Optical Engineering

SPIEDigitalLibrary.org/oe

Enhanced discriminative locality alignment and its kernel extension

Zhonghua Liu
Zhong Jin

SPIE

Enhanced discriminative locality alignment and its kernel extension

Zhonghua Liu
Nanjing University of Science and Technology
School of Computer Science and Technology
Nanjing, Jiangsu 86 China
and
Electronic Information Engineering College
Henan University of Science and Technology
Luoyang, China
E-mail: lzhua@yahoo.com

Zhong Jin
Nanjing University of Science and Technology
School of Computer Science and Technology
Nanjing, Jiangsu 86 China

Abstract. Although discriminative locality alignment (DLA), which is based on the idea of part optimization and whole alignment, has better performance than classical methods in feature extraction, DLA is too overly sensitive to the values of the parameters and falls short of exploiting the full supervision information. We propose a novel supervised feature extraction method, named enhanced discriminative locality alignment (EDLA), for robust feature extraction. EDLA is not sensitive on the choice of the parameters, and both the local structure and class label information are taken into consideration in EDLA algorithm. Moreover, a kernel version of EDLA, named kernel EDLA, is developed through applying the kernel trick to EDLA to increase its performance on nonlinear feature extraction. Experiments on the face databases demonstrate the effectiveness of our methods. *© 2011 Society of Photo-Optical Instrumentation Engineers (SPIE).* [DOI: 10.1117/1.3605477]

Subject terms: discriminative locality alignment; part optimization; whole alignment; enhanced discriminative locality alignment; kernel enhanced discriminative locality alignment.

Paper 110173R received Feb. 21, 2011; revised manuscript received May 9, 2011; accepted for publication Jun. 8, 2011; published online Jul. 28, 2011.

1 Introduction

Recently, graph-based methods for linear dimensionality reduction[1–4] have attracted much attention. These methods typically rely on a graph to capture the salient geometric relations of the data in the high-dimensional space. It has been observed that, in general, supervised graph-based methods outperform their unsupervised ones in various classification tasks. It is common practice to construct a class graph by only setting the adjacent nodes from the same class. That is, points from the same class will be mapped to be close to each other in the low-dimensional space. However, points from different classe's but nearby, may be projected to points that are close by in the low-dimensional space. On the basis of this fact, a methodology based on a repulsion graph to enhance the graph-based methods for dimensionality reduction was proposed.[5] The main idea is to repel points from different classes that are nearby in the input space, while at the same time taking advantage of the full available supervision information (class graph). The idea in Ref. 5 has been used in another context in graph drawing techniques,[6,7] in which the main idea is that multiclass data points in high-dimensional space tend to move due to local intraclass attraction or interclass repulsion, and after being embedding into a low-dimensional subspace, data points in the same class form a compact submanifold, whereas the gaps between submanifolds corresponding to different classes become wider than before. The main difference is that the method in Refs. 6 and 7 does not use the class graph to define the attraction forces. Instead it relies entirely on the k-NN graph to model both the attraction part and repulsion parts. Hence, it falls short of exploiting the full supervision information. However, those points that belong to the same class and are not neighbors can bring significant discriminant information to the dimensionality reduction matrix.

0091-3286/2011/$25.00 © 2011 SPIE

Discriminative locality alignment (DLA)[8,9] is a recently proposed dimensionality reduction method. The main idea of DLA is generally described as three steps. First, discriminative information is imposed over patches, each of which is associated with one sample and its neighbors. Then, each part optimization is weighted by margin degree. Finally the alignment trick is used to align all of the weighted part optimizations to the whole optimization. The advantages of DLA are as follows: First, it distinguishes the contribution of each sample for discriminative subspace selection, which overcomes the nonlinearity of the distribution of samples. Second, it preserves discriminative information over local patches. Third, it avoids the matrix singularity problem. However, the DLA algorithm does not exploit the full supervision information, and the recognition performance is heavily reliant on the choice of the parameters.

Therefore, an improved method, named enhanced discriminative locality alignment (EDLA), is first proposed in this paper. The key idea of the proposed EDLA algorithm is that both the local structure and all class label information are taken into consideration. Though the proposed EDLA model can preserve the local structure of the data space, it is intrinsically linear. The EDLA model may fail to extract the desired features when the data structure is nonlinear. Kernel-based approaches[10–14] can discover the nonlinear structure of the images. That is, it can change the distribution of samples by nonlinear mapping. Thus, some linearly inseparable samples in the original feature space may become linearly separable in the high-dimensional feature space. Therefore, kernel-enhanced discriminative locality alignment (KEDLA) is further proposed in this paper.

The remainder of this paper is organized as follows: Section 2 introduces discriminative locality alignment method. Section 3 analyzes deficiencies of DLA and proposes EDLA method. In Sec. 4, in order to attack the distribution nonlinearity of the data, KEDLA is further proposed.

Section 5 demonstrates the detailed experimental results on image databases. The final section gives our conclusions.

2 Discriminative Locality Alignment

In DLA algorithm,[8,9] for each sample, one patch is first built by the given sample and its neighbors, which may include the samples from not only the same class but also different classes from the given sample. For each patch, an objective function is designed to preserve the local discriminative information first. Then, margin degree, which is defined for each sample as a measure of the sample importance in contributing classification, is used to weight each patch. Finally, all the weighted part optimizations are integrated together to form a global coordinate according to the alignment trick[15–17] and the projection matrix can be obtained by solving a standard eigendecomposition problem.

2.1 *Part Optimization*

Consider a data set Xdata, which consists of N samples in a high-dimensional space R^{m}. That is, $X\text{data} = [x_1,\ldots,x_N] \in R^{m\times N}$. For a given sample x_i, k_1 nearest neighbors with respect to x_i from samples in the same class with x_i are selected and denoted by $x_{i^1},\ldots,x_{i^{k1}}$. k_2 nearest neighbors with respect to x_i from samples in different classes with x_i are selected and denoted by $x_{i_1},\ldots,x_{i_{k_2}}$. By putting $x_{i^1},\ldots,x_{i^{k_1}}$ and $x_{i_1},\ldots,x_{i_{k_2}}$ together, the local patch for the sample x_i can be expressed as $X_i = [x_i, x_{i^1},\ldots,x_{i^{k_1}}, x_{i_1},\ldots,x_{i_{k_2}}]$.

For each patch, the corresponding output in the low dimensional space is denoted as $Y_i = [y_i, y_{i^1},\ldots,y_{i^{k1}}, y_{i_1},\ldots,y_{i_{k_2}}]$. In the low-dimensional space, we expect that distances between the given sample and neighbor samples of a same class are as small as possible, while distances between the given sample and the neighbor samples of different classes are as large as possible. That is, intraclass compactness and interclass separability are achieved, simultaneously. Thus, we have

$$\begin{aligned}&\arg\min_{y_i}\sum_{j=1}^{k_1}\|y_i - y_{i^j}\|^2\\ &\arg\max_{y_i}\sum_{p=1}^{k_2}\|y_i - y_{i_p}\|^2.\end{aligned} \tag{1}$$

Because the patch formed by the local neighborhood can be regarded approximately linear,[18] the part discriminator by using the linear manipulation is formulated as follows:

$$\arg\min_{y_i}\left(\sum_{j=1}^{k_1}\|y_i - y_{i^j}\|^2 - \beta\sum_{p=1}^{k_2}\left\|y_i - y_{i_p}\right\|^2\right), \tag{2}$$

where $\beta(\beta\in[0,1])$ is the penalty parameter. The weight vectors are defined as follows:

$$W_i = [\overbrace{1,\ldots,1}^{k_1}, \ -\beta*\overbrace{(1,\ldots,1)}^{k_2}]. \tag{3}$$

Then, Eq. (2) reduces to

$$\begin{aligned}&\arg\min_{y_i}\left[\sum_{j=1}^{k_1}\|y_i - y_{i^j}\|^2 W_i(j) + \sum_{p=1}^{k_2}\|y_i - y_{i_p}\|^2 W_i(p+k_1)\right]\\ &= \arg\min_{y_i}\sum_{j=1}^{k_1+k_2}\|y_{F_i\{1\}} - y_{F_i\{j+1\}}\|^2 W_i(j),\\ &= \arg\min_{Y_i} tr\left(Y_i L_i Y_i^T\right)\end{aligned} \tag{4}$$

where $F_i = \{i, i^1,\ldots,i^{k_1}, i_1,\ldots,i_{k_2}\}$ is the index set for the i'th

$$\text{patch, } L_i = \begin{bmatrix}\sum_{j=1}^{k_1+k_2} W_i(j) & -W_i\\ -W_i^T & \operatorname{diag}(W_i)\end{bmatrix}.$$

2.2 *Sample Weighting*

Margin degree is used to quantify the importance of a sample x_i for discriminative subspace selection and is defined as follows:

$$m_i = \exp\left[-\frac{1}{(n_i+\delta)t}\right], \qquad i = 1,\ldots,N, \tag{5}$$

where n_i is the number of samples with different class with the given x_i in the ε-ball centered at x_i, δ is a regularization parameter, and t is a scaling factor.

Therefore, when the part optimization of the i'th patch is weighted by the margin degree of the i'th sample, Eq. (4) can be written as

$$\arg\min_{Y_i} m_i tr\left(Y_i L_i Y_i^T\right) = \arg\min_{Y_i} tr\left(Y_i m_i L_i Y_i^T\right). \tag{6}$$

2.3 *Whole Alignment*

For each patch X_i, there is a low-dimensional representation Y_i. All Y_i can be unified together as a whole by assuming that the coordinate for the i'th patch Y_i is selected from the global coordinate $Y = [y_1,\ldots,y_N]$, such that

$$Y_i = Y S_i, \tag{7}$$

where $S_i \in R^{N\times(k_1+k_2+1)}$ is the selection matrix and an entry is defined as

$$(S_i)_{pq} = \begin{cases}1, & \text{if} \quad p = F_i\{q\}\\ 0, & \text{else.}\end{cases}. \tag{8}$$

Then, Eq. (6) can be written as

$$\arg\min_{Y} tr\left(Y S_i m_i L_i S_i^T Y^T\right). \tag{9}$$

By summing over all the part optimizations described as Eq. (9), the whole alignment is

$$\begin{aligned}&\arg\min_{Y} tr(Y S_i m_i L_i S_i^T Y^T)\\ &= \arg\min_{Y} tr\left[Y\left(\sum_{i=1}^{N} S_i m_i L_i S_i^T\right)Y^T\right],\\ &= \arg\min_{Y} tr(Y L Y^T)\end{aligned} \tag{10}$$

where $L = \sum_{i=1}^{N} S_i m_i L_i S_i^T \in R^{N\times N}$ is the alignment matrix. It is obtained based on an iterative procedure,

$$L(F_i, F_i) \leftarrow L(F_i, F_i) + m_i L_i \tag{11}$$

for $i = 1,\ldots, N$, with the initialization $L = 0$.

To obtain the linear and orthogonal projection matrix U, such as $Y = U^T X$, Eq. (10) is deformed as follows:

$$\arg\min_U tr(U^T X L X^T U) \quad s.t.\ U^T U = I. \tag{12}$$

The transformation matrix U that minimizes the objective function is given by the minimum eigenvalue solution to the standard eigenvalue problem,

$$XLX^T U = \lambda U. \tag{13}$$

3 Enhanced Discriminative Locality Alignment

In DLA algorithm, DLA has three particular advantages: (*i*) It can deal with the nonlinearity of the distribution of samples while preserving the discriminative information because of focusing on the local patch of each sample; (*ii*) because the neighbor measurements of different classes are considered, it well preserves discriminability of classes; and (*iii*) because it obviates the need to compute the inverse of a matrix, it avoids the small-sample-size problem. However, DLA has some deficiencies as follows: (*i*) The local patch for any sample x_i is constructed by k_1 nearest neighbors with same class label of x_i and k_2 nearest neighbors with different class label of x_i. That is to say, k_1 and k_2 are invariable for any sample; (*ii*) the resulting performance is too overly sensitive to the values of the parameters k_1 and k_2; (*iii*) it only use the part class graph to define the attraction forces; and (*iv*) the weight vectors W makes two points become the same point in the feature space if they belong to the same class or they are neighbor samples from different classes.

The main goal of graph-based methods is to preserve the properties of a graph representing the affinity between data points in local neighborhoods of the high-dimensional space. It has been observed that supervised graph-based methods outperform their unsupervised methods in classification tasks. On the basis of this fact, a method based on class graph and repulsion graph was proposed.[5] The method has high recognition performance. It is also generic and can be applied to any graph-based method for linear dimensionality reduction.

Therefore, a novel method called EDLA is proposed in this paper. Both the local structure and all class label information are taken enough consideration in EDLA. Because the second and third stages of the proposed EDLA are the same as those of DLA, and the difference between DLA and EDLA mainly focuses on the first stage—part optimization.

Consider a data set represented by the columns of a matrix $X\text{data} = [x_1, x_2,\ldots, x_N] \in R^{m\times N}$. Each sample x_i belongs to one of the c classes, and n_i is the number of data samples, which belong to the i'th class. For a given sample x_i, we select those samples that belong to the same class with x_i and term them class samples denoted by $x_{i^1},\ldots, x_{i^{n_i-1}}$. In order to determine how each sample is influenced by its class samples, Gaussian weights are often used to do it. The weight matrix is defined as follows:

$$W_{ij}^{(c)} = \exp\left(\frac{-\|x_i - x_j\|^2}{t}\right) \quad (\text{if } c_i = c_j), \tag{14}$$

where $1 \le i \le N$, $1 \le j \le n_i - 1$. The Euclidean distance $\|x_i - x_j\|^2$ is in the exponent, and the parameter t is used as a regulator. It is difficult to determine the appropriate value of the parameter t. However, in experiment part, a method for determining a good value for the width t of the Gaussian envelope is employed, and it usually gives a good and reasonable estimate for the value of t.[5]

A repulsion graph is one that is extracted from the k-NN graph, based on class label information. For a given sample x_i, its k nearest neighbors is used, and we select those samples that retain among the k nearest neighbors of a sample x_i and are not in the same class as sample x_i term repulsion samples denoted by $x_{i_1},\ldots, x_{i_{m_i}}$ $(0 \le m_i \le k)$. We define the weight coefficients of sample x_i and repulsion samples as follows:

$$W_{ij}^{(r)} = 1 \quad [x_i \in knn(j) \vee x_j \in knn(i)] \wedge (c_i \ne c_j). \tag{15}$$

Alternative weights were proposed for the repulsion graph in Ref. 5, and it was defined as follows:

$$W_{ij}^{(r)} = \frac{1}{\sigma + \|x_i - x_{i_j}\|^2 / \left(\|x_i\|^2 + \|x_{i_j}\|^2\right)}, \tag{16}$$

where $1 \le i \le N$, $1 \le j \le m_i$, $\|\cdot\|$ is the L2 norm, and the parameter σ is used as a regulator. Although it requires an additional parameter, the authors experimentally have demonstrated that the resulting performance is not too overly sensitive on the choice of σ, and what's more, it often gives markedly better results than constant weights. In experiments, the parameter σ is set to 7.

The local patch for the sample x_i is constructed by putting $x_{i^1},\ldots, x_{i^{n_i-1}}$ and $x_{i_1},\ldots, x_{i_{m_i}}$ together as $X_i = [x_i, x_{i^1},\ldots, x_{i^{n_i-1}}, x_{i_1},\ldots, x_{i_{m_i}}]$. The corresponding weight vector is

$$W_i^* = \left[W_i^{(c)}(i,:), -\beta \cdot W_i^{(r)}(i,:)\right]. \tag{17}$$

Therefore, the patch objective function of EDLA can be expressed as

$$\arg\min_{y_i} \sum_{j=1}^{k_1+k_2} \|y_i - y_{ij}\|^2 W_i^*(j). \tag{18}$$

4 Kernel Enhanced Discriminative Locality Alignment

The kernel trick is widely used to enhance the separability of the linear supervised dimensionality reduction algorithms. The EDLA can be further improved by using the kernel trick.

The nonlinear mapping Φ is used to map the input data R^N into a Hilbert space [i.e., $x \mapsto \Phi(x)$]. We implement EDLA in the Hilbert space $\Phi(x) = [\Phi(x_1),\ldots,\Phi(x_n)]$ The eigenvector problem of Eq. (13) in the Hilbert space can be written as follows:

$$[\Phi(X) L^{\Phi} \Phi^T(X)] U = \lambda U. \tag{19}$$

We formulate EDLA with the dot product to generalize it to the nonlinear case. The dot product in the Hilbert

space is presented with a kernel function [i.e., $K(x_i,x_j) = \Phi(x_i)^T\Phi(x_j)$]. Because any eigenvector may be expressed by a linear combination of the observations in feature space, there exist coefficients v_i, $i = 1,\ldots,N$ such that

$$U = \sum_{i=1}^{N} v_i \Phi(x_i), \tag{20}$$

where $v = [v_1,\ldots,v_N]^T$. Thus, Eq. (19) can be rewritten as

$$\begin{aligned}
&[\Phi(X)L^{\Phi}\Phi^T(X)]\sum_{i=1}^{N} v_i\Phi(x_i) = \lambda\sum_{i=1}^{N} v_i\Phi(x_i)\\
&[\Phi(X)L^{\Phi}\Phi^T(X)]\Phi(X)v = \lambda\Phi(X)v\\
&[\Phi^T(X)\Phi(X)]L^{\Phi}[\Phi^T(X)\Phi(X)]v = \lambda[\Phi^T(X)\Phi(X)]v\\
&KL^{\Phi}Kv = \lambda Kv.
\end{aligned} \tag{21}$$

The solution of Eq. (21) is denoted by $v^* = [v^1,\ldots,v^d]$ whose column vectors are the d eigenvectors corresponding to the first d smallest eigenvalues. For a test sample x, we compute projections onto the eigenvectors U^k according to

$$[U^k \cdot \Phi(x)] = \sum_{i=1}^{N} v_i^k[\Phi(x)\cdot\phi(x_i)] = \sum_{i=1}^{N} v_i^k K(x,x_i), \tag{22}$$

where v_i^k is the i'th element of the vector v^k.

Note that the distance matrix for KEDLA may be different from that for EDLA and can be computed by kernel function $K(x, y)$ [assume that $K[x_i, x_j] = \Phi(x_i)^T\Phi(x_j)$]. The distance between sample x_i and x_j in the higher dimensional Hilbert space is obtained as follows:

$$\begin{aligned}
D(x_i,x_j) &= (\Phi(x_i)-\Phi(x_j))^T(\Phi(x_i)-\Phi(x_j))\\
&= \Phi(x_i)^T\Phi(x_i) - 2*\Phi(x_i)^T\Phi(x_j) + \Phi(x_j)^T\Phi(x_j)\\
&= K(x_i,x_i) - 2K(x_i,x_j) + K(x_j,x_j).
\end{aligned} \tag{23}$$

5 Experiments and Results

In this section, we implemented experiments on Olivetti Research Laboratory (ORL), YALE, and UMIST databases to evaluate the proposed algorithm. First, we determine the parameters (i.e., the parameter k in k-NN graph-based methods, the penalty parameter β, kernel parameters) and, second, we evaluated the performance of the proposed algorithms on recognition accuracy.

For the sake of completeness, we compare the performance of PCA,[19] LDA,[20] DLA, and the proposed EDLA. Especially for DLA, we select the optimal values of k_1 (the number of neighbor samples from an identical class) and k_2 (the number of neighbor samples from different classes) in the experiments.

In order to avoid singularity problem and reduce computational complexity, for all algorithms but PCA itself, the first step is the PCA projection. We retain $N - C$ dimensions in the PCA step for LDA method, and PCA subspace is set $N - 1$ dimensions for other methods.

5.1 *Selection of Parameters*

The parameter t in Gaussian weights is selected as follows: 1000 sample points are randomly from the data set (if your data set has <1000 points, then all of them are use.) and the pairwise distances among them are computed. Then, t is set to half the median of those pairwise distances.[5] The parameter k in the k-NN graph is selected as follows: for simplicity, we assume that each class has the same sample number m. Therefore, we generally select $k = m - 1$.

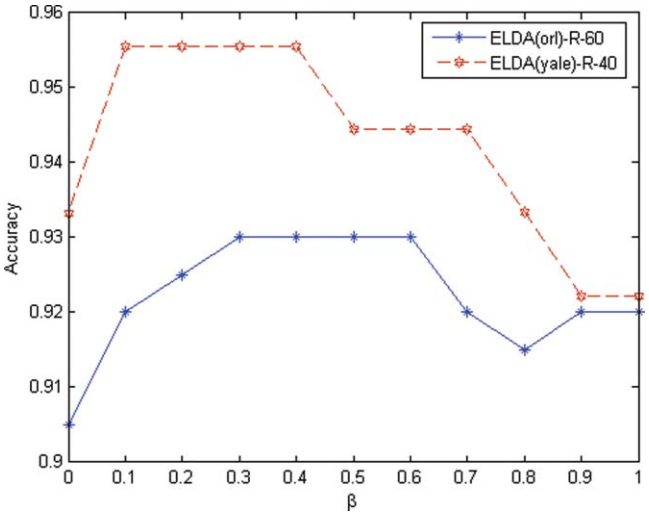

Fig. 1 Performance of EDLA for varying β.

There is still one free parameter to be determined in EDLA algorithm. It is the penalty parameter β, which is a common parameter in both EDLA and DLA. In order to illustrate the behavior of the EDLA with respect to the penalty parameter β, we use the ORL face database and the Yale database to test it. We randomly choose five different images per individual to form the training set. The rest of each person is used for the test. The experiment is repeated for 10 times on each database. The reduced dimension in experiments is set to 60 for the ORL face database and to 40 for the Yale face database. β is varied from 0 to 1, and Fig. 1 plots the average recognition accuracy of EDLA with respect to β.

It can be seen from Fig. 1 that the recognition rates of EDLA vary along with the change in β. The average recognition rates of EDLA in two face databases reach the highest recognition rate when β is 0.3 or 0.4. Therefore, we can set β to 0.3 or 0.4 for experiments. For simplicity, we generally select $\beta = 0.3$ in the next experiments.

5.2 *Face Recognition Results*

In this section, experiments are designed to evaluate the efficacy of EDLA and KEDLA algorithms. We use three data sets that are publicly available for the experiments. PCA and LDA are taken as the baseline methods for comparison. In addition, two popular kernels are involved in our experiments. One is the polynomial kernel $k(x, y) = (1 + x^{Ty})^d$, and the other is the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/t)$. For KEDLA, we use these two kernels, respectively. The optimal kernel parameters are selected.

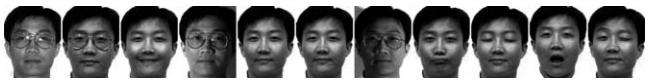

Fig. 2 Sample images of one person on Yale database.

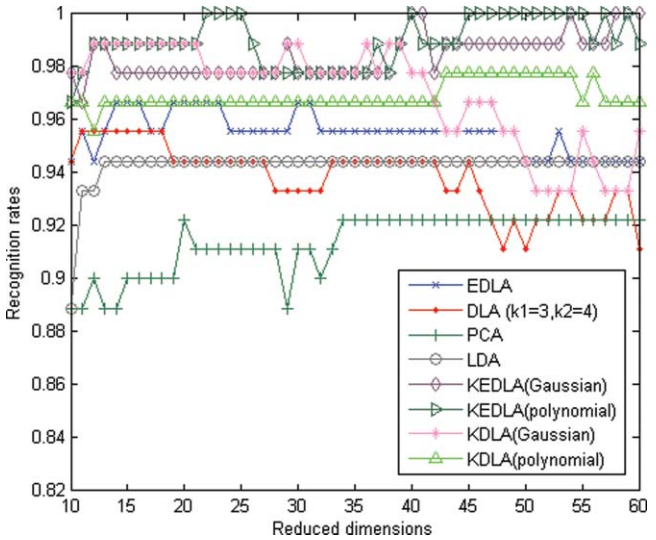

Fig. 3 Average recognition rates versus the dimensions on Yale database.

5.2.1 *Experiments on Yale face database*

The Yale database[20] contains 165 gray-scale images of 15 individuals. The images demonstrate variations in lighting, facial expression, and faces with or withou glasses. In our experiment, every image was manually cropped and resized to 100 × 80 pixels. Eleven sample images of one individual are displayed in Fig. 2.

In the experiment, five images per individual are randomly chosen for training and the remaining six images are used for testing. The experiment is repeated for 10 times. We experiment with the dimension of the reduced space $d = [10{:}60]$ (in Matlab notation), and for each value of d, we plot the recognition rate. Figure 3 shows the average recognition rates versus the dimensions. From Fig. 3, we can see three main points. First, the proposed EDLA outperforms DLA, PCA, and LDA algorithms. Second, no matter which kernel is used, our KEDLA almost outperforms other six algorithms. From the first and second points, we can see that kernel approach can indeed improve the face recognition accuracy. The maximal recognition rates and the corresponding dimensions are given in Table 1. From Table 1, it can be seen that the maximal average recognition rates of KEDLA with polynomial kernel and Gaussian kernel achieve 100%, respectively.

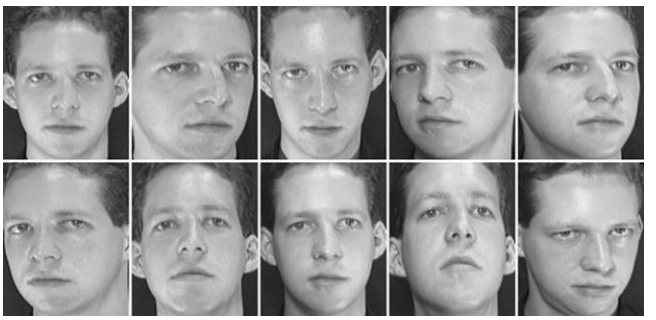

Fig. 4 Sample images of one person on ORL database.

5.2.2 *Experiments on Olivetti Research Laboratory face database*

The ORL face database[21] contains 400 images of 40 individuals (10 samples per person). The images are captured at different times and have different variations, including expressions (open or closed eyes, smiling or nonsmiling) and facial details (glasses or no glasses). The images were taken with a tolerance for some tilting and rotation of the face up to 20 deg. The size of each face image is 112×92 pixels. Ten sample images of one individual are displayed in Fig. 4.

Table 1 Maximal average recognition rates of EDLA, DLA, PCA, LDA, KEDLA, and KDLA on Yale database and the corresponding dimensions and parameters.

	EDLA	DLA	PCA	LDA	KEDLA (%)		KDLA (%)	
Method	(%)	(%)	(%)	(%)	Polynomial	Gaussian	Polynomial	Gaussian
Recognition rate	96.67	95.56	92.22	94.44	100	100	97.78	98.89
Dimension	14	11	20	13	22	40	43	12
Parameter	$\beta = 0.3$	$k1 = 3$, $k2 = 4$	/	/	$d = 0.2$	$t = 2$	$d = 0.8$	$t = 3$

Table 2 Maximal average recognition rates of EDLA, DLA, PCA, LDA, KEDLA, and KDLA on ORL database and the corresponding dimensions and parameters.

	EDLA	DLA	PCA	LDA	KEDLA (%)		KDLA (%)	
Method	(%)	(%)	(%)	(%)	Polynomial	Gaussian	Polynomial	Gaussian
Recognition rate	93.5	91.5	90.5	90.5	95	93.5	90	91
Dimension	19	32	78	39	111	33	30	42
Parameter	$\beta = 0.3$	$k1 = 2$, $k2 = 3$	/	/	$d = 0.4$	$t = 3$	$d = 1.5$	$t = 1$

Table 3 Maximal average recognition rates of EDLA, DLA, PCA, LDA, KEDLA, and KDLA on UMIST database and the corresponding dimensions and parameters

	EDLA	DLA	PCA	LDA	KEDLA (%)		KDLA (%)	
Method	(%)	(%)	(%)	(%)	Polynomial	Gaussian	Polynomial	Gaussian
Recognition rate	97.01	94.02	90.57	95.86	99.08	97.93	93.79	94.71
Dimension	74	22	63	12	81	15	30	77
Parameter	$\beta = 0.3$	$k1 = 2$, $k2 = 2$	/	/	$d = 0.7$	$t = 4$	$d = 0.1$	$t = 3$

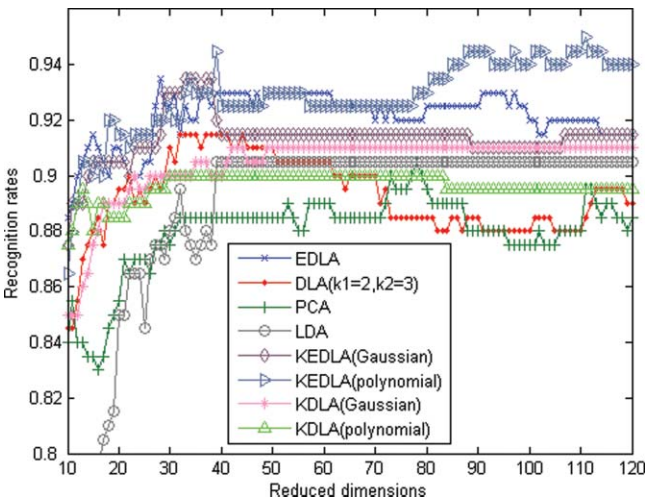

Fig. 5 Average recognition rates versus the dimensions on ORL database.

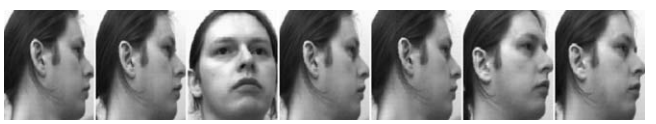

Fig. 6 Some sample images of one person on UMIST database.

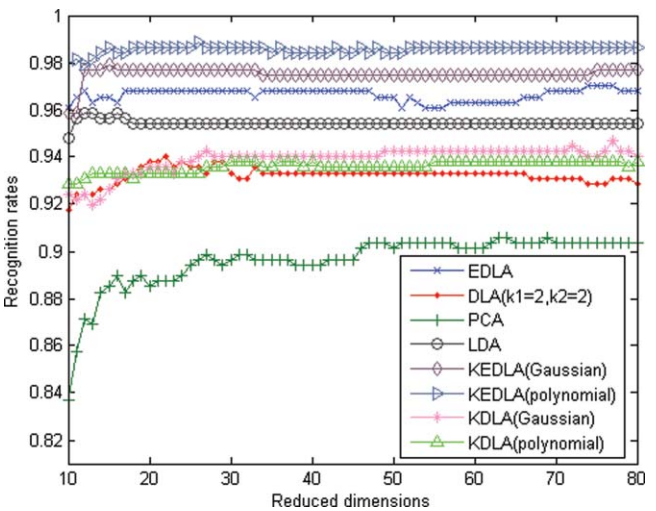

Fig. 7 Average recognition rates versus the dimensions on UMIST database.

For this experiment, five images per individual are randomly chosen for training and the remaining five images are used for testing. The experiment is repeated for 10 times. Figure 5 illustrates the average recognition rates of each method versus the dimensions. Table 2 lists the maximal average recognition rate of each method across 10 runs and the corresponding dimension and parameter. From Fig. 5, it can be seen that EDLA outperforms DLA, PCA, and LDA algorithms irrespective of the variation of the dimensions. From Table 2, we can also see that the recognition rates of KEDLA and EDLA are close and KEDLA with polynomial kernel performs a little better.

5.2.3 *Experiments on UMIST face database*

The UMIST database[22] consists of a total 575 face images of 20 people. The individuals are a mix of race, sex, and appearance and are photographed in a range of poses from profile to frontal views. The number of different views per subject varies from 19 to 48. The size of each face image is 112×92 pixels. Figure 6 shows sample images of one person.

For this experiment, we form the training set by a random subset of seven different poses per subject and the rest of the database is used for testing. The experiment is repeated for 10 times. The recognition results are illustrated in Fig. 7. It is shown that the proposed EDLA and KEDLA consistently outperform other five algorithms irrespective of the variation of dimensions. Table 3 lists the maximal recognition rate and the corresponding dimension of each classification method. It shows that KEDLA has the best performance.

6 Conclusions

In this paper, we develop a new linear dimensionality reduction algorithm called EDLA for feature extraction. Similar to DLA, EDLA is also based on the idea of part optimization and whole alignment. However, significantly differing from the DLA, both the local structure and all class label information are taken into consideration in EDLA. EDLA can preserve more discriminative information and achieve the higher recognition performance. We also extend EDLA to the kernel space and develop a KEDLA algorithm. Experimental results demonstrate the effectiveness of our EDLA and KEDLA. Researches 9 and 23 show that unlabeled samples may be helpful to improve the classification performance. In the future, we will generalize EDLA by taking unlabeled samples into account and propose semisupervised EDLA.

Acknowledgments

This project is supported by NSFC of China (Grants No. 60632050, No. 60705006, No. 60873151, and No. 60973098).

References

1. M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. of Conf. on Advances in Neural Information Processing Syst.*, pp. 585–591 (2002).
2. S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science* **290**, 2323–2326 (2000).
3. S. Yan, D. Xu, et. al., "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(1), 40–51 (January 2007).
4. S. Yan, D. Xu, B. Zhang, and H. Zhang, "Graph-embedding: a general framework for dimensionality reduction," in *Proc. of IEEE Int. Conf. Comput. Vis. Pattern Recog.*, pp. 830–837 (2005).
5. E. Kokiopoulou and Y. Saad, "Enhanced graph-based dimensionality reduction with repulsion Laplaceans," *Pattern Recogn.* **42**, 2392–2402 (2009).
6. W. Zhang, X. Xue, H. Lu, and Y. Guo, "Discriminant neighborhood embedding for classification," *Pattern Recogn.* **39**(11), 2240–2243 (2006).
7. A. Noack, "An energy model for visual graph clustering," in *Proc. of 11th Int. Symp. on Graph Drawing*, Lecture Notes in Computer Science 2912, Springer, Berlin, pp. 425–436 (2004).
8. T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.* **21**(9), 1299–1313 (2009).
9. T. Zhang, D. Tao, and J. Yang, "Discriminative locality alignment," in *Proc. 10th European Conf. on Computer Vision*, pp. 725–738 (2008).
10. X. He and P. Niyogi, "Locality preserving projections," in *Proc. of Conf. on Advances in Neural Information Processing Systems*, pp. 585–591 (2003).
11. J. Li, S. Chu, et. al., "Adaptive data-dependent matrix norm based Gaussian kernel for facial feature extraction," *Int. J. Innovative Computing, Information and Control* **3**(5), 1263–1272 (2007).
12. S. Mika, G. Ratsch, J. Weston, and B. Scholkopf, "Fisher discriminant analysis with kernels," in *Proc. of IEEE Int. Workshop on Neural Networks for Signal Processing*, Vol. IX, pp. 41–48 (1999).
13. K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Networks* **12**, 181–201 (2001).
14. J. Li, J. Pan, and S. Chu, "Kernel class-wise locality preserving projection," *Inf. Sci.* **178**, 1825–1835 (2008).
15. Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM J. Sci. Comput.* **26**(1), 313–338 (2005).
16. D. Zhao, Z. Lin, and X. Tang, "Laplacian PCA and its applications," in *Proc. of IEEE Int. Conf. on Computer Vision*, pp. 1–8 (2007).
17. T. Zhang, D. Tao, X. Li, and J. Yang, "A unifying framework for spectral analysis based dimensionality reduction," in *Proc. of Int. Conf. on Neural Networks*, pp. 1670–1677 (2008).
18. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science* **290**, 2323–2326 (2000).
19. M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 586–591 (1991).
20. P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 711–720 (1997).
21. F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," presented at 2nd IEEE Workshop on Applications of Computer Vision, Sarasota, FL (December 1994).
22. D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," in *Face Recognition: From Theory to Applications*, Vol. 163, pp. 446–456, NATO ASI Series F, Computer and Systems Sciences (1998).
23. M. Belkin, P. Niyogi, and V. Sindhwani, "On manifold regularization," in *Proc. of 10th Int. Workshop on Artificial Intelligence and Statistics*, pp. 371–378 (2005).

Zhonghua Liu received his BS in computer engineering from the First Aeronautical Institute of the Air Force and his MS in computer software and theory from Xihua University in 1998 and 2005, respectively. He is pursuing his PhD in pattern recognition and intelligent systems from Nanjing University of Science and Technology, China. His current research interests include face recognition and detection, and image processing.

Zhong Jin received his BS in mathematics, MS in applied mathematics, and PhD in pattern recognition and intelligence systems from Nanjing University of Science and Technology (NUST), China, in 1982, 1984, and 1999, respectively. He is a professor in the Department of Computer Science, NUST, and previously was a research assistant at the Department of Computer Science and Engineering, Chinese University of Hong Kong from 2000 to 2001. He visited the Laboratoire HEUDIASYC, Universite de Technologie de Compiegne, France, from October 2001 to July 2002. He visited the Centre de Visio per Computador, Universitat Autonoma de Barcelona, Spain, as the Ramon y Cajal Program Research Fellow from September 2005 to October 2005. His current interests are in the areas of pattern recognition, computer vision, face recognition, facial expression analysis, and content-based image retrieval.