 | 87.51 | 87.28  | 88.17 | 89.83        |
| 15        | 92.11 | 92.22 | 91.36  | 91.74 | 92.73        |

3.2 Convergence validation
The convergence of the proposed FDLSR algorithm on five databases is illustrated in Fig. 4. As expected, the convergence of the objective function (11) is dropped fast within only a few number of iterations, which verifies the effectiveness of the used alternating optimization method in solving the FDLSR problem.

4 Conclusion
In this paper, we propose a novel Fisher discriminative criterion based LSR with self-adaptive weighting method (FDLSR) for multiclass face recognition. Different from previous LSR based classification methods, FDLSR aims to improve the classification accuracy by encouraging the discriminability of the learned features. Specifically, FDLSR introduces a Fisher discrimination criterion to promote the transformed features to have small within-class but large inter-class scatter, thus resulting in discriminative transformation. Besides, FDLSR uses a non-negative weighting matrix to self-adaptively enlarge the margins of true and false classes. Experimental results demonstrate the superior performance of FDLSR against other state-of-the-art LSR based classification methods.

ACKNOWLEDGMENTS
This work is supported by the National Natural Science Foundation of China under Grant Nos. 61672265, U1836218, the 111 Project of Ministry of Education of China under Grant No. B12018, and UK EPSRC Grant EP/N007743/1, MURI/EPSRC/DSTL GRANT EP/R018456/1.
Fig. 4. Objective function value versus the number of iterations of the proposed FDLISR method on five face databases.

REFERENCES

[1] Xu, Y., Fang, X., Zhu, Q., Chen, Y., You, J., and Liu, H., “Modified minimum squared error algorithm for robust classification and face recognition experiments,” Neurocomputing, 135(C), 253-261, 2014.
[2] F. Nie, H. Huang, X. Cai, and C. H. Ding, “Efficient and robust feature selection via joint 2,1-norms minimization,” in Proc. NIPS, 2010, pp. 1813-1821.
[3] Feng, Q., Zhou, Y., and Lan, B., “Pairwise linear regression classification for image set retrieval,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4865-4872, 2016.
[4] T. Strutz, “Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond,” Wiesbaden, Germany: Vieweg, 2010.
[5] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” Technometrics, vol. 12, no. 1, pp. 55-67, 1970.
[6] R. Tibshirani, “Regression shrinkage and selection via the LASSO,” J. Roy. Statist. Soc. B, Methodol., vol. 58, no. 1, pp. 267-288, 1996.
[7] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, “Applied Logistic Regression,” New York, NY, USA: Wiley, 2013.
[8] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273-297, 1995.
[9] L. Jiao, L. Bo, and L. Wang, “Fast sparse approximation for least squares support vector machine,” IEEE Trans. Neural Netw., vol. 18, no. 3, pp. 656-697, May 2007.
[10] Ruppert, D., Sheather, S. J., and Wand, M. P., “An effective bandwidth selector for local least squares regression,” J. Amer. Statist. Assoc., 90(432), 1257-1270, 1995.
[11] Abdi, H., “Partial least squares regression and projection on latent structure regression (pls regression),” Wiley Interdiscip. Rev. Comput. Stat., 2(1), 971-106, 2010.
[12] Gao, J., Shi, D., and Liu, X., “Significant vector learning to construct sparse kernel regression models,” Neural Netw., 20(7), 791798, 2007.
[13] I. Naseem, R. Togneri, and M. Bennamoun, “Linear regression for face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 11, pp. 2106-2112, 2010.
[14] J. Wright, A.Y. Yang, A. Ganesh, et al, “Robust face recognition via sparse representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 2, pp. 210-227, 2009.
[15] L. Zhang, M. Yang, and X. Feng, “Sparse representation or collaborative representation: Which helps face recognition?,” in Proc. of IEEE Int. Conf. Comput. Vis., pp. 471-478, 2011.
[16] Cai, X., Ding, C., Nie, F., and Huang, H, “On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions,” In ACM SIGKDD international conference on knowledge discovery and data mining, pp. 11241132, 2013.
[17] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, “Discriminative least squares regression for multiclass classification and feature selection,” IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 11, pp. 17381754, Nov. 2012.
[18] Fang, X., Xu, Y., Li, X., Lai, Z., Wong, W. K., and Fang, B., “Regularized label relaxation linear regression,” IEEE Transactions on Neural Networks and Learning Systems, 2017.
[19] X.-Y. Zhang, L. Wang, S. Xiang, and C.-L. Liu, “Retargeted least squares regression algorithm,” IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 9, pp. 2206-2213, Sep. 2015.
[20] Wang, L., and Pan, C., “Groupwise retargeted least-squares regression,” IEEE Transactions on Neural Networks and Learning Systems, 2017.
[21] H. L. Yuan, J. J. Zheng, L. L. Lai, Y. Y. Tang, “A constrained least squares regression model,” Information Sciences, vol. 429, pp. 247-259, 2018.
[22] M. Yang, L. Zhang, X.C. Feng, and D. Zhang, “Fisher discrimination dictionary learning for sparse representation,” In Proc. ICCV, 2011.
[23] Jie Wen, Bob Zhang, Yong Xu, Jian Yang, Na Han, “Adaptive weighted nonnegative low-rank representation,” Pattern Recognition, 81: 326-340, 2018.
[24] J. Zheng, P. Yang, S. Chen, G. Shen, W. Wang, “Iterative re-constrained group sparse face recognition with adaptive weights learning,” IEEE Trans. Image Process., 26 (5), pp. 2408-2423, 2017.
[25] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via alternating direction method of multipliers,” Found. Trends Mach. Learn., vol. 3, no. 1, pp. 1122, 2011.
[26] T. H. Vu and V. Monga, “Fast low-rank shared dictionary learning for image classification,” IEEE Trans. on Image Proc., vol. 26, no. 11, pp. 51605175, 2017.
[27] F. Nie, X. Wang, M.I. Jordan, H. Huang, “The constrained Laplacian rank algorithm for graph-based clustering,” in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 1969-1976, 2016.

[28] Jie Wen, Yong Xu, Zuoyong Li, Zhongli Ma, Yuanrong Xu, “Inter-class sparsity based discriminative least square regression”, Neural Networks, 102: 36-47, 2018.

[29] F. S. Samaria and A. C. Harter, Parameterisation of a stochastic model for human face identification, in Proc. 2nd IEEE Workshop Appl. Comput. Vis., Dec. 1994, pp. 138-142.

[30] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 643-660, Jun. 2001.

[31] A. M. Martinez and R. Benavente, The AR face database, CVC, New Delhi, India, Tech. Rep. 24, 1998.

[32] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, College Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[33] T. Sim, S. Baker, and M. Bsat, The CMU pose, illumination, and expression database, IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1615-1618, Dec. 2003.