



Schatten-p Norm Based Linear Regression Discriminant Analysis for Face Recognition

Lijiang Chen, Wentao Dou, and Xia Mao (✉)

School of Electronic and Information Engineering, Beihang University,
Beijing 100191, China
moukyou@buaa.edu.cn

Abstract. Locality-regularized linear regression classification (LLRC) shows good performance on face recognition. However, it solely performs on the original space, which results in degraded classification efficiency. To solve this problem, we propose a dimensionality reduction algorithm named Schatten-p norm based linear regression discriminant analysis (SPLRDA) for image feature extraction. First, it defines intra-class and inter-class scatters based on Schatten-p norm, which improves the capability to deal with illumination changes. Then the objective function which incorporates discriminant analysis is derived from the minimization of intra-class compactness and the maximization of inter-class separability. Experiments carried on some typical databases validate the effectiveness and robustness of our method.

Keywords: Dimensionality reduction · Schatten-p norm · Linear regression
Feature extraction · Face recognition · Discriminant analysis

1 Introduction

Dimensionality reduction has played a key role in many fields such as machine learning, face recognition and data mining. It devotes to excavate the low dimensional features from high dimensional data while preserving the intrinsic information existed in data.

During the past few decades, a lot of dimensionality reduction algorithms using for feature extraction has been proposed. Principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2] are two typical methods. As an unsupervised approach, PCA projects high dimensional data onto a variance preserving subspace. In contrast, LDA, as a supervised method, aims to minimize within-class scatter and maximize between-class scatter to extract more discriminative features using labeled class information. However, it often suffers from small sample size (sss) [4] problem resulting from the singularity of within-class scatter. However, the above linear algorithms do not have the capabilities to capture the nonlinear structure embedded in image matrix. A lot of nonlinear feature extraction algorithms have been proposed to excavate the manifold structure of data, such as isometric mapping (ISOMAP) [5], Laplacian eigenmaps (LE) [6] and locally linear embedding (LLE) [7]. ISOMAP extends multidimensional scaling by incorporating the geodesic distances imposed by a weighted graph. LE finds the low-dimensional manifold structure by building a graph,

whose node and connectivity are data points and proximity of neighboring points. LLE determines the low-dimensional representations of data by focusing on how to preserve the locally linear structure and minimizes the linear reconstruction error. However, the so called out-of-sample problem often occurs in those nonlinear algorithms. To solve the out-of-sample problem, locality preserving projection (LPP) [8] was proposed. LPP minimizes the local reconstruction error to preserve local structure and obtain the optimal projection matrix.

After feature extraction, data classification is another step for face recognition. Many classification methods have been proposed, such as nearest neighbor classifier (NNC) [9] and linear regression classification (LRC) [10]. More recently, Brown et al. proposed a locality-regularized linear regression classification (LLRC) [11] method using a specific class as the neighbor of a training sample to classify and improve the accuracy of classification.

Although the above feature extraction and classification methods obtain great performances, each of them are designed independently. Therefore, they may be not fit each other perfectly. Using the rule of LRC, Chen et al. proposed reconstructive discriminant analysis (RDA) [12]. In 2018, Locality-regularized linear regression discriminant analysis (LLRDA) [13] deriving from LLRC was proposed to extract features. LLRDA ameliorates the LLRC by performing intra-class and inter-class scatter in the feature subspace, which brings more appropriate features for LLRC. However, in the feature subspace, LLRDA measures the reconstruction error utilizing L2 norm, which causes the strong sensitiveness to illumination changes and outliers. To alleviate the deficiency of L2 norm based methods, many algorithms basing on schatten- p norm have been developed. To improve the robustness to illumination changes and outliers, two-dimensional principal component analysis based on schatten- p norm (2DPCA-SP) [14] was presented using schatten- p norm to measure the reconstruction error. Incorporating the discriminant analysis and schatten- p norm to extract discriminative and robust features, two-dimensional discriminant analysis based on schatten- p norm (2DDA-SP) [15] was proposed. In 2018, Shi et al. proposed robust principal component analysis via optimal mean by joint 2, 1 and schatten p -norms minimization (RPOM) [16] imposing an schatten- p norm based regularized term to suppress the singular values of reconstructed data. Motivated by the above methods, we propose an LLRC based feature extraction method named schatten- p Norm based linear regression discriminant analysis (SPLRDA) utilizing schatten- p norm to improve the robustness to illumination changes. The main advantages of our algorithm are listed below: (1) Features are directly extracted from matrix rather than vectors reshaped from original image; (2) A specific class is assumed to be the neighborhood of a training sample instead of selecting from all samples when calculating the reconstruction vector β ; (3) Discriminant analysis is incorporated to obtain discriminative features; (4) In the feature subspace, we measure the similarity distances by schatten- p norm whose parameter p is adjustable, which is more robust to illumination changes and outliers.

The rest parts of this paper are organized as follows. Sect. 2 briefly reviews the background knowledge of LRC and LLRC. The presented SPLRDA method is introduced in Sect. 3. The experimental results and analysis are arranged in Sect. 4. Finally, Sect. 5 concludes this paper.

2 Related Work

Suppose $X = [x_1, x_2, \dots, x_n]$ be a set of n training images of C classes. Given the number of images in i th class is n_i , therefore, we have $\sum_{i=1}^C n_i = n$. x_i denotes the i th reshaped image whose dimension N is the product of the row and column numbers of original i th image. In this section, LRC and LLRC are reviewed briefly.

2.1 LRC

LRC [10] is based on the assumption that a sample can be represented as a linear combination of samples from same class. The task of LRC is finding which class the testing sample y belongs to. Let y be a test sample from the i th class, then it can be reconstructed approximately by:

$$y = X_i \beta_i \quad (1)$$

where $\beta_i \in R^{m \times 1}$ is the reconstruction coefficient vector with respect to training image set of class i . β_i is calculated by least square estimation (LSE) method as:

$$\beta_i = (X_i^T X_i)^{-1} X_i^T y \quad (2)$$

Utilizing the estimated β_i , y can be reconstructed as:

$$\hat{y}_i = X_i (X_i^T X_i)^{-1} X_i^T y \quad (3)$$

Since \hat{y}_i should approximate to y , the reconstruction error based on Euclidean norm is defined as:

$$\begin{aligned} l(y) &= \min_i \|y - \hat{y}_i\|^2 \\ &= \min_i \left\| y - X_i (X_i^T X_i)^{-1} X_i^T y \right\|^2 \end{aligned} \quad (4)$$

where $l(y)$ represents the class label of y .

2.2 LLRC

Different from LRC, LLRC [11] pays more attention to the local linearity of each sample and considers it to be more important than global linearity. Therefore, images from a specific class instead of all samples are supposed to be the neighborhood of a image sample based on this principle, Brown *et al.* presented an constraint of locality regularization on LRC by solely involving k closest images to the query image based on Euclidean distance measure.

The k nearest neighbors set of testing sample y in class i is denoted by $\tilde{X}_i = [x_{i1}, x_{i2}, \dots, x_{ik}] \in R^{N \times k}$. Similar to the LRC, the label of y is computed by minimizing the reconstruction error as below:

$$l(y) = \min_i \left\| y - \tilde{X}_i (\tilde{X}_i^T \tilde{X}_i)^{-1} \tilde{X}_i^T y \right\|^2 \quad (5)$$

3 Our Method

3.1 Problem Formulation

Suppose $x_i^j \in R^{a \times b}$ be the j th training sample of i th class, then its intra-class and inter-class reconstruction error are defined respectively as:

$$\sum_{i,j} e_i^j = \sum_{i,j} \|x_i^j - \tilde{X}_i^j \tilde{\beta}_i^j\|_{sp}^p \quad (6)$$

$$\sum_{i,j,m} e_{im}^j = \sum_{i,j,m} \|x_i^j - \tilde{X}_{im}^j \tilde{\beta}_{im}^j\|_{sp}^p \quad (7)$$

where \tilde{X}_i^j and \tilde{X}_{im}^j denote the k nearest neighbors set of x_i^j in class i and class m respectively. Class m is one of the K nearest heterogeneous subspaces of x_i^j . The intra-class scatter characterizes the compactness of each training samples class, while the inter-class scatter describes the separability between different classes. $\|\bullet\|_{sp}$ denotes the schatten- p norm. Since we know that the singular value decomposition of x is defined as:

$$U \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \sigma_{\min(m,n)} \end{bmatrix} V^T = SVD(x) \quad (8)$$

The schatten- p norm can be represented by [15]:

$$\|x\|_{sp} = \left(\sum_{i=1}^{\min(m,n)} \sigma_i^p \right)^{\frac{1}{p}} = \left[Tr(xx^T)^{\frac{p}{2}} \right]^{\frac{1}{p}} \quad (9)$$

where σ_i is the i th singular value of x . If the parameter p was set to be 1, the schatten- p norm becomes the nuclear norm, which is famous for its capability for solving illumination changes. In the experiments, we can adjust p to the value that attains best performance for face recognition. Suppose the optimal projection matrix be denoted by $A \in R^{b \times s}$ ($s < b$). In the feature subspace, the corresponding data can be replaced by:

$$y_i^j = x_i^j A \quad (10a)$$

$$Y_i^j = \left[\tilde{X}_i^{j(1)} A, \tilde{X}_i^{j(2)} A, \dots, \tilde{X}_i^{j(k)} A \right] \quad (10b)$$

$$Y_{im}^j = \left[\tilde{X}_{im}^{j(1)} A, \tilde{X}_{im}^{j(2)} A, \dots, \tilde{X}_{im}^{j(k)} A \right] \quad (10c)$$

where $y_i^j \in R^{a \times s}$, $Y_i^j \in R^{a \times ks}$ and $Y_{im}^j \in R^{a \times ks}$. Therefore, to find an optimal projection matrix which projects the original data into feature subspace, the function performed in the feature subspace should be maximized:

$$f(A) = \left[\sum_{i,j,m} \|y_i^j - Y_{im}^j \tilde{\beta}_{im}^j\|_{sp}^p - \sum_{i,j} \|y_i^j - Y_i^j \tilde{\beta}_i^j\|_{sp}^p \right] \quad \text{s.t. } A^T A = I_s \quad (11)$$

3.2 Problem Solving

As analyzed in the last subsection, the optimization problem can be formulated and simplified as:

$$\begin{aligned} \arg \max_A f(A) &= \arg \max_A \left[\sum_{i,j,m} \|y_i^j - Y_{im}^j \tilde{\beta}_{im}^j\|_{sp}^p - \sum_{i,j} \|y_i^j - Y_i^j \tilde{\beta}_i^j\|_{sp}^p \right] \\ &= \arg \max_A \left[\sum_{i,j,m} \|x_i^j A - \tilde{X}_{im}^j \tilde{\beta}_{im}^j A\| - \sum_{i,j} \|x_i^j A - \tilde{X}_i^j \tilde{\beta}_i^j A\|_{sp}^p \right] \\ &= \arg \max_A \left\{ \sum_{i,j,m} \text{Tr} \left[B_{ij}^m A A^T (B_{ij}^m)^T \right]^{\frac{p}{2}} - \sum_{i,j} \text{Tr} \left[W_{ij} A A^T (W_{ij})^T \right]^{\frac{p}{2}} \right\} \\ &\quad \text{s.t. } A^T A = I_s \end{aligned} \quad (12)$$

where $B_{ij}^m = x_i^j - \tilde{X}_{im}^j \tilde{\beta}_{im}^j$ and $W_{ij} = x_i^j - \tilde{X}_i^j \tilde{\beta}_i^j$. Based on the objective function, the Lagrangian function can be built as:

$$L(A, \Lambda) = \sum_{i,j,m} \text{Tr} \left[B_{ij}^m A A^T (B_{ij}^m)^T \right]^{\frac{p}{2}} - \sum_{i,j} \text{Tr} \left[W_{ij} A A^T (W_{ij})^T \right]^{\frac{p}{2}} - \text{Tr} [\Lambda (A^T A - I_s)] \quad (13)$$

Taking the derivative of L with respect to A , we have:

$$\frac{\partial L}{\partial A} = 2 \sum_{i,j,m} (B_{ij}^m)^T D_{ij}^m B_{ij}^m A - 2 \sum_{i,j} (W_{ij})^T H_{ij} W_{ij} A - 2 \Lambda A \quad (14)$$

where $D_{ij}^m = \frac{\nu}{2} \left[B_{ij}^m A A^T (B_{ij}^m)^T \right]^{\frac{\nu-2}{2}}$ and $H_{ij} = \frac{\nu}{2} \left[W_{ij} A A^T (W_{ij})^T \right]^{\frac{\nu-2}{2}}$. $\Lambda \in R^{s \times s}$ is the symmetric Lagrangian multiplier matrix. To look for the maximum point of objective function, (14) is set to be zero, then the Eq. (14) is changed to:

$$(S_b - S_w)A = A\Lambda \quad (15)$$

where $S_b = \sum_{i,j,m} (B_{ij}^m)^T D_{ij}^m B_{ij}^m$ and $S_w = \sum_{i,j} (W_{ij})^T H_{ij} W_{ij}$. Both S_b and S_w rely on A . If $(S_b - S_w)$ is a known constant matrix and considering the orthogonal constraint $A^T A = I_s$, the optimal projection matrix A can be calculated by solving the eigen value decomposition problem as below:

$$A^T (S_b - S_w) A = \Lambda \quad (16)$$

According to (15), since $(S_b - S_w)$ is symmetric and each element in the diagonal of Λ is an eigen value of $(S_b - S_w)$, A is formed by the s eigen vectors corresponding to s largest eigen values of $(S_b - S_w)$. Since the objective function is bounded by $A^T A = I_s$ and it increases after each iteration, the convergence of this algorithm can be realized. Based on these analyses, we propose an iterative algorithm to obtain the optimal projection matrix. The algorithm is concluded in Algorithm 1.

Algorithm 1. An efficient iterative algorithm for solving the Eq. (12)

Input: N training image matrices $\{x_i^j\}$, $i = 1, \dots, C, j = 1, \dots, l$ (l denotes the number of training samples per individual); Values of k, K, p, ε ; Initialize A such that $A^T A = I_s$; Set $t = 0$.

While not converge:

1. Compute $B_{ij}^m, W_{ij}, D_{ij}^m, H_{ij}, S_b$ and S_w ;
2. Perform eigen value decomposition of $(S_b - S_w)$, A is formed by the s eigen vectors corresponding to s the largest eigen values of $(S_b - S_w)$;
3. Check the convergence situation $\|A^{t+1} - A^t\| < \varepsilon$;
4. $t = t + 1$.

End while

Output: the optimal projection matrix $A = A^t$.

4 Experiments

In this section, extensive experiments are conducted on ORL [17] and CMU PIE [18] databases to testify the effectiveness of our method in the condition that $p = 1/4, 1/2, 3/4$ respectively. Meanwhile, we compare SPLRDA with other state-of-art feature extraction methods such as PCA [1], LDA [2], LPP [8] and RDA [12]. Different classifiers are adopted to measure the performance of each method. All the algorithms have been run for five times independently to obtain average recognition rates. We only exhibit the highest results for comparison and analysis.

4.1 Experiments on ORL Database

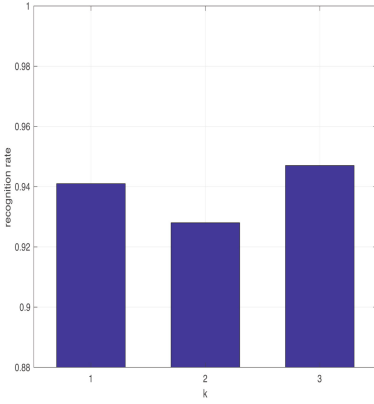
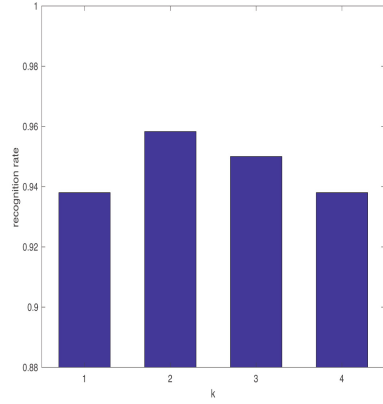
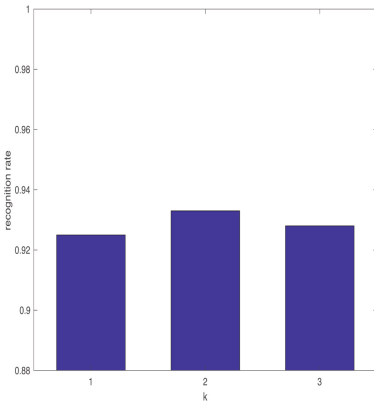
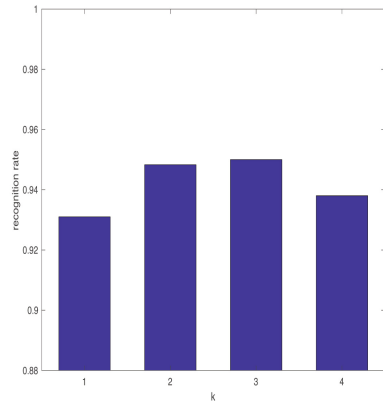
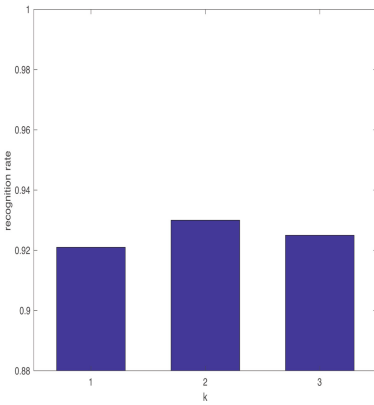
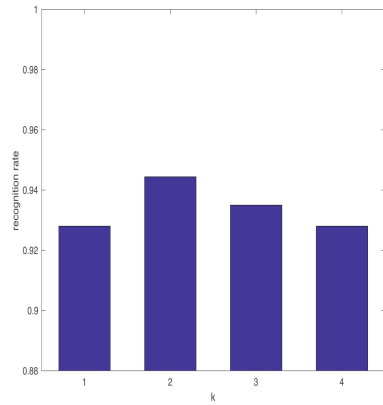
The ORL face database includes 400 face images belonging to 40 individuals. Each person has 10 images distinct from view direction, facial expression (mouth opened or closed, laughing or calm), facial details (with or without glasses) and illumination. Each image has been normalized to the size of 112×92 with 256 gray levels.

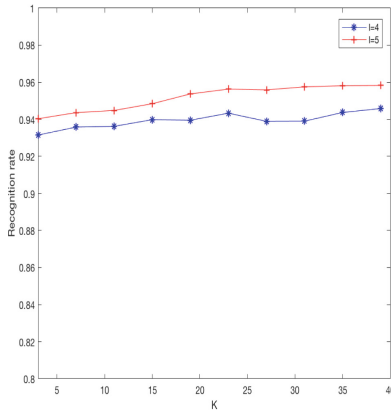
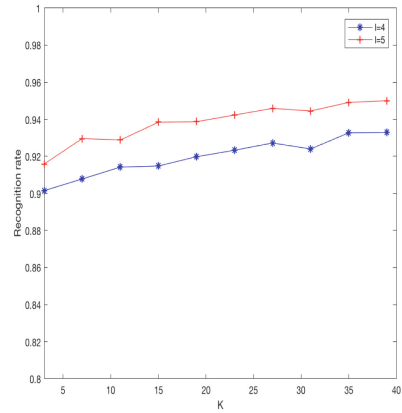
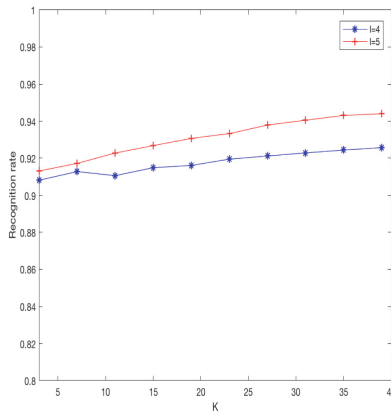
Ten samples of one individual are displayed in Fig. 1. In our experiments, we cropped each image manually and resized it to 32×32 pixels. l ($l = 4, 5$) images per person are randomly selected for training while the remainders for testing. Note that k ranges from 1 to $l - 1$ and K ranges from 1 to $C - 1$. To observe the effect of k , K is fixed to $\lfloor C/2 \rfloor + 1$ and k is varied from 1 to $l - 1$.



Fig. 1. Ten samples of one individual in ORL database

Figure 2 shows the recognition rates of SPLRDA plus LLRC with varied k . The results indicate that our method achieves best performances when $k = 2$ and $k = 3$ associating with $l = 5$ and $l = 4$, which demonstrates the effectiveness of exploiting neighborhood structure. Therefore, to detect the effect of K , we fixed k to 3 and 2 corresponding to $l = 4$ and $l = 5$ respectively, and varied K from 3 to $C - 1$ in increments of 4. Figure 3 displays the recognition rates of SPLRDA plus LLRC with varied K . From Fig. 3, it can be seen that parameter K does affect the performance of our method and they all gain best recognition rates when K achieves highest value. Average recognition rates of all methods are displayed in Table 1. In the experiments, we tested these feature extraction methods under three different classifiers. Experimental results verify that classifier matters the performance of face recognition and our method fits LLRC better than other related methods. From Table 1, we can see that when $l = 4$ and $l = 5$, SPLRDA achieves best recognition rates in the condition that $p = 1/4$, which demonstrates the effectiveness of our method.

(a) $p=1/4, l=4$ (b) $p=1/4, l=5$ (c) $p=1/2, l=4$ (d) $p=1/2, l=5$ (e) $p=3/4, l=4$ (f) $p=3/4, l=5$ **Fig. 2.** Recognition rates of SPLRDA plus LLRC with varied k

(a) $p=1/4$ (b) $p=1/2$ (c) $p=3/4$ **Fig. 3.** Recognition rates of SPLRDA plus LLRC with varied K**Table 1.** Recognition accuracy of each method in ORL database

		PCA	LDA	LPP	RDA	SPLRDA ($p = 1/4$)	SPLRDA ($p = 1/2$)	SPLRDA ($p = 3/4$)
$l = 4$	NNC	0.8892	0.9125	0.8489	0.9050	0.9050	0.8991	0.8808
	LRC	0.8892	0.9000	0.9000	0.9175	0.9333	0.9300	0.9175
	LLRC	0.8925	0.9083	0.8925	0.9192	0.9458	0.9333	0.9300
$l = 5$	NNC	0.9300	0.9250	0.8808	0.9400	0.9350	0.9300	0.9240
	LRC	0.9500	0.9200	0.9200	0.9500	0.9500	0.9400	0.9410
	LLRC	0.9500	0.9333	0.9200	0.9400	0.9583	0.9500	0.9440

4.2 Experiments on CMU PIE Database

The CMU PIE face database contains 68 different individuals with more than 40,000 face images. Each image of an individual differs from others in poses, illumination and expression. This database is stipulated that 4 different expressions, 43 different illumination changes and 13 different poses should be satisfied for images of an individual. This database includes five near-frontal poses (C05, C07, C09, C27 and C29). We choose C05 subset for our experiments. The subset contains 1632 images of 68 individuals. In the experiments, all the images were cropped and resized to 64×64 pixels. Figure 4 shows ten samples of one individual in CMU PIE database.



Fig. 4. Ten samples of one individual in CMU PIE database

In the experiments, we randomly choose l (6, 7, 8) images per person for training and the remainders for testing. Table 2 reports the recognition rates with different classifiers. From Table 2, some observations are concluded as follow: (1) unsupervised methods (PCA and LPP) perform worse than other supervised methods because labeled information are not be exploited; (2) Although RDA plus LRC gains high recognition rate, it performs worse than SPLRDA plus LLRC since neighborhood structure of data is not utilized; (3) SPLRDA achieves better performance than other methods, which verifies that our method is better than other related methods.

Table 2. Recognition accuracy of each method in CMU PIE database

		PCA	LDA	LPP	RDA	SPLRDA ($p = 1/4$)	SPLRDA ($p = 1/2$)	SPLRDA ($p = 3/4$)
$l = 6$	NNC	0.3302	0.5497	0.4166	0.5229	0.5987	0.6012	0.5997
	LRC	0.5231	0.7116	0.5602	0.7211	0.7271	0.7285	0.7255
	LLRC	0.5044	0.6895	0.5379	0.7071	0.7419	0.7426	0.7408
$l = 7$	NNC	0.3837	0.5988	0.4612	0.4533	0.5687	0.5672	0.5683
	LRC	0.6058	0.7532	0.6168	0.8426	0.8345	0.8357	0.8339
	LLRC	0.6113	0.7791	0.6174	0.8391	0.8627	0.8581	0.8576
$l = 8$	NNC	0.4912	0.7527	0.5713	0.5947	0.5969	0.5955	0.5937
	LRC	0.7645	0.8504	0.7792	0.9215	0.9224	0.9207	0.9194
	LLRC	0.7539	0.8583	0.7825	0.9203	0.9298	0.9226	0.9226

5 Conclusion

In this paper, a novel feature extraction method named Schatten-p norm based linear regression discriminant analysis (SPLRDA) has been proposed. It not only incorporates Schatten-p norm reducing the interference of illumination changes but also exploits neighborhood structure, which fits LLRC well. Experiments has demonstrated the reliability and effectiveness of our method. It performs better than other related methods.

References

1. Turk, M., Pentland, A.: Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**(1), 71–86 (1991)
2. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 711–720 (1997)
3. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788 (1999). EP
4. Raudys, S.J., Jain, A.K.: Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(3), 252–264 (1991)
5. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500), 2319–2323 (2000)
6. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**(6), 1373–1396 (2003)
7. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**(5500), 2323–2326 (2000)
8. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face recognition using Laplacianfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 328–340 (2005)
9. Cover, T., Hart, P.: Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**(1), 21–27 (1967)
10. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(11), 2106–2112 (2010)
11. Brown, D., Li, H., Gao, Y.: Locality-regularized linear regression for face recognition. In: *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012*, pp. 1586–1589 (2012)
12. Chen, Y., Jin, Z.: Reconstructive discriminant analysis: a feature extraction method induced from linear regression classification. *Neurocomputing* **87**, 41–50 (2012)
13. Huang, P., Li, T., Shu, Z., Gao, G., Yang, G., Qian, C.: Locality-regularized linear regression discriminant analysis for feature extraction. *Inf. Sci.* **429**, 164–176 (2018)
14. Du, H., Hu, Q., Jiang, M., Zhang, F.: Two-dimensional principal component analysis based on Schatten p-norm for image feature extraction. *J. Vis. Commun. Image Represent.* **32**, 55–62 (2015)
15. Du, H., Zhao, Z., Wang, S., Hu, Q.: Two-dimensional discriminant analysis based on Schatten p-norm for image feature extraction. *J. Vis. Commun. Image Represent.* **45**, 87–94 (2017)
16. Shi, X., Nie, F., Lai, Z., Guo, Z.: Robust principal component analysis via optimal mean by joint $l_{2,1}$ and Schatten p-norms minimization. *Neurocomputing* **283**, 205–213 (2018)

17. Samaria, F.S., Harter, A.C.: Parameterisation of a stochastic model for human face identification. In: Proceedings of 1994 IEEE Workshop on Applications of Computer Vision, pp. 138–142 (1994)
18. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, pp. 46–51 (2002)