Linear Discriminant Analysis with Maximum Correntropy Criterion

Wei Zhou and Sei-ichiro Kamata

Waseda University, Japan

Abstract. Linear Discriminant Analysis (LDA) is a famous supervised feature extraction method for subspace learning in computer vision and pattern recognition. In this paper, a novel method of LDA based on a new Maximum Correntropy Criterion optimization technique is proposed. The conventional LDA, which is based on L2-norm, is sensitivity to the presence of outliers. The proposed method has several advantages: first, it is robust to large outliers. Second, it is invariant to rotations. Third, it can be effectively solved by half-quadratic optimization algorithm. And in each iteration step, the complex optimization problem can be reduced to a quadratic problem that can be efficiently solved by a weighted eigenvalue optimization method. The proposed method is capable of analyzing non-Gaussian noise to reduce the influence of large outliers substantially, resulting in a robust classification. Performance assessment in several datasets shows that the proposed approach is more effectiveness to address outlier issue than traditional ones.

1 Introduction

In many data measurement problems, observation data often lies in a lower dimensional subspace which can be obtained from the original high dimensional data space. Such a lower dimensional subspace, especially the linear subspace, has many important applications in computer vision or pattern recognition, such as object recognition [1], motion estimation [2]. Among these subspace methods, linear discriminant analysis (LDA) [3] is one of the most popular methods. LDA tries to find a set of projections that maximize the ratio of the between-class distance to the within-class distance. These projections constitute a low-dimensional linear subspace by which the data structure in the original input space can be effectively captured.

In general, LDA approaches [3] [4] utilize the Frobenius norm (L2-Norm) (we call it LDA-L2 in the following) to measure the between-class and within-class distances. Thus, the process of training may be dominated by outliers since the between-class or within-class distances is determined by the sum of squared distances. Recently, in order to solve the outlier problem, Li [5] proposed rotation invariant L1-norm (notated as R1-norm) based linear discriminant analysis (we call it LDA-R1 in the following). The R1-norm is determined by the sum of elements without being squared. Thus, the R1 norm is less sensitive to outliers than L2-norm. However, in the spatial dimension, squared data is still used.

K.M. Lee et al. (Eds.): ACCV 2012, Part I, LNCS 7724, pp. 500–511, 2013.

Moreover, LDA-R1 takes a lot of time to achieve convergence for a large dimensional input space and it can not effectively handle large outlier problem. In this paper, instead of maximizing variance which is based on L2-norm, maximum correntropy criterion (MCC) [6] based linear discriminant analysis (we denote it as LDA-MCC) is proposed, which is a useful measurement to handle non-Gaussian noise with large outliers. From the viewpoint of Information Theoretic Learning (ITL), LDA-MCC is a natural extension of LDA by replacing MSE criterion by MCC and has several appealing advantages: 1) It is robust to outliers as well as rotationally invariant. 2) Optimal solutions of the proposed method are the principal eigenvectors of a robust covariance matrix corresponding to the largest eigenvalues.

The remainder of this paper is organized as follows: Problem formulation will be described in section 2. In section 3, the solution of the proposed method will be introduced and experiments are presented in section 4. Finally, conclusions and future work are discussed in section 5.

2 Problem Formulation

Assume we have a set of samples $X = \{\{x_i^l\}_{i=1}^{N_l}\}_{l=1}^{C} \in \mathbb{R}^{d\times n}$, N_l of which belong to class ω_l $(l = 1, 2, ..., C)$, where n and d denote the number of samples and the dimension of the original input space, respectively. And $n = \sum_{l=1}^{C} N_l$. In LDA-L2, the objective is to seek t projections $Y = \{\{y_i^l\}_{i=1}^{N_l}\}_{l=1}^{C} \in \mathbb{R}^{t\times n}$ by means of t linear transformation vectors $W \in \mathbb{R}^{d\times t}$, which embeds the original d dimension into t dimension vector space such that $t < d$. Let Tr(.) be the trace of its matrix argument, S_b be the between-class scatter matrix, and S_w be the within-class scatter matrix, which are formulated as: $S_b = \sum_{l=1}^{C}(m_l - m)(m_l - m)^T$ and $S_w = \sum_{l=1}^{C}\sum_{i=1}^{N_l}(x_i^l - m_l)(x_i^l - m_l)^T$. Here $m_l = (1/N_l)\sum_{i=1}^{N_l} x_i^l$ is the mean of the samples belonging to class ω_l, and $m = (1/n)\sum_{l=1}^{C} N_l m_l$ is the global mean of the samples. LDA-L2 aims to find an optimal transformation W by maximizing the ratio of $\mathrm{Tr}(S_b)$ and $\mathrm{Tr}(S_w)$ as following problem

$$\begin{aligned} \max_W J_{L2} &= \max_W \frac{\mathrm{Tr}(S_b)}{\mathrm{Tr}(S_w)} \\ &= \frac{W^T S_b W}{W^T S_w W} \end{aligned} \tag{1}$$

The denominator of the objective function J_{L2} can be simply to $W^T S_w W = I$, since it is invariant with respect to rescaling of the vectors $W \to \beta W$ (β is some coefficient). Thus, the problem of maximizing J_{L2} can be converted into the following constrained optimization problem:

$$\begin{aligned} \max_W \quad & W^T S_b W \\ s.t. \quad & W^T S_w W = I \end{aligned} \tag{2}$$

It is known that the L2-norm is sensitive to outliers and recently, R1-norm approach [5] was presented to solve this problem. In this case, the problem

becomes finding W that maximizes the following objective function:

$$\begin{aligned} \max_{W} J_{R_1} = (1-\alpha)\sum_{l=1}^{C}\sqrt{||W^T(m_l-m)||^2} - \\ \alpha\sum_{l=1}^{C}\sum_{i=1}^{N_l}\sqrt{||W^T(x_i^l-m_l)||^2} \end{aligned} \tag{3}$$

However, for a large dimensional input space, it takes a lot of time to achieve convergence, and in the spatial dimension, squared data is still used. Thus, R1-norm approach is not effective and efficient for larger outlier problems. In this paper, we try to use Maximum Correntropy Criterion (MCC) to measure the between-class scatter instead of Mean Square Error (MSE). In practice, the correntropy is defined as a generalized similarity measure between two arbitrary random variables A and B:

$$V_{n,\sigma}(A,B) = \frac{1}{n}\sum_{l=1}^{n} k_\sigma(a_l - b_l) \tag{4}$$

When kernel function $k_\sigma(.)$ is Gaussian kernel $g(x) = exp(-x^2/2\sigma^2)$, then

$$V_{n,\sigma}(A,B) = \frac{1}{n}\sum_{l=1}^{n} g(a_l - b_l) \tag{5}$$

In order to measure the similarity of two random variables A and B, MSE uses all the samples in the input space while correntropy is just determined by kernel function along the line $a_l = b_l$. This important property intuitively explains the reason that the correntropy is superior to MSE if the residual of $A - B$ is non-symmetric or with nonzero mean.

In ITL, it has been pointed out that MSE is a global measurement while MCC is a local measurement [6]. By global, that means all the data points in the joint space will contribute equally to the value of the measurement and the locality of MCC means that the value is mainly determined by the kernel function. Since an outlier is far away from the data cluster, then its contribution to estimating correntropy will be smaller so that it always receives a low value in the matrix. Therefore, the outliers will have weaker influence on the estimation as correntropy increases. As a result, LDA-MCC is robust against outliers even large outliers occur.

Substituting $a_l = (m_l - m)$ and $b_l = WV_l$ into Eq.(5), here, $V_l = W^T(m_l - m)$ is a projected vector, and we can obtain a novel maximum correntropy criterion based LDA as follows:

$$\begin{aligned} \max_{W} \quad & J_{MCC} = \sum_{l=1}^{C} g((m_l - m) - WV_l) \\ s.t. \quad & W^T S_w W = I \end{aligned} \tag{6}$$

Since W is orthonormal and then

$$\begin{aligned} & g((m_l - m) - WV_l) = g(\sqrt{||(m_l - m) - WW^T(m_l - m)||^2}) \\ & = g(\sqrt{(m_l - m)^T(m_l - m) - (m_l - m)^T WW^T(m_l - m)}) \end{aligned} \tag{7}$$

let $M_l = (m_l - m)$then the Eq.(6) can be converted into following objective function:

$$\begin{array}{ll} \underset{W}{max} & J_{MCC} = \sum_{l=1}^{C} g(\sqrt{M_l^T M_l - M_l^T W W^T M_l}) \\ s.t. & W^T S_w W = I \end{array} \tag{8}$$

3 Linear Discriminant Analysis with Maximum Correntropy Criterion

Recently, Information theoretic learning (ITL) has been proved more efficient to data analysis problems. ITL utilizes probability density function of the data, estimated by Parzen kernel estimator [7], as the cost function.

3.1 Optimization

In ITL, the half-quadratic technique [8] [9] is often used to solve nonlinear ITL optimization problem. And in our study, half quadratic based algorithm is also applied to solve Eq.(8). According to the theory of convex conjugated functions [8], we can get the following proposition.

Proposition: There exists a convex conjugated function φ of $g(x)$ such that

$$g(x) = \underset{p'}{max}(p' \frac{||x||^2}{\sigma^2} - \varphi(p')) \tag{9}$$

where $p' \in R$ is a scalar variable, and for a fixed x, the maximum is reached at $p' = -g(x)$ [9]. Substituting Eq.(9) into Eq.(8), we can get an augmented objective function in the enlarged parameter space then the Eq.(6) can be converted into

$$\begin{array}{ll} \underset{W,P}{max} & J_{MCC} = \sum_{l=1}^{C} (p_l (M_l^T M_l - M_l^T W W^T M_l) - \varphi(p_l)) \\ s.t. & W^T S_w W = I \end{array} \tag{10}$$

where $P = [p_1, p_2, ..., p_C]$ is storing the auxiliary variables introduced in the Half-Quadratic optimization. Consequently, we can optimize (W, P) by iterations as:

$$\underset{W,P}{max}\, \mathcal{L} = J_{MCC} - \lambda(W^T S_w W - I) \tag{11}$$

Then, according to Lagrangian method, a weighted traditional LDA problem can be obtained as follows

$$(S_w)^{-1} S_b P W = \lambda W \tag{12}$$

where P is a diagonal matrix whose diagonal entity $p(l,l) = -p_l$ and $p_l = -g(\sqrt{M_l^T M_l - M_l^T W W^T M_l})$. Thus, the final algorithm of LDA-MCC is listed in Algorithm 1.

Algorithm 1. LDA-MCC

Require: $X = \{\{x_i^l\}_{i=1}^{N_l}\}_{l=1}^{C} \in \mathbb{R}^{d\times n}, t \leq d$
Initialization: $W = [w_1, w_2, ..., w_t] \in \mathbb{R}^{d\times t}, W^T W = I$
while not converge **do**
1. Calculate $p_l = -g(\sqrt{M_l^T M_l - M_l^T W W^T M_l})$
2. Update W according to $(S_w)^{-1} S_b P W = \lambda W$
end while
return $W \in \mathbb{R}^{d\times t}$

3.2 Convergence of LDA-MCC

Let r be the iteration number of Algorithm 1. then

$$\begin{aligned} J_{MCC}^{r+1} - J_{MCC}^{r} &= J_{MCC}(W^{r+1}, P^{r+1}) - J_{MCC}(W^r, P^r) \\ &= [J_{MCC}(W^{r+1}, P^{r+1}) - J_{MCC}(W^r, P^{r+1})] \\ &\quad + [J_{MCC}(W^r, P^{r+1}) - J_{MCC}(W^r, P^r)] \end{aligned} \tag{13}$$

Based on the Proposition and Eq.(12), W^{r+1} and P^{r+1} is the optimization value for J_{MCC}^{r+1} and J_{MCC}^{r}, respectively. Then $J_{MCC}(W^{r+1}, P^{r+1}) - J_{MCC}(W^r, P^{r+1}) \geq 0$ and $J_{MCC}(W^r, P^{r+1}) - J_{MCC}(W^r, P^r) \geq 0$. So $J_{MCC}^{r+1} - J_{MCC}^{r} \geq 0$. That is, the objective function $J_{MCC}^{r}|_{r=1,2,...}$ increases monotonically. In the other side, apparently, $J_{MCC}^{r}|_{r=1,2,...}$ function has an upper bound Thus, we can get that $J_{MCC}^{r}|_{r=1,2,...}$ converges.

4 Experiments

In this section, the proposed approach is applied to some pattern recognition problems and the performance is compared with those of LDA-L2 and LDA-R1. This work follows the lines of correntropy [6] and estimates the bandwidth σ by Silvermans rule [10].

4.1 Toy Set

The first experiment is based on a toy set composed of ten samples clustered into two category with an additional large outlier as shown in Fig.1(a).

To evaluate the effectiveness of LDA-MCC which is less sensitivity to outlies, the outlier sample (plotted as triangle at the top-right corner of Fig.1(a)) is intentionally added into the training samples of Class 1 before classification. For this kind of data, LDA-L2, LDA-R1 and LDA-MCC are applied and the projection vectors are $w_{L2} = [-0.7071, 0.7071]^T$, $w_{R1} = [-0.76431, 0.6448]^T$ and $w_{MCC} = [-0.8784, 0.4779]^T$. The final learning results are plotted as 1-dimensional signals in Fig.1(b) , Fig.1(c) and Fig.1(d) corresponding to LDA-L2, LDA-R1 and LDA-MCC, respectively. After the step of dimension reduction. Clearly, the between-class scatter of the two-class samples except for the outlier sample in Fig.1(d) is much larger than that in Fig.1(b) and Fig.1(c). In this

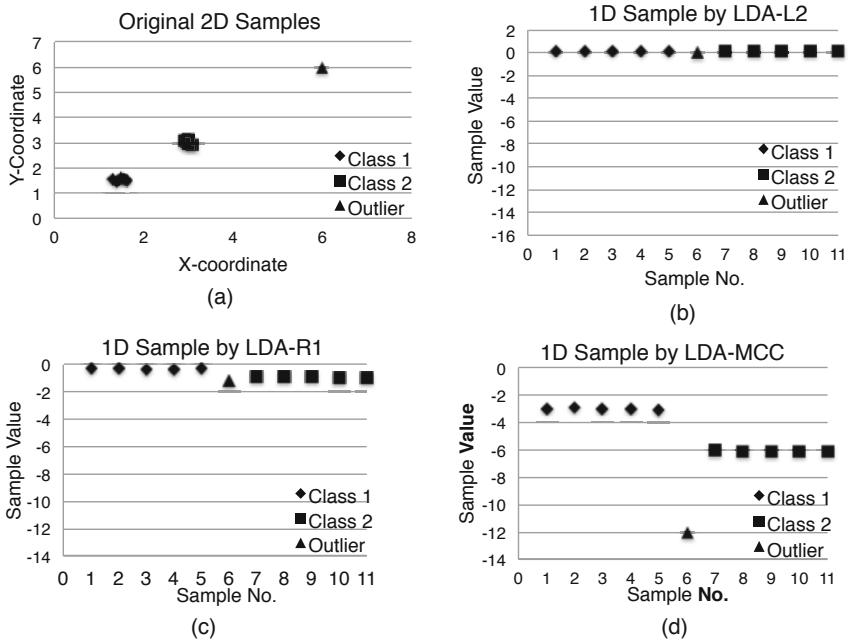

Fig. 1. (a) Samples in toy set (b) Results of LDA-L2 (c) Results of LDA-R1 (d) Results of LDA-MCC

experiment, LDA-MCC is randomly initialized and only two iterations are taken for convergence, while LDA-R1 converges in four iterations. Thus, the proposed method is more powerful to address the outlier problem.

4.2 Brodatz Texture Dataset

The second experiment is to evaluate the classification performance over the subset of Brodatz Texture Dataset [11]. In this dataset, 20 images are selected as category("real" images) and one image is selected as outlier image(shown in Fig.2).

At first, each image is normalized into 128 × 128 size, and then is non-overlapping divided into 16 regions. 5 regions per category and 1 region in outlier image are used as gallery and others per category are treated as probe. The final classification results are shown in Fig.3, where x-axis corresponds to the reduced dimension and y-axis is associated with the accuracy. From this figure, we can see the proposed method is less sensitive to outlier than the other two traditional approaches. In average, the proposed method can achieve about 6 percent higher than LDA-R1 and 11 percent higher than LDA-L2 approach. Moreover, form Dim. 30 to Dim. 40, the accuracy of LDA-R1 drops significantly, that means the projection weights from Dim. 30 to Dim. 40 obtained by LDA-R1 are very sensitivity to the outlier while LDA-MCC is much stable.

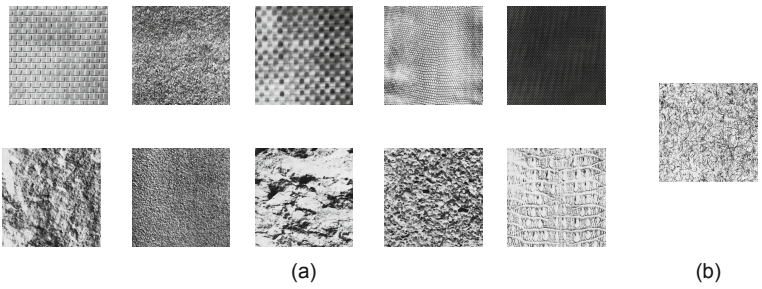

Fig. 2. Samples in Brodatz Texture Dataset (a) "real" images (b)outlier image

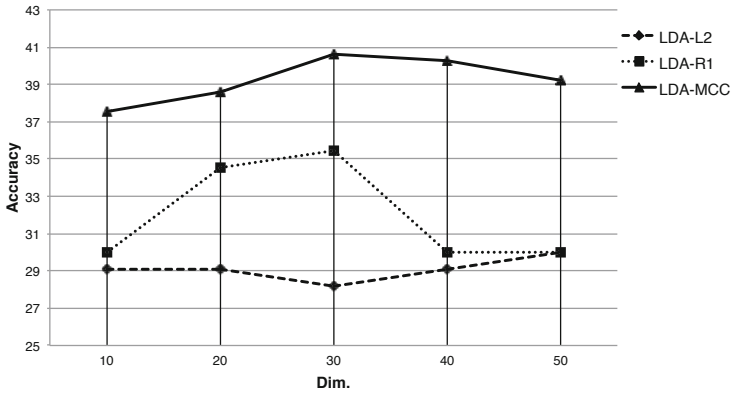

Fig. 3. Accuracy in Brodatz Texture Dataset

4.3 ORL Dataset

The third experiment is evaluated over the ORL dataset [12]. All images are gray scale and normalized to a resolution of 32 × 32 pixels. Among these 400 images, 30 percent were randomly selected and occluded with a rectangular noise consisting of random black and white dots whose size was 10 × 10, located at a random position. For a better illustration, some training samples are shown in Fig.4. 3 images per person are used for training and others are for testing. Simple 1-nearest-neighbor(1NN) classifier is used for the final classification. The performance is shown in Fig.5. The average number of iterations for LDA-MCC is 6.25 while 9.7 for LDA-R1. From this figure, we can see that the proposed method is the outstanding one and can obtain about 10 percent or 35 percent than LDA-R1 and LDA-L2, respectively.

Moreover, in this figure, when the reduced dimension is very small, the proposed method can get significant performance. In order to see how the accuracy changes in small dimension, another experiment is carried out and the result is

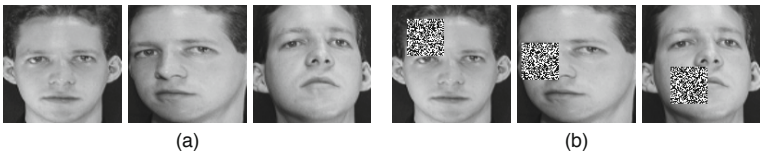

Fig. 4. ORL dataset (a) Original Images (b) Corresponding Images with occlusion

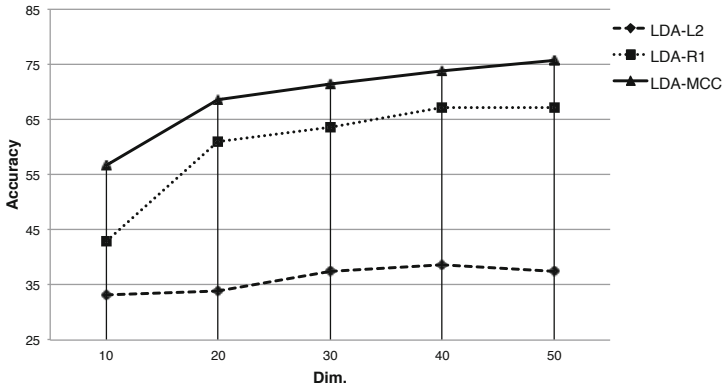

Fig. 5. Accuracy in ORL Dataset

shown in Fig.6. From this figure, we can see more clear about the effectiveness of the proposed method.

Finally, the accuracy on ORL dataset and the average training time are concluded in Table 1, here, PCA-L2 [13] means L2 norm based PCA while PCA-L1 [14] is L1 norm based PCA. From this table, we can see that our proposed method has higher performance than traditional ones.

Table 1. Recognition rate and computation cost on ORL dataset

method	Recognition Rate	Average number of iterations	Average time (s)
LDA-L2	38.6	/	/
LDA-R1	67.1	9.7	21.8
PCA-L2 [13]	49.5	/	/
PCA-L1 [14]	68.1	/	/
LDA-MCC	75.7	6.25	10.9

In next experiment, the proposed method is applied to face reconstruction problem and the performance is compared with those of other methods. We applied LDA-L2, LDA-R1, LDA-MCC and extracted various numbers of features.

By using only a fraction of features, we could compute the average reconstruction error with respect to the original unoccluded images as Eq.(14).

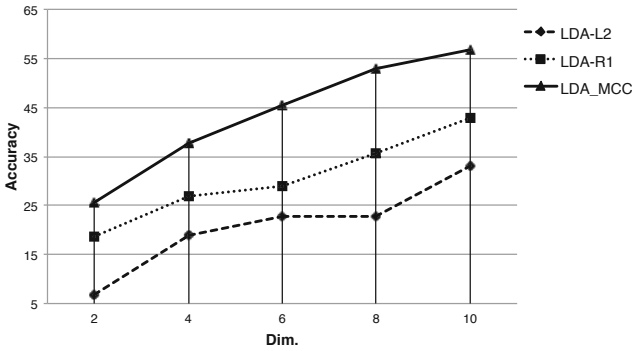

Fig. 6. Classification results for small dimension in ORL Dataset

$$e(m) = \frac{1}{n}\sum_{i=1}^{n} ||X^{org} - WW^T X||_2 \tag{14}$$

Here, n is the number of samples, which is 400 in this case, X^{org} and X are the original unoccluded image and the image used in the training, respectively. Fig.7 shows the average reconstruction errors for various numbers of extracted features. In this figure, even when the number of extracted features is small, the average reconstruction error of the proposed method is smaller than LDA-L2 and LDA-R1 approaches.

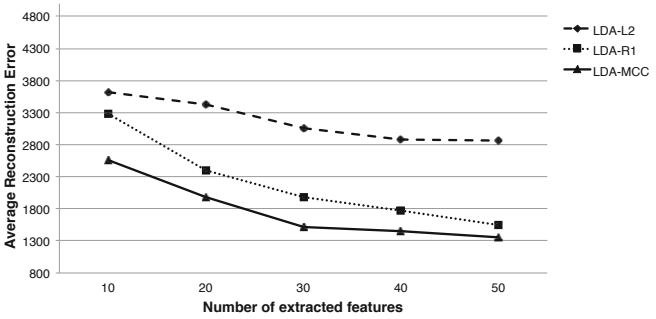

Fig. 7. Average reconstruction errors for ORL dataset

4.4 AR Dataset

The AR [15] dataset consists of over 3,200 color images of the frontal images of faces of 126 subjects. Each subject has 26 different images, including frontal views of with different facial expressions, lighting conditions and occlusions. For each subject, these images were recorded in two different sessions which are separated by two weeks, each session consisting of 13 images. For the experiments reported in this section, 60 different individuals were randomly selected from this database. Then there are 1560 images in our experiments. All images were manually cropped and resized to 80 by 60. Some example images of one person are shown in Fig.8.

Fig. 8. Some samples from AR dataset

In this evaluation, the recognition performances of the different algorithms on AR database are compared. Six samples of each individual are randomly selected as gallery (training images), and the remaining ones are used for probe (testing images). In our study, we perform 5 times to randomly choose the training set and calculate the average recognition rates. Some classification results are listed in Fig.9, where we can see that the proposed method has higher performance than LDA-L2 and LDA-R1. In general, LDA-MCC can obtain about 10 percent or 20 percent than LDA-R1 and LDA-L2, respectively. And the average number of iterations for LDA-MCC is 10.1 while 25.3 for LDA-R1. Thus, we can see clearly that LDA-R1 takes much more computation cost to achieve convergence in larger dimensional input space, such as face recognition application, than LDA-MCC. Base on this evaluation, our proposed method is more effective and efficient than the traditional approaches to solve facial expression, illumination or occlusions issues.

In Fig.10, only low-dimensional space is focused on since we want to make a comparison of the most discriminant features for the proposed method and some related algorithms. Same as Fig.9, the proposed methods can extract more discriminant features.

Finally, the average accuracy and time cost for training on AR dataset is concluded in Table 2, and our proposed methods are superior to the traditional approaches.

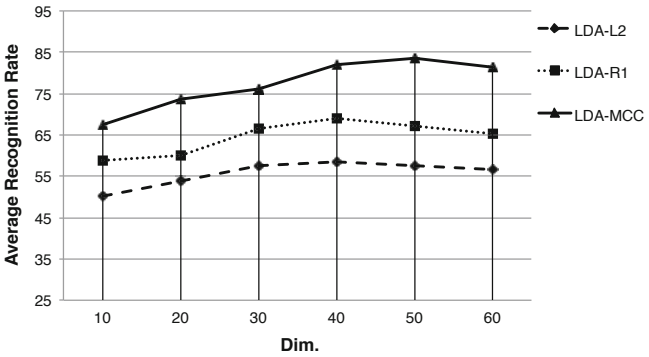

Fig. 9. Classification results in AR Dataset

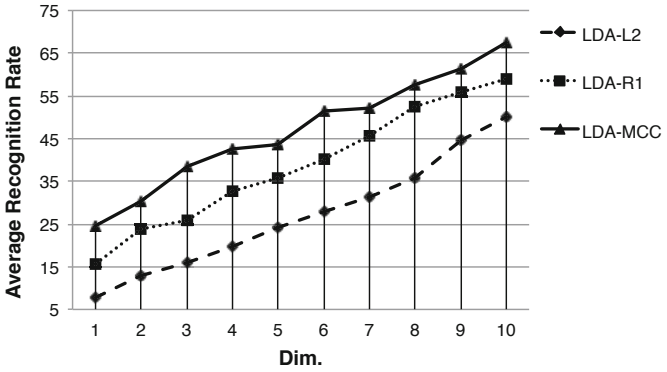

Fig. 10. Classification results for small dimension in AR Dataset

Table 2. Recognition rate and computation cost on AR dataset

method	Recognition Rate	Average number of iterations	Average time (s)
LDA-L2	58.6	/	/
LDA-R1	69.1	25.3	87.8
PCA-L2 [13]	55.2	/	/
PCA-L1 [14]	73.2	/	/
LDA-MCC	83.7	10.1	30.5

5 Conclusions and Future Work

In this paper, we proposed a novel method of LDA with MCC, which better characterizes the between-class separability. The proposed objective function is robust to outliers and can be efficiently optimized by the half-quadratic optimization technique. For each iteration step, the complex correntropy objective can

be reduced to a weighted traditional LDA optimization problem. The proposed subspace method not only successfully suppresses the negative effects of outliers but also it is invariant to rotations. Experimental results have demonstrated the effectiveness of the proposed method compared to the existing approaches.

In out future work, first, how to apply MCC to with-class distance and how to extend MCC to matrix or tensor based LDA will be studied. Second, some specific applications, such as facial expression recognition, using the proposed method will be evaluated.

References

1. Geng, Y., Shan, C., Hao, P.: Square loss based regularized lda for face recognition using image sets. In: CVPRW, pp. 99–106 (2009)
2. Landon, J., Jeffs, B., Warnick, K.: Model-based subspace projection beamforming for deep interference nulling. IEEE Transactions on Signal Processing 60, 1215–1228 (2012)
3. McLachlan, G.J.: Discriminant analysis and statistical pattern recognition. Wiley (1992)
4. Zhao, W., Chellappa, R., Krishnaswamy, A.: Discriminant analysis of principal components for face recognition. In: 3rd International Conference on Automatic Face and Gesture Recognition (1998)
5. Li, X., Hu, W., Wang, H., Zhang, Z.: Linear discriminant analysis using rotational invariant l1 norm. Neurocomputing, 2571–2579 (2010)
6. Liu, W., Pokharel, P.P., Principe, J.C.: Correntropy: Properties and applications in non-gaussian signal processing. IEEE Trans. Signal Process 55, 5286–5298 (2007)
7. Parzen, E.: On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33, 1065–1076 (1962)
8. Rockfellar, R.: Convex analysis. Princeton Univ., Princeton (1970)
9. Yuan, X., Hu, B.: Robust feature extraction via information theoretic learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1193–1200 (2009)
10. Silverman, B.W.: Density estimation for statistics and data analysis. Chapman and Hall, London (1986)
11. `http://www.ux.uis.no/tranden/brodatz.html`
12. `http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html`
13. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: IEEE Conference on Computer Vision and Pattern Recognition (1991)
14. Kwak, N.: Principal component analysis based on L1-norm maximization. IEEE Trans. Pattern Anal. Mach. Intell. 30, 1672–1680 (2008)
15. Martnez, A., Benavente, R.: The ar-face database. CVC Technical Report 24 (1998)