Contextual Constraints based Linear Discriminant Analysis

Zhen Lei*, Stan Li

Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun Donglu, Beijing 100190, China.

Abstract

Linear feature extraction methods such as LDA have achieved great success in pattern recognition and image processing area. For most existing methods, the image data is usually transformed into a vector representation and the contextual information among pixels is not exploited. However, image data distribute sparsely in high-dimension feature space and the dependence among neighboring pixels is important to represent a natural image. Therefore, in this paper, we propose a novel image contextual constraint based linear discriminant analysis (CCLDA) method by taking into account the pixel dependence of an image in subspace learning process. In this way, a more discriminative subspace could be learned especially in the case of small sample size. Extensive experiments on ORL, Extended Yale-B, PIE and FRGC databases validate the efficacy of the proposed method.

Keywords: Contextual constraint, Discriminant analysis, Face recognition

*Corresponding author. Tel: +86-10-62658306. Fax: +86-10-62632259
Email address: zhen.ray@gmail.com (Zhen Lei)

1. Introduction

Subspace learning, due to its efficacy and efficiency, has achieved great success in pattern recognition and computer vision area, such as face recognition and image retrieval etc. Linear Discriminant Analysis (LDA) (Belhumeur et al. (1997)) is one of the representative methods. The objective of LDA is to find a subspace that maximizes the sample distance from different classes and meanwhile minimizes the sample distance from the same class. Thus, the derived subspace is discriminative to classify different samples correctly. However, in real applications, due to the high dimension of feature (e.g. for one 100×100 image, the number of pixels is 10,000) and usually small number of samples, the classical LDA always fails; this is called small sample size (SSS) problem. To address this problem, a lot of work has been investigated and many LDA-variants have been proposed such as Null-LDA (Chen et al. (2000)), Direct LDA (Yu and Yang (2001)), Regularized LDA (Friedman (1989); Lu et al. (2004)) etc..

All the above LDA methods are based on vector representation. That is, before LDA, the images are transformed into vectors by a certain order. Recently, 2D-LDA (Ye et al. (2004)), which takes operations on image matrix directly was proposed. The most advantage of 2D-LDA is its low computational cost and it implicitly avoids small sample size (SSS) problem since the within class scatter is usually non-singular. Moreover, the 2D-LDA is claimed to retain the image spatial structure so that it is more appropriate than 1D-LDA to address image related problem.

Most of the existing subspace learning methods like PCA (Turk and Pentland (1991)), LDA, LLE (Roweis and Saul (2000)), LPP (He et al. (2005b)),

(a) (b)

Figure 1: A face image (left) and its re-shuffled one (right). For 1D-LDA classification result, there is no difference between these two forms. However, people almost cannot get any information except some noise from the right one.

NPE (He et al. (2005a)), LPDP (Gui et al. (2010)), consider the pixels in image independently, not taking into account their spatial relationship. Fig. 1 shows an original face image and its relative in which the pixels are re-shuffled. Obviously, we cannot recognize any face information from the right image. However, in vector based discriminant learning methods, if the pixels in every image are re-shuffled in the same way, it would result in the same classification performance even though their is no semantic sense in shuffled images any more. It is well known that images with certain pattern occupy specific manifold, which is constrained by contextual information, in high-dimension feature space. Therefore, contextual constraint in image is important for image understanding and will provide useful information for classification. One of the most successful work to model the contextual information is the Markov Random Fields (MRFs) (Huang et al. (2004); Dass and Jain (2001); Dass et al. (2002)) which derive the results by maximizing the posterior probability in Bayesian deduction framework. However, the

optimization by MRF is somewhat computational expensive and is easy to converge into local minima that limits its application.

Wang et al. (Wang et al. (2005)) proposed a novel image matching distance considering the spatial information. However, they didn't demonstrate how to integrate the contextual information into dimensionality reduction problem. In this paper, we propose contextual constraints based linear discriminant analysis (CCLDA) which incorporates the contextual information into linear discriminant analysis. The main difference of proposed method from the existing ones is that it takes into account the image contextual constraints during feature dimensionality reduction. Moreover, we study the relationship of CCLDA with LDA and 2D-LDA and indicate that LDA and 2D-LDA are essentially certain versions of CCLDA. It should be noted the proposed contextual constraints consideration could be incorporated not only into LDA, but also into other general subspace learning methods.

The remainder of this paper is organized as follows. Section 2 briefly reviews some related work of LDA and 2D-LDA. Section 3 details the CCLDA and its relationship with LDA and 2D-LDA. Experiments and discussion on ORL, Extended Yale-B, PIE, and FRGC databases are demonstrated in Section 4 and in Section 5, we conclude the paper.

2. Related Works

Linear discriminant analysis (LDA), due to its simplicity and effectiveness, achieved great success in pattern recognition. The essential idea of LDA is to find the optimal subspace that gathers the samples from the same class and meanwhile disperses the samples from the different ones. Given

Notation	Description
N	number of images in dataset
L	number of classes in dataset
C_i	the i-th class
N_i	number of images in class C_i
$\mathbf{X}_i^k$	i-th image sample in class C_k
$\mathbf{x}_i^k$	i-th sample in vector representation in class C_k
r	number of rows in an image
c	number of columns in an image
$\mathbf{u}_i$	the mean vector of samples in class C_i
$\mathbf{u}$	the mean vector of samples in dataset
$\mathbf{U}_i$	the mean matrix of samples in class C_i
$\mathbf{U}$	the mean matrix of samples in dataset

Figure 2: Notations

the data set $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}]$, the between class scatter matrix $\mathbf{S}_b$ and within class scatter matrix $\mathbf{S}_w$ are defined as

$$\begin{aligned} \mathbf{S}_w &= \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i}(\mathbf{x}_j^i - \mathbf{u}_i)(\mathbf{x}_j^i - \mathbf{u}_i)^T \\ \mathbf{S}_b &= \frac{1}{N}\sum_{i=1}^{L} N_i(\mathbf{u}_i - \mathbf{u})(\mathbf{u}_i - \mathbf{u})^T \end{aligned} \tag{1}$$

LDA aims to learn the projective directions $\mathbf{W}$ which maximize the ratio of between class scatter matrix to within class scatter one as

$$J = \frac{|\tilde{\mathbf{S}}_b|}{|\tilde{\mathbf{S}}_w|} = \frac{|\mathbf{W}^T\mathbf{S}_b\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_w\mathbf{W}|} \tag{2}$$

The optimal projection matrix $\mathbf{W}_{opt}$ can be obtained by solving the following eigen-value problem

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{W} = \mathbf{W}\Lambda \tag{3}$$

where Λ is the diagonal matrix whose diagonal elements are eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

In real applications, especially the image involving problem, the dimension of feature is usually much larger than the sample size, so the within scatter matrix is singular. It is called small sample size or under sample problem. On the other hand, even if the $\mathbf{S}_w$ is full, due to the high dimension of feature, the sample size is still too small to reflect the class manifold in feature space and brings out the over-fitting problem.

Later, some researchers pointed out that the vectorization in traditional LDA disturbed the structure of image which is important for classification. 2D-LDA which takes operation on image matrices directly rather than vectors is consequently proposed. Let the sample set be $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n]$, and $\mathbf{X}_i$ is the i-th image matrix with $r \times c$ size. The within-class scatter matrix $\mathbf{S}_w$ and the between-class scatter matrix $\mathbf{S}_b$ based on image matrix are defined as follows:

$$\begin{aligned} \mathbf{S}_w^{2d} &= \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i}(\mathbf{X}_j^i - \mathbf{U}_i)(\mathbf{X}_j^i - \mathbf{U}_i)^T \\ \mathbf{S}_b^{2d} &= \frac{1}{N}\sum_{i=1}^{L} N_i(\mathbf{U}_i - \mathbf{U})(\mathbf{U}_i - \mathbf{U})^T \end{aligned} \tag{4}$$

The 2D-LDA searches such optimal projections that after projecting the original data onto these directions, the trace of the ratio of the resulting between-class scatter matrix to the within-class scatter matrix is maximized.

Let $\mathbf{W}^{2d}$ denote a $r \times d$ ($d < r$) projection matrix, and the $r \times c$ image matrix $\mathbf{X}$ is projected onto $\mathbf{W}^{2d}$ through the following linear transformation:

$$\mathbf{Y} = \mathbf{W}^{2d^T}\mathbf{X} \tag{5}$$

where the resulting $\mathbf{Y}$ is a $d \times c$ matrix with smaller size than $\mathbf{X}$. 2D-LDA chooses $\mathbf{W}^{2d}$ so that the following object function is maximized:

$$J = \frac{|\mathbf{W}^{2d^T}\mathbf{S}_b^{2d}\mathbf{W}^{2d}|}{|\mathbf{W}^{2d^T}\mathbf{S}_w^{2d}\mathbf{W}^{2d}|} \tag{6}$$

The optimal projection matrix $\mathbf{W}_{opt}^{2d}$ can be obtained by solving the following eigen-value problem

$$\mathbf{S}_w^{2d^{-1}}\mathbf{S}_b^{2d}\mathbf{W}^{2d} = \mathbf{W}^{2d}\Lambda \tag{7}$$

where Λ is the diagonal matrix whose diagonal elements are eigenvalues of $\mathbf{S}_w^{2d^{-1}}\mathbf{S}_b^{2d}$.

The process described above is called the left unilateral 2D-LDA (Li and Yuan (2005); Xiong et al. (2005)). There are also other 2D-LDA versions like right unilateral 2D-LDA and bilateral 2D-LDA (Ye et al. (2004); Yang et al. (2005); Kong et al. (2005)) etc.

3. Contextual Constraints based Linear Discriminant Analysis (C-CLDA)

3.1. Formulation of CCLDA

As described in Section 1, it is believed that the contextual information could be utilized to improve the classification performance. As we all know, the essence of most subspace learning methods is to assign weights to different

pixels respectively and then sum them up to project the samples into low-dimension subspace to be best classified. Direct modeling the contextual information on pixels seems a little difficult. In this work, we adopt an alternative way to impose the contextual constraints on the weights instead of the pixels. The idea is that if the pixels in original image have consistent properties, the weights corresponding to them should also have similar values. In this way, we transfer the contextual constraints from the image to the weight. That is, if we arrange the weights to form a image (called weight image) according to the order of the pixels in original image, there is also contextual constraints on the weight image.

As stated above, intuitively, if the pixels are of the similar property or reflect the similar structure, the weights on them would have strong relationship, otherwise the weights on independent pixels would also be weakly related. Following this rational, we impose a constraints $J_2(\mathbf{w}) = \frac{1}{2}\sum_{i,j}(w_i - w_j)^2 S_{ij}$ [1] on traditional LDA to reformulate the object of discriminant analysis as

$$J = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{(\mathbf{w}^T \mathbf{S}_w \mathbf{w} + \eta J_2(\mathbf{w}))} \tag{8}$$

where S_{ij} describes the similarity of pixels i and j, and η is a coefficient to balance the trade-off between the training discriminant power and contextual constraints. The constraint function $J_2(\mathbf{w})$ gives a high penalty when the weights of related pixels differ too much. Due to the symmetry of S_{ij} in general case, the contextual constraints $J_2(\mathbf{w})$ on weight image can be

[1] For ease of interpretation, we discuss the one-dimensional subspace case here and it is easy to be extended to d-dimensional case.

formulated using the matrix operations further as follows.

$$\begin{aligned} J_2(\mathbf{w}) &= \frac{1}{2}\sum_{i,j}(w_i - w_j)^2 S_{ij} \\ &= \frac{1}{2}(\sum_{i,j} w_i^2 S_{ij} + \sum_{i,j} w_j^2 S_{ij} - 2\sum_{i,j} w_i w_j S_{ij}) \\ &= \sum_{i,j} w_i^2 S_{ij} - \sum_{i,j} w_i w_j S_{ij} \\ &= \sum_{i} w_i^2 D_{ii} - \sum_{i,j} w_i w_j S_{ij} \\ &= \mathbf{w}^T\mathbf{D}\mathbf{w} - \mathbf{w}^T\mathbf{S}\mathbf{w} \\ &= \mathbf{w}^T\mathbf{L}^w\mathbf{w} \end{aligned} \tag{9}$$

where $\mathbf{L}^w = \mathbf{D} - \mathbf{S}$ is the Laplacian matrix[2], and $\mathbf{D}$ is a diagonal matrix where $D_{ii} = \sum_j S_{ij}$. Thus, the objective of CCLDA (Eq. 10) can be reformulated as

$$J = \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{(\mathbf{w}^T\mathbf{S}_w\mathbf{w} + \eta\mathbf{w}^T\mathbf{L}^w\mathbf{w})} \tag{10}$$

The optimal projection $\mathbf{w}$ can be obtained by solving the following generalized eigen-value problem.

$$\mathbf{S}_b\mathbf{w} = \lambda(\mathbf{S}_w + \eta\mathbf{L}^w)\mathbf{w} \tag{11}$$

For a nature image, one assumption often used is that the pixels in local neighboring region have the consistent properties and reflect the similar image structure. Therefore, one way to define the contextual matrix $\mathbf{S}$ among

[2]The formulation of $J_2(\mathbf{w})$ is very similar to LPP. However, they are totally different. LPP describes the neighboring sample structure in high-dimensional feature space while in this paper, we use this formulation to model the weight contextual information among neighboring pixels in one image.

different weights of pixels is

$$S_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are neighbors} \\ \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

If pixel i is in the neighboring region of j or pixel j is in the neighboring region of i, the corresponding similarity of w_i and w_j is set to be 1, otherwise, the similarity is set to be 0. In this way, all the pixels in a local adjacent regions are treated equally and the pixels in different regions are assumed to be independent. It is known as a rigid constraint. However, sometimes the pixel values in neighboring regions may change rapidly such as pixels around the edge. It might be improper to impose the same weight on these pixels. Therefore, we could define the similarity matrix of weights in another way named soft constraint as follows.

$$S_{ij} = \begin{cases} e^{-||\mathbf{f}_i - \mathbf{f}_j||^2/\sigma^2} & \text{if } i \text{ and } j \text{ are neighbors} \\ \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where f_i and f_j are the feature vectors extracted at position i and j respectively to describe the texture and spatial relationship between positions i and j. It should be noted the definition of weight similarity matrix $\mathbf{S}$ is not limited to above two versions. Different definitions are possible according to different problems.

3.2. Relationship of CCLDA, LDA and 2DLDA

It is obvious that when $\eta = 0$, CCDA will degenerated into classical LDA. In (Zheng et al. (2008)), it is revealed that the formulations of LDA and 2DLDA have the following relationship.

Theorem 1. *Let* $\boldsymbol{w} = [\tilde{\boldsymbol{w}}_1^T, \cdots, \tilde{\boldsymbol{w}}_c^T]^T$. *If* $\tilde{\boldsymbol{w}}_1, \cdots, \tilde{\boldsymbol{w}}_c$ *are imposed to be equivalent, i.e.,*

$$\boldsymbol{w}^{2d} = \tilde{\boldsymbol{w}}_1 = \cdots = \tilde{\boldsymbol{w}}_c \in \boldsymbol{R}^{r\times 1}$$

then the following relations are valid:

$$\begin{aligned}
\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} &= \boldsymbol{w}^{2dT} \boldsymbol{S}_b^{2d} \boldsymbol{w}^{2d} + \boldsymbol{w}^{2dT} \{\sum_{k=1}^{L} \frac{N_k}{N} \sum_{j=1,h=1,j\neq h}^{c} (\boldsymbol{U}_k(j) \\
&- \boldsymbol{U}(j))(\boldsymbol{U}_k(h) - \boldsymbol{U}(h))^T\} \boldsymbol{w}^{2d} \\
\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} &= \boldsymbol{w}^{2dT} \boldsymbol{S}_w^{2d} \boldsymbol{w}^{2d} + \boldsymbol{w}^{2dT} \{\frac{1}{N} \sum_{k=1}^{L} \sum_{i=1}^{N_k} \sum_{i=1,h=1,j\neq h}^{c} (\boldsymbol{X}_i^k(j) \\
&- \boldsymbol{U}_k(j))(\boldsymbol{X}_i^k(h) - \boldsymbol{U}_k(h))^T\} \boldsymbol{w}^{2d}.
\end{aligned}$$

Therefore, if the columns of image are independent and the projection weights for each column are the same, the result of 2DLDA is equivalent with that of LDA. However, as pointed in (Zheng et al. (2008)), Theorem 1 indicates that 2DLDA drops the information that characterizes the covariance of columns of image, which may be helpful for classification. Therefore, the performance of 2DLDA is not always better than LDA.

In CCLDA, we impose a constraint on projection to generate the weights in neighboring regions such that they are close together. Particulary, if the similarity matrix $\mathbf{S}$ is constructed in the way that the similarities of pixels from the same row are set to be 1 and others are 0, we have the following theorem.

$$S_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are in the same row} \\ \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

Corollary 1. *When η in Eq. 10 is large enough and the $\boldsymbol{S}$ is defined as Eq. 14, the CCLDA will approach to 2DLDA if the following equations hold:*

$$\sum_{k=1}^{L} \frac{N_k}{N} \sum_{j=1,h=1,j\neq h}^{c} (\boldsymbol{U}_k(j) - \boldsymbol{U}(j))(\boldsymbol{U}_k(h) - \boldsymbol{U}(h))^T\} = 0$$

$$\frac{1}{N} \sum_{k=1}^{L} \sum_{i=1}^{N_k} \sum_{i=1,h=1,j\neq h}^{c} (\boldsymbol{X}_i^k(j) - \boldsymbol{U}_k(j))(\boldsymbol{X}_i^k(h) - \boldsymbol{U}_k(h))^T = 0$$

The proof of this corollary is straightforward. The definition in Eq. 14 assures that the LDA weights of pixels from the same row are as close as possible. When η is large, more concern will be paid on the contextual constraint and hence the derived solution will approach to the constraint of 2DLDA if the columns of images are assumed to be independent.

Compared with 2DLDA, CCLDA utilizes the contextual information in a more reasonable and flexible way. Generally, in the formulation of CCLDA, it models the contextual information as well as preserves the covariance information among different columns, so it incorporates the advantage of LDA and 2DLDA and should achieve better performance than LDA and 2DLDA in practice.

Another interesting point is the relationship between CCLDA and regularized LDA (R-LDA) (Friedman (1989)). The formulations of these two methods are very similar. The CCLDA utilizes the contextual information by introducing a laplacian matrix while the R-LDA regularizes the LDA solution space by adding the identity matrix onto the within class scatter matrix. The characteristic of CCLDA is that it explicitly models contextual information in its formulation and hence owns more semantic meanings than R-LDA.

4. Experiments

Four face databases, ORL, Extended Yale-B, PIE, and FRGC are used to evaluate the performance of CCLDA with Fisher LDA (FLDA) (Belhumeur et al. (1997)), Null LDA (N-LDA) (Chen et al. (2000)), Direct LDA (D-LDA) (Yu and Yang (2001)), R-LDA (Friedman (1989)) and Lu's R-LDA (R-LDA^L) (Lu et al. (2004)) etc. methods. For CCLDA, we use the soft constraints defined in Eq. 13 and extract the gray values of images at position i to form the feature vector f_i. The 2DLDA is implemented following the paper (Ye et al. (2004)).

4.1. Data Preparation

4.1.1. ORL

There are 40 persons in ORL database (Samaria and Harter (1994)), each of which contains 10 images. The images are all frontal with slight tilt of the head, including lighting and expression variations. All the images are cropped to 32×32 size.

4.1.2. Extended Yale-B

The Extended Yale-B database (Georghiades et al. (2001)) contains 2414 images from 38 persons each of which contains nearly 64 ones. All the images are rotated, scaled and cropped to 32×32 size.

4.1.3. PIE

The PIE database (Sim et al. (2003)) consists of $41,368$ images from 68 people under different poses, illumination conditions and expressions. Five near frontal poses(C05, C07, C09, C27, C29) and all the images under different

illuminations and expressions are selected. So there are totally 170 images for each individual. All the images are cropped to 32×32 in experiments.

4.1.4. FRGC

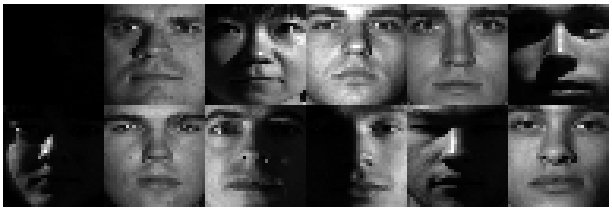
For FRGC database (Phillips et al. (2005)), we follow the experiment 1&4 protocols to evaluate different algorithms. There are $12,776$ face images from 222 individuals in training set, including 6360 controlled images and 6416 uncontrolled ones. In experiment 1, both target and query images contains 16028 controlled images and in experiment 4, which is considered as the most challenge case, there are 16028 controlled images as the target images and 8014 query images which are uncontrolled ones, from 466 persons. The images are captured over several sessions. All the faces are cropped to 64×64 size images.

4.2. Parameter Selection

There are mainly two parameters in our proposed method that affecting the performance of algorithm. One is the value of σ in Eq. 13, another is the contextual information regularized coefficient η in Eq. 10. In this experiment, the feature vector f in Eq. 13 is extracted by grouping the the pixel values at corresponding position from all the training images. The parameter σ is empirically set to be the average distance among these feature vectors. For the parameter η, we try to examine the impact of different values of η on the performance of algorithm and then choose the best one.

Fig. 4 shows the recognition rate of CCLDA on the PIE face database as a function of η. It is easy to see that the CCLDA can achieve better recognition performance than LDA when the value of η is approximately

(a) ORL



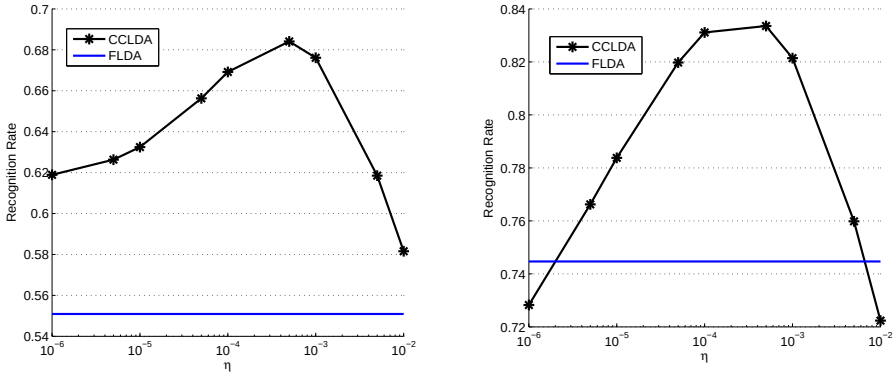
(b) Extended Yale-B

(c) PIE

(d) FRGC

Figure 3: Sample examples from (a) ORL, (b) Extended Yale-B, (c) PIE and (d) FRGC databases.

between 0.00005 and 0.005, and it achieves its highest recognition rate at the point of $\eta = 0.0005$. Therefore, in all the following experiments, the



(a) PIE 5 train (b) PIE 10 train

Figure 4: Parameter η selection for CCLDA. The curve shows the recognition rate with respect to η. The solid line shows the recognition rate of FLDA.

parameter η is set to 0.0005. For R-LDA and R-LDA^L, the regularized coefficient is set to 0.01 empirically in the same way.

4.3. Recognition Results

For the experiments on ORL, Extended Yale-B and PIE databases, the database is randomly partitioned into gallery and probe sets. For ease of representation, in following results, G_N denotes that N images per person are selected to form gallery set and the left ones consist the probe set. The gallery set is used for training, and in test phase, the images in probe set are compared with the images in gallery set and the nearest neighboring (NN) classifier is used to do the classification. All experimental results are reported as the average of 10 different runs on different splits. For FRGC, we follow the experiment 1 and 4 protocols and report the verification rate when the false accept rate is 0.001 in ROC I, ROC II and ROC III.

For FLDA, the PCA is first applied by keeping 98% energy, followed by LDA. The reduced dimension of FLDA, N-LDA, D-LDA, R-LDA and CCLDA are chosen to be $C-1$ to retain the most discriminative information, where C is the class number. For 2DLDA, we finally keep 10×10 dimension for the classification.

Table 1: ORL: Classification accuracy (mean±std%) with different training set size.

Method	G2	G3	G4	G5	G6
FLDA	61.09 ± 4.88	77.93 ± 6.13	92.00 ± 1.88	95.20 ± 1.55	96.56 ± 1.82
N-LDA	82.94 ± 2.59	91.00 ± 2.33	94.92 ± 1.16	96.90 ± 1.31	97.13 ± 2.40
D-LDA	34.63 ± 13.01	83.93 ± 2.65	91.83 ± 2.18	95.45 ± 1.30	96.38 ± 1.90
2DLDA	78.28 ± 3.50	87.57 ± 2.21	93.46 ± 2.21	95.15 ± 1.73	96.19 ± 1.90
R-LDA	82.88 ± 2.74	91.14 ± 2.23	95.50 ± 1.49	97.00 ± 1.03	$\mathbf{97.88} \pm 1.89$
R-LDA^L	76.06 ± 4.78	84.18 ± 2.76	91.88 ± 2.18	95.45 ± 1.30	96.44 ± 1.93
CCLDA	$\mathbf{85.94} \pm 2.59$	$\mathbf{92.25} \pm 2.08$	$\mathbf{96.08} \pm 1.31$	$\mathbf{97.30} \pm 1.03$	97.81 ± 1.62

4.4. Analysis and Discussion

As mentioned in Section 3.1, the essence of CCLDA is to impose the smooth constraint on the weight image $\mathbf{w}$ so that the weight values in local region do not fluctuate too much. This is consistent with the property of natural images. Fig. 5 shows the first 5 projections of FLDA and CCLDA in the form of images. It shows that compared to the result of FLDA, the projections of CCLDA are much smoother; that is, the values of the elements

Table 2: Extended Yale-B: Classification accuracy (mean±std%) with different training set size.

Method	G10	G20	G30	G40	G50
FLDA	79.08 ± 1.38	88.07 ± 0.58	92.17 ± 0.64	93.70 ± 0.59	94.34 ± 1.16
N-LDA	79.03 ± 2.31	83.23 ± 1.39	N/A	N/A	N/A
D-LDA	77.30 ± 1.16	84.92 ± 0.61	88.97 ± 0.77	89.99 ± 0.88	91.11 ± 1.28
2DLDA	81.45 ± 1.66	88.47 ± 0.89	92.27 ± 0.57	93.20 ± 0.57	94.26 ± 1.31
R-LDA	82.61 ± 1.53	$\mathbf{91.29} \pm 0.64$	94.60 ± 0.53	$\mathbf{95.92} \pm 0.62$	$\mathbf{96.48} \pm 0.86$
R-LDA^L	64.06 ± 2.43	77.30 ± 1.17	84.92 ± 0.61	88.97 ± 0.77	89.97 ± 0.87
CCLDA	$\mathbf{82.66} \pm 1.47$	91.23 ± 0.77	$\mathbf{94.62} \pm 0.63$	95.87 ± 0.53	96.28 ± 1.04

Table 3: PIE: Classification accuracy (mean±std%) with different training set size.

Method	G10	G20	G30	G40	G50
FLDA	74.47 ± 0.83	87.72 ± 0.90	91.85 ± 0.64	93.85 ± 0.41	95.36 ± 0.36
N-LDA	69.95 ± 1.05	N/A	N/A	N/A	N/A
D-LDA	73.96 ± 0.94	85.11 ± 0.89	89.68 ± 0.56	92.07 ± 0.46	93.77 ± 0.37
2DLDA	74.44 ± 0.91	85.08 ± 0.59	89.32 ± 0.44	91.69 ± 0.44	93.19 ± 0.39
R-LDA	83.17 ± 0.97	91.89 ± 0.67	94.35 ± 0.27	95.56 ± 0.27	96.54 ± 0.32
R-LDA^L	58.95 ± 1.49	73.94 ± 0.94	85.10 ± 0.89	89.68 ± 0.56	92.07 ± 0.46
CCLDA	$\mathbf{83.35} \pm 0.89$	$\mathbf{92.04} \pm 0.69$	$\mathbf{94.50} \pm 0.31$	$\mathbf{95.71} \pm 0.31$	$\mathbf{96.70} \pm 0.29$

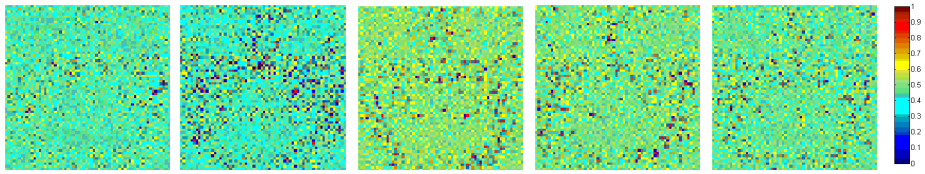
Table 4: FRGC: Verification rates when false accept rate is at 0.001.

Method	Experiment 1			Experiment 4		
	ROC I	ROC II	ROC III	ROC I	ROC II	ROC III
FLDA	0.7932	0.7746	0.7531	0.4514	0.4464	0.4393
D-LDA	0.8714	0.8438	0.8378	0.4240	0.4092	0.3879
2DLDA	0.6384	0.5745	0.5059	0.1735	0.1612	0.1481
R-LDA	0.8715	0.8520	0.8378	0.4844	0.4738	0.4633
R-LDA^L	0.8720	0.8439	0.8173	0.4241	0.4085	0.3868
CCLDA	**0.9222**	**0.9111**	**0.9003**	**0.6022**	**0.6005**	**0.6005**

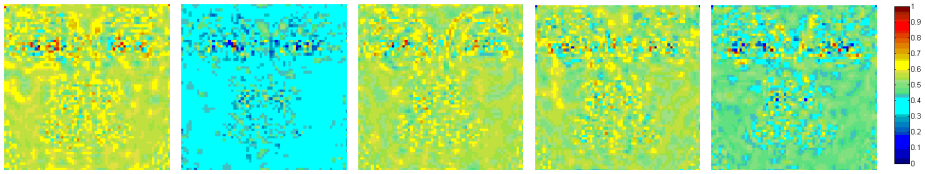
in neighboring areas are much closer, thus satisfying the motivation behind this work.

Tables 1-4 illustrate the recognition performance of CCLDA and other popular methods on ORL, Extended Yale-B, PIE and FRGC databases. The results reveal some interesting points.

1. In almost all experiments, the performance of CCLDA is always better than that of LDA, N-LDA, D-LDA, 2DLDA and R-LDA^L. It proves that the CCLDA method is effective for solving image related pattern recognition problem such as face recognition.
2. Compared with FLDA, which is based on vector representation, 2DLDA and CCLDA, which consider the image contextual constraints, significantly improve the recognition performance when the sample size is small. It indicates that the contextual information is important and can be considered as a good prior for image recognition. Moreover, the



(a)



(b)

Figure 5: The first five weight images of FLDA (a) and CCLDA (b) on FRGC training set. The weight images of CCLDA are much smoother than those of FLDA.

performance of CCLDA is always better than 2DLDA, which proves the CCLDA could better utilize the contextual information than 2DLDA does.

3. Compared CCLDA to R-LDA, the performance of these two methods are very close. Overall, the performance of CCLDA is slightly better than that of R-LDA, but not significant. The advantage of CCLDA is that it explicitly models the contextual information in its formulation and it carries more definite semantic meanings than R-LDA.

4. On FRGC, where the experiment is considered as the most challenge one, CCLDA significantly enhances the recognition performance (much better than R-LDA in this case). It is surprising that the performance of 2DLDA is much worse than that of FLDA in this case. One possible explanation is that the training set size of FRGC is large enough for

FLDA. Compared with 2DLDA, CCLDA takes into account the contextual information in a more reasonable and flexible way and achieves better results.

5. Conclusions

In this paper, we propose contextual constraints based linear discriminant method and apply it to image recognition problem. The motivation is that the contextual information is important for image understanding. Therefore, it is possible that contextual information could provide useful clues that help improve image classification performance. Extensive experiments validate our motivation. The proposed contextual constraints could be incorporated into not only linear discriminant analysis, but also other subspace learning methods like LPP, NPE etc. One of our future work is to extend the contextual information to the non-linear (kernel) formulation so as to improve the classification performance further.

References

Belhumeur, P., Hespanha, J., Kriegman, D., 1997. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. PAMI 19 (7), 711–720.

Chen, L., Liao, H., Ko, M., Lin, J., Yu, G., October 2000. A new lda-based face recognition system which can solve the small sample size problem. Pattern Recognition 33 (10), 1713–1726.

Dass, S. C., Jain, A. K., July 9-12 2001. Markov face models. In: ICCV. Vancouver, pp. 112–116.

Dass, S. C., Jain, A. K., Lu, X., August 2002. Face detection and synthesis using Markov random field models. In: ICPR. Quebec City, pp. 112–116.

Friedman, J. H., March 1989. Regularized discriminant analysis. Journal of the American Statistical Association 84 (405), 165.

Georghiades, A., Belhumeur, P., Kriegman, D., 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. PAMI 23 (6), 643–660.

Gui, J., Jia, W., Zhu, L., Wang, S., Huang, D., 2010. Locality preserving discriminant projections for face and palmprint recognition. Neurocomputing 73 (13-15), 2696–2707.

He, X., Cai, D., Yan, S., Zhang, H., 2005a. Neighborhood preserving embedding. In: ICCV. pp. 1208–1213.

He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H., 2005b. Face recognition using laplacianfaces. IEEE Trans. PAMI 27 (3), 328–340.

Huang, R., Pavlovic, V., Metaxas, D., August 2004. A hybrid face recognition method using markov random fields. In: ICPR. Cambridge, UK, pp. 157–160.

Kong, H., Wang, L., Teoh, E. K., Wang, J.-G., Venkateswarlu, R., 2005. A framework of 2d fisher discrminant analysis: Application to face recognition with small number of training samples. In: CVPR. Vol. 2. pp. 1083–1088.

Li, M., Yuan, B., 2005. 2d-lda: A novel statistical linear discriminant analysis for image matrix. Pattern Recognition Letters 26 (5), 527–532.

Lu, J., Plataniotis, K. N., Venetsanopoulos, A. N., 2004. Regularization studies on lda for face recognition. In: Proceedings of IEEE International Conference on Image Processing. pp. 63–66.

Phillips, P. J., Flynn, P. J., Scruggs, W. T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W. J., 2005. Overview of the face recognition grand challenge. In: CVPR. pp. 947–954.

Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (22), 2323–2326.

Samaria, F., Harter, A., 1994. Parameterisation of a stochastic model for human face identification. In: Proceedings of 2nd IEEE Workshop on Applications of Computer Vision. Sarasota FL, pp. 138–142.

Sim, T., Baker, S., Bsat, M., 2003. The CMU pose, illumination, and expression database. IEEE Trans. PAMI 25 (12), 1615–1618.

Turk, M. A., Pentland, A. P., 1991. Face recognition using eigenfaces. In: CVPR. pp. 586–591.

Wang, L., Zhang, Y., Feng, J., 2005. On the euclidean distance of images. IEEE Trans. PAMI 27 (8), 1334–1339.

Xiong, H., Swamy, M. N. S., Ahmad, M. O., July 2005. Two-dimensional fld for face recognition. Pattern Recognition 38, 1121–1124.

Yang, J., Zhang, D., Yong, X., Yang, J.-Y., 2005. Two-dimensional discriminant transform for face recognition. Pattern Recognition 38, 1125–1129.

Ye, J., Janardan, R., Li, Q., 2004. Two-dimensional linear discriminant analysis. In: NIPS. Vol. 17. pp. 1569–1576.

Yu, H., Yang, J., October 2001. A direct lda algorithm for high-dimensional data with application to face recognition. Pattern Recognition 34 (10), 2067–2070.

Zheng, W.-S., Lai, J. H., Li, S. Z., 2008. 1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based? Pattern Recognition 47, 2156–2172.