Discriminant Analysis for Perceptionally Comparable Classes

Bingpeng Ma[1,2], Shiguang Shan[1], Xilin Chen[1], Wen Gao[1,3]
[1] Key Lab of Intelligent Information Processing of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, 100190, China
[2] Graduate School of the Chinese Academy of Sciences, CAS, Beijing, 100039, China
[3] School of Electronic Engineering and Computer Science, Peking University, Beijing, 100871, China
{bpma, sgshan, xlchen, wgao}@jdl.ac.cn

Abstract

Traditional discriminate analysis treats all the involved classes equally in the computation of the between-class scatter matrix. However, we find that for many vision tasks, the classes to be processed are not equal in perception, i.e. a distance metric can be defined between the classes. Typical examples include head pose classification and age estimation. Aiming at this category of classification problem, this paper proposes a novel discriminant analysis method, called Class Distance based Discriminant Analysis (CDDA). In CDDA, the perceptional distance between two classes is exploited to weight the outer product in the between-class scatter computation, to concentrate more on the classes difficult to separate. Another novelty of CDDA is that to preserve the within-class local structure of multimodal labeled data, the within-class scatter is re-defined by complementing the similarity of the samples pairs in the nearby classes. The method is then applied to head pose classification and age estimation problem, and experimental results demonstrate the effectiveness of CDDA.

1. Introduction

The goal of dimensionality reduction is to embed high-dimensional data samples in a low-dimensional space so that most of "intrinsic information" contained in the data is preserved [7] [10] [2]. Once dimensionality reduction is carried out appropriately, the compact representation of the data can be used for various succeeding tasks such as visualization, classification, etc. In this paper, we consider the supervised dimensionality reduction problem.

Fisher Discriminant Analysis (FDA) [4] is a popular method for linear supervised dimensionality reduction and seeks for an embedding transformation to maximize between-class scatter and minimize within-class scatter. Besides the well-known limitation that the dimension of the features is at most $c-1$, where c is the number of the classes, FDA has two other disadvantages. One of the disadvantages is that it is difficult to deal with the within-class multimodality. The multimodality means that the samples belonging to one class form several separate clusters and can be observed in many practical applications. By combining the ideas of FDA and Locality Preserving Projections (LPP) [5], M. Sugiyama proposes a new method called Local Fisher Discriminant Analysis (LFDA) [9], which not only maximizes between-class separability but also simultaneously preserves the within-class local structure. The paper reports that LFDA is useful for dimensionality reduction of multimodal labeled data.

The other disadvantage of FDA is that all the classes are equally considered in the computation of the between-class scatter matrix $\mathbf{S}^b$, if only they have equal number of training samples. In FDA, $\mathbf{S}^b$ can be defined as follows [12]:

$$\mathbf{S}^b = \frac{1}{2n}\sum_{i,j=1}^{c} n_i n_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \qquad (1)$$

where c is the total number of classes; n is the total number of training samples; n_i and n_j are the number of training samples for class i and class j respectively; $\mathbf{m}_i$ and $\mathbf{m}_j$ are the mean of the class i and class j respectively.

From Equation (1), it is easy to see the phenomena that all the classes are equally considered. In theory, such a strategy may cause problems in case the class means are not uniformly distributed, for instance, with one class very far away from all the other classes. In this situation, the Fisher criterion may be attracted too much by the "outlier" class, since Fisher criterion is Euclidean distance-based rather than classifiability-based. A synthetic example is shown in Figure 1, in which two classes are nearby while the third class is far from the other two classes. It is clear that the first projection pursued by FDA method is far from the optimal discriminatory projection to separate all the three classes.

For the above problem, M. Loog et al. propose a mod-

978-1-4244-2154-1/08/$25.00 2008 IE

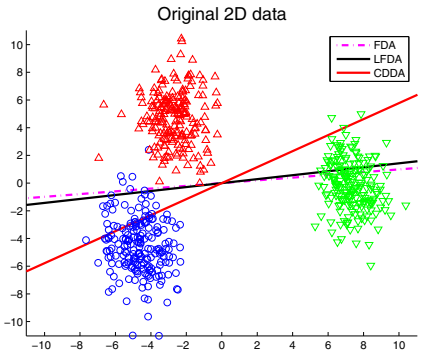

Figure 1. The projection direction of FDA, LFDA and CDDA.

ification of the Fisher criterion, in which the Mahanalobis distances between the class means are used as the weight of the contribution to $\mathbf{S}^b$ depends on the Bayes error rate [6]. In practice, there is also a category of classification applications in which the classes are not equal in perception. In other words, the classes to be processed are comparable; even some distance measurement between class labels can be defined. Typical examples include head pose estimation and age estimation. For head pose classification, when only the head rotation in yaw are considered as the classes, the labels of the classes are ranged in $[-90°, 90°]$, which are continuous and comparable labels. Easy to understand, the distance between class of $20°$ and $25°$ is evidently smaller than that between $20°$ and $50°$. Similarly, in age estimation, ages can also be seen as the comparable labels.

To overcome these drawbacks of FDA, in this paper, a new discriminant analysis method named Class Distance based Discriminant Analysis (CDDA) is proposed. Similarly to LFDA, there is no limitation that the dimension of the features is at most $c-1$ for CDDA. Therefore, CDDA can be practically employed for dimensionality reduction into any dimensional spaces. The main contribution of CDDA is, for the sample pairs from the different class labels, the weights in the re-defined between-class scatter matrix are adjusted according to their class distances, which can improve the discriminant ability of features. In other words, the weights of the sample pairs from the nearby classes are increased to improve the discriminant ability of the nearby classes. The adjustment is based on the facts that the samples between the nearby classes are more similar and the misclassified samples are often misclassified into the nearby classes. Compared with the methods using the distances of the class means proposed by M. Loog, the performance of CDDA is more better for that the class distances introduce more useful information.

The other contribution of CDDA is that in the re-defined within-class scatter matrix $\hat{\mathbf{S}}^w$, the weights of all the sample pairs, unlike only the pairs in the same class in FDA and LFDA, are modified by the new similarity of samples to preserve the within-class local structure of the multimodal samples. In CDDA, the new similarity is computed based on the combination of the image appearance distances (the Euclidean distance) and the class distances. Then, for the sample pairs in the same class, the larger the Euclidean distance, the smaller the weight; for the sample pairs from the different classes, the larger the class distance, the smaller the weight of the sample pairs. By this way, the comparability of the nearby classes is complemented in $\hat{\mathbf{S}}^w$. Compared with LFDA, for the samples far apart the class mean, the combination further reduces their influence on $\hat{\mathbf{S}}^w$. In this way, the within-class local structure is kept well.

To validate the effectiveness, CDDA is applied to the above-mentioned toy samples and two real problems: head pose estimation and age estimation. In Figure 1, the projected direction of CDDA is the same with the optimal projected direction. The results of the experiments illustrate that CDDA can improve the discriminant ability.

2. Background

Since CDDA is mainly motivated by FDA and LFDA, this section briefly introduce these two methods.

2.1. Fisher Discriminant Analysis

FDA is a popular method for linear supervised dimensionality reduction. Through a linear transformation, the original feature representation is projected into a new subspace where the between-class scatter matrix $\mathbf{S}^b$ is maximized while the within-class scatter matrix $\mathbf{S}^w$ is minimized by maximizing the Fisher separation criterion. FDA is interpreted as keeping the sample pairs of the same class close and the sample pairs of different classes apart. Let $\mathbf{x}_i \in \mathbb{R}^d (i = 1, 2, \cdots, n)$ be the sample in the d-dimensional space, $y_i \in \{1, 2, \cdots, c\}$ the label of $\mathbf{x}_i$, n the total number of samples, c the total number of the classes, n_p the number of the samples in class p. $\mathbf{S}^w$ and $\mathbf{S}^b$ can be defined in pairwise manner [9]:

$$\mathbf{S}^w = \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{i,j}^w (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{2}$$

$$\mathbf{S}^b = \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{i,j}^b (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{3}$$

where

$$\mathbf{W}_{i,j}^w = \begin{cases} 1/n_p & if \quad y_i = y_j = p \\ 0 & if \quad y_i \neq y_j \end{cases} \tag{4}$$

$$\mathbf{W}_{i,j}^b = \begin{cases} 1/n - 1/n_p & if \quad y_i = y_j = p \\ 1/n & if \quad y_i \neq y_j \end{cases} \tag{5}$$

Finally, the optimal projection matrix $\mathbf{T}_{\text{FDA}}$ can be obtained by solving the following optimization problem:

$$\mathbf{T}_{\text{FDA}} = \underset{\mathbf{T}\in\mathbb{R}^{d\times r}}{\text{argmax}}[\text{tr}\left(\mathbf{T}^T\mathbf{S}^b\mathbf{T}(\mathbf{T}^T\mathbf{S}^w\mathbf{T})^{-1}\right)] \tag{6}$$

2.2. Local Fisher Discriminant Analysis

By effectively combining the ideas of FDA and LPP, LFDA aims at reducing the dimensionality of multimodal labeled data appropriately by maximizing between-class separability and preserving the within-class local structure at the same time. In LFDA, $\tilde{\mathbf{S}}^w$ and $\tilde{\mathbf{S}}^b$ are re-defined by replacing $\mathbf{W}^w_{i,j}$ and $\mathbf{W}^b_{i,j}$ by $\tilde{\mathbf{W}}^w_{i,j}$ and $\tilde{\mathbf{W}}^b_{i,j}$:

$$\tilde{\mathbf{W}}^w_{i,j} = \begin{cases} \tilde{A}_{i,j}/n_p & if \quad y_i = y_j = p \\ 0 & if \quad y_i \neq y_j \end{cases} \tag{7}$$

$$\tilde{\mathbf{W}}^b_{i,j} = \begin{cases} \tilde{A}_{i,j}(1/n - 1/n_p) & if \quad y_i = y_j = p \\ 1/n & if \quad y_i \neq y_j \end{cases} \tag{8}$$

where $\tilde{A}_{i,j}$ is the weight of the sample pairs belonging to the same class and employed by the local scaling method [13]:

$$\tilde{A}_{i,j} = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\tilde{\sigma}_i\tilde{\sigma}_j}) \tag{9}$$

where $\tilde{\sigma}_i$ represents the local scaling of the samples around $\mathbf{x}_i$ and is determined by

$$\tilde{\sigma}_i = ||\mathbf{x}_i - \mathbf{x}_i^{(k)}|| \tag{10}$$

where $\mathbf{x}_i^{(k)}$ is the k-th nearest neighbor of $\mathbf{x}_i$. In LFDA, far apart sample pairs in the same class have less influence on $\tilde{\mathbf{S}}^b$ and $\tilde{\mathbf{S}}^w$. More details of LFDA are introduced in [9].

3. Class Distance based Discriminant Analysis

In this section, we present the proposed CDDA method in detail, especially how the pre-defined class distance can be involved in the Fisher separation criterion. We first present the general form of CDDA, and then discuss how the parameters should be set in CDDA. The computational efficiency problem is also discussed in the last sub-section.

3.1. Class Distance based Discriminant Analysis

As introduced in Section 1, though FDA has achieved great successes in many areas, it has some drawbacks. Firstly, the dimension of the features is at most $c-1$. Secondly, all the classes are equally considered and there is no an effective way to use the comparability of the classes. Finally, it is difficult to preserve the within-class local structure of the multimodal labeled data because $\mathbf{S}^b$ and $\mathbf{S}^w$ are evaluated globally.

In this paper, CDDA is proposed to overcome the above drawbacks of FDA. In CDDA, similarly to the pairwise manner of FDA, the within-class scatter matrix $\hat{\mathbf{S}}^w$ and the between-class scatter matrix $\hat{\mathbf{S}}^b$ are re-defined by replacing $\mathbf{W}^w_{i,j}$ and $\mathbf{W}^b_{i,j}$ by $\hat{\mathbf{W}}^w_{i,j}$ and $\hat{\mathbf{W}}^b_{i,j}$:

$$\hat{\mathbf{W}}^w_{i,j} = \hat{A}_{i,j}/\sqrt{n_p n_q} \quad y_i = p, y_j = q \tag{11}$$

$$\hat{\mathbf{W}}^b_{i,j} = \begin{cases} \hat{A}_{i,j}(1/n - 1/n_p) & if \quad y_i = y_j = p \\ \hat{B}_{p,q}/n & if \quad y_i = p, y_j = q, p \neq q \end{cases} \tag{12}$$

where $\hat{A}_{i,j}$ is the similarity weight and $\hat{B}_{p,q}$ is the dissimilarity weight of the sample pair $\mathbf{x}_i$ and $\mathbf{x}_j$. Finally, the optimal projection matrix $\mathbf{T}_{\text{CDDA}}$ can be obtained by solving the following equation:

$$\mathbf{T}_{\text{CDDA}} = \underset{\mathbf{T}\in\mathbb{R}^{d\times r}}{\text{argmax}}[\text{tr}\left(\mathbf{T}^T\hat{\mathbf{S}}^b\mathbf{T}(\mathbf{T}^T\hat{\mathbf{S}}^w\mathbf{T})^{-1}\right)] \tag{13}$$

Clearly, if $\hat{A}_{i,j}$ is set to 1 for the sample pairs in the same class and 0 for the pairs from the different classes, and $\hat{B}_{p,q}$ is set to 1 for all the sample pairs, $\hat{\mathbf{S}}^w$ and $\hat{\mathbf{S}}^b$ become the same as $\mathbf{S}^w$ and $\mathbf{S}^b$ in traditional FDA respectively, i.e., CDDA is reduced to the original FDA. Therefore, CDDA might be regarded as a natural localized variant of FDA.

3.2. Dissimilarity Weight $\hat{B}_{p,q}$

In FDA, there are only two relations for two class labels: same or different, i.e., the degree of variance between the classes is not considered. For example, in head pose estimation, the difference of angle 0° from 1° and the difference of angle 0° from 90° are equally treated as the between-class differences and have the same influence on $\mathbf{S}^b$.

In this paper, we argue that the difference degrees of the classes are very important for classification. Generally, the misclassified samples are often misclassified into the nearby classes but hardly the faraway classes. In other words, the smaller the class distance, the more similarly the sample pair and the more possible the samples be misclassified. Therefore, in CDDA, $\hat{B}_{p,q}$ is set to increase with the decrease of the class distances. Such a strategy implies that sample pairs with the smaller class distances have more influence on the CDDA Fisher criterion. Then, by maximizing the criterion, the pursued projection directions can more accurately distinguish the samples from the nearby classes.

In practice, $\hat{B}_{p,q}$ can be designed to meet all kinds of special cases. In this paper, we adopt a unified definition:

$$\hat{B}_{p,q} = \exp(-l^\alpha_{p,q})/\bar{B} \tag{14}$$

where α is a constant; $\bar{B}$ is the mean of $\hat{B}$; $l_{p,q}$ is the class distance between class p and class q. In head pose estimation, $l_{p,q}$ can be defined as $l_{p,q} = |y_i - y_j|$, where y_i and y_j are the pose angle of $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively.

3.3. Similarity Weight $\hat{A}_{i,j}$

In Equation (11) and (12), $\hat{A}_{i,j}$ is the similarity weight of $\mathbf{x}_i$ and $\mathbf{x}_j$, which is used to preserve the within-class local structure. On one hand, since Euclidean distances are not directly related to the separability of samples, it is not enough that only Euclidean distances are taken as the metric of sample similarity. On the other hand, because of the class comparability, it is more reasonable that the similarity metric should be designed more related to the class distances. In other words, the sample pairs belonging to the nearby classes should have the larger similarity. Thus in CDDA, the similarity metric $\hat{d}_{i,j}$ of samples is defined as the combination of Euclidean distances and the class distances:

$$\hat{d}_{i,j} = ||\mathbf{x}_i - \mathbf{x}_j||^2 \times L_{p,q} \tag{15}$$

where $L_{p,q}$ is the correctional form of the original metric $l_{p,q}$ to ensure that the biased distance values are well-separated for different classes and is defined as:

$$L_{p,q} = \begin{cases} l_{p,q}/(\mathrm{L} - l_{p,q}) & if \quad l_{p,q} \in [0, \mathrm{L}) \\ \mathrm{L} & otherwise \end{cases} \tag{16}$$

where L is a constant. Then the local scaling $\hat{\sigma}_i$ of the data samples around $\mathbf{x}_i$ is defined as:

$$\hat{\sigma}_i = \hat{d}^{1/2}_{i,i^{(k)}} = ||\mathbf{x}_i - \mathbf{x}_i^{(k)}|| \times L^{1/2}_{p,q^{(k)}} \tag{17}$$

where $\mathbf{x}_i^{(k)}$ is the k-th nearest neighbor of $\mathbf{x}_i$, and the parameter k is a tuning parameter. For the new metric $\hat{d}_{i,j}$, the nearby sample pairs in the same class are closer and the pairs from the different classes are separated from each other correspondingly, which makes the task of classification much easier. Finally, in CDDA, $\hat{A}_{i,j}$ is defined as:

$$\hat{A}_{i,j} = \exp(-\frac{\hat{d}_{i,j}}{\hat{\sigma}_i \hat{\sigma}_j}) \tag{18}$$

The above equations tell us that $\hat{A}_{i,j}$ is in the range $(0, 1]$. The larger the class distance $l_{p,q}$, the larger $L_{p,q}$ and $\hat{d}_{i,j}$, and vice versa. At the same time, the larger $l_{p,q}$ is, the closer $\hat{A}_{i,j}$ is to 0; and the smaller $l_{p,q}$, the closer $\hat{A}_{i,j}$ is to 1.

Equation (3) shows that in FDA, the sample pairs belonging to the same class have the same influence on $\mathbf{S}^w$ and $\mathbf{S}^b$, which cause the destruction of the local structure to some extent. In LFDA, the local structure can be kept by the idea that the influences of the sample pairs in the same class are decreased with the increases of their Euclidean distances. However, Equation (9) shows that the local scaling $\tilde{\sigma}$ also has the influence on $\tilde{\mathbf{S}}^w$: when $||\mathbf{x}_i - \mathbf{x}_j||$ is the constant, the larger $\tilde{\sigma}_i$, the smaller $\tilde{A}_{i,j}$. When the neighborhoods are limited in the samples in the same class, the larger $\tilde{\sigma}$ increases the influence of the far apart sample pairs on $\tilde{\mathbf{S}}^w$ and $\tilde{\mathbf{S}}^b$, which disagrees the above-mentioned idea and limits the effect of $\tilde{A}_{i,j}$ in keeping the local structure.

In CDDA, the neighborhoods are extended to all the samples and the Euclidean distances of the sample pairs are combined by their class distances. In some sense, we actually divide samples into three types: the inner sample (whose class label is the same as that of all its neighbors), the outlier (all of its neighbors have different class labels from itself) and the boundary sample (part of whose neighbors come from the same class as itself, while others not). According the definition of $\hat{\sigma}$ in CDDA, for the inner samples, their $\hat{\sigma}$ are the same to the $\tilde{\sigma}$ in LFDA; for the boundary samples, their $\hat{\sigma}$ are much smaller than $\tilde{\sigma}$. In practice, to remove the influence of outliers, the $\hat{\sigma}$ of the outliers are just set to 0. Compared with LFDA, the influences of the far apart sample pairs are decreased further in CDDA. By this way, CDDA keeps the within-class local structure more accurately and improves the discriminant ability of features.

Equation (11) shows that the sample pairs from the different classes have also the contribution on $\hat{\mathbf{S}}^w$. We argue that this extension is significant when there are the comparable classes. On one side, since the samples of the nearby classes are more similarly, $\hat{\mathbf{S}}^w$ can be complement by the samples of the nearby classes when there are a few samples in the training database. In the extreme, when there is only 1 training sample for each class, $\mathbf{S}^w$ can not be computed in FDA but can be computed in CDDA for this complement. On the other side, since $\hat{A}_{i,j}$ is decreased greatly with the increase of the class distances, for the contribution on $\hat{\mathbf{S}}^w$, the proportion of the samples pairs from the different classes are much smaller than that of the pairs in the same class when there are enough samples in the training set. On the whole, in CDDA, the extension can be seen as the significative complement on $\hat{\mathbf{S}}^w$.

4. Experiment

In this section, we evaluate the performance of CDDA and other methods in head pose estimation and age estimation. From these experiments, we can draw the conclusion that CDDA can improve the discriminant ability by using the class comparability and preserving the within-class local structure.

The head angles in yaw and the ages are seen as the class labels in head pose estimation and age estimation respectively. Then, the within-class multimodality can be observed since the distance of the samples from the same person with nearby classes is smaller than that of the different person with the same class. Since the classes are near the continuity, the estimation problem can be taken as the classification problem for simpleness.

We compare CDDA with the following methods: PCA, FDA, aPAC, and LFDA. PCA and FDA have achieved the great successes in face recognition, head pose estimation

and other related areas [11] [1] [3] [8]. aPAC is the method proposed by M. Loog [6], in which S_b is redefined by introducing the Mahanalobis distance of the class means. To show the performance of using the sample pair from the different classes to complement $\hat{\mathbf{S}}^w$, we also design a new method named the same class CDDA (sCDDA), whose $\hat{B}_{i,j}$ is the same as the $\hat{B}_{i,j}$ of CDDA, but $\hat{A}_{i,j}$ is set to 0 if y_i is different with y_j.

For all the input images, firstly, face region is located by face detection method. Secondly, face regions are normalized to the same size of 32×32 and histogram equalization is used to reduce the influence of lighting variations. Then face regions are transformed to vectors. Thirdly, PCA is used to reduce the dimension of image vectors to reduce the computational complexity. Fourthly, different methods are applied to extract features. Finally, Nearest Neighbor (NN) classifier is used as the classifier to predict the labels of the testing samples. Since there is no specific gallery set, the training set is also taken as the gallery set.

For all the samples, 3-fold cross-validation is used to overcome over-training. The images in the database are sorted by their subject labels and divided into three parts. Two parts are taken as the training set and the other part is taken as the testing set. In this way, the subjects for training and testing are different. Repeat it three times, so each part has been taken as the testing set. The final results can be achieved by computing the mean of all the test sets.

The values of some parameters are set as follows: 95% of the total energy of eigenvalues is kept in PCA; the tuning parameter k is set to 7 because it works well by and large [13]; the constant α is set to 0.02; and L is set to $2c$.

4.1. Experiment on Head Pose Estimation

We first compare the performances of different methods on head pose estimation. We build a multi-view face database with close to the continuous poses, which is named the Multi-Pose database in this paper. The database consists of $3,030$ images taken from 102 subjects with consistent lighting conditions and background. The yaw angles and the pitch angles range within $[-50°, +50°]$ and $[-45°, +45°]$ with intervals of $1°$ respectively. In addition, some images contain faces wearing glasses. The image number for each subject is different. The images of one subject and the results of face detection are shown in Figure 2.

Since the angles in yaw are nearly continuous, we exploit the error between the ground-truth and the predicted angle. The error mean M is computed as follows:

$$M = \frac{1}{n}\sum_{i=1}^{n}|y_i' - y_i| \tag{19}$$

where n is the total number of the testing samples, y_i and y_i' are the ground-truth and the predicted angle of the ith sample respectively. Thus, the unit of this measurement is actually rotation degree in yaw. Figure 3 shows the plots of the error mean versus the dimension of subspace. The best error mean of all the dimensions is shown in Table 1.

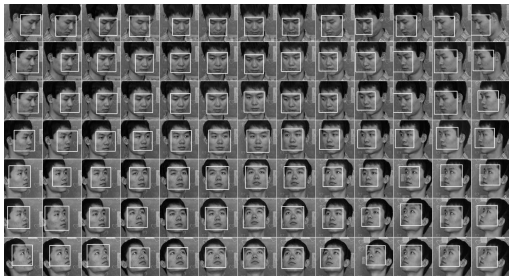

Figure 2. The results of face detection on the Multi-Pose database.

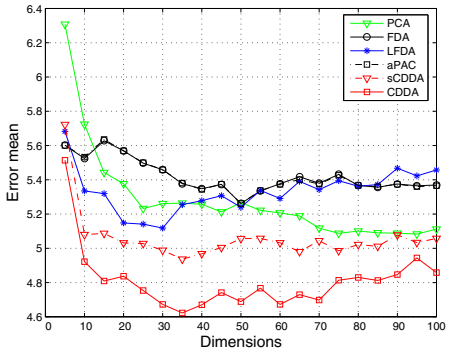

Figure 3. The error mean vs. dimension on head pose estimation.

We can get several points from Table 1 and Figure 3. Firstly, the performances of CDDA are always the best in all the methods for all the dimensions. In CDDA, the similarity metric of samples is the combination of the Euclidean distances and the class distances, which keeps the within-class local structure more accurately. The weight $\hat{B}_{i,j}$ is based on the class distances, which improves the discriminant ability to the sample pairs of the nearby classes. The combination of the similarity and dissimilarity weight improves the discriminant ability of features greatly.

Secondly, the performance of sCDDA is much better than that of other methods expect CDDA, which means that the discriminant ability can be improved by using the comparability of the labels; on the other hand, the performance of sCDDA is much worse than that of CDDA, which shows that the complement of the similarity of the sample pairs from the nearby classes in $\hat{\mathbf{S}}^w$ is effectively.

Thirdly, the performances of FDA and aPAC are near coincide with each other, which shows the modify of $\mathbf{S}^b$ by the distance of the class means is very limited. Besides pose angles, there is the identification in the input image. The distance of two images of the same person and two nearby

problem	measure	PCA	FDA	LFDA	aPAC	sCDDA	CDDA
pose estimation	degree($^\circ$)	5.08	5.26	5.12	5.26	4.94	4.62
age estimation	year	5.36	4.67	4.61	4.68	4.54	4.37

Table 1. The best error mean of different methods on head pose estimation and age estimation.

yaw angles may smaller than that of the same yaw angle and two different persons. In other words, there is the large within-class difference and the large overlap between the nearby classes in head pose estimation. In conditions, the distances between the class means are not availably improve the discriminant ability.

Finally, the performances of FDA are much worse than that of PCA when the dimensions are larger than 10, which disagrees with the common understand that the supervised FDA is much better than the unsupervised PCA. In head pose estimation, since the angles in pitch are different for the samples with the same angle in yaw, the similarity of the samples from the same class may be less than that from the different classes. Therefore, the poor performance of FDA can be attributed to the large within-class scatter.

4.2. Experiment on Age Estimation

To show the generalization of CDDA, we also compare the performances of different methods on age estimation on our own age database. In the database, the images are taken with consistent lighting conditions and background; some images contain faces wearing glasses; there is only one image for each subject; the ages range within $[25, 40]$ with intervals of 1. A subset including 1600 images (100 images for each age) is selected randomly to reduce the computational complexity. For all the methods, the dimension of features is set to 15. The error means of the different methods are shown in Table 1. The result analysis is near the same as the analysis on head pose estimation.

5. Conclusion

In this paper, a new discriminant analysis method named CDDA is proposed to improve the discriminant ability by using the comparability of the classes and preserve the within-class local structure. The method is especially effective for the applications in which a distance measure between the classes can be pre-defined as a prior. Effectiveness of CDDA is extensively validated by the performance comparisons on head pose estimation and age estimation. Experimental results show that CDDA outperforms PCA, FDA, aPAC and LFDA.

Acknowledgements

This paper is partially supported by National Natural Science Foundation of China under contract No.60332010, and No.60772071; Hi-Tech Research and Development Program of China under contract No.2006AA01Z122 and No.2007AA01Z163; 100 Talents Program of CAS; and ISVISION Technology Co. Ltd.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Laplacian eigenmaps and spectral techniques for embedding and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7):711–720, 1997.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 15:585–591, 2001.

[3] L. Chen, L. Zhang, Y. Hu, M. Li, and H. Zhang. Head pose estimation using fisher manifold learning. *Proc. IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 203–207, 2003.

[4] R. A. Fisher. The use of multiple measures in taxonomic problems. *Ann. Eugenics*, 7:179–188, 1936.

[5] X. He and P. Niyogi. Locality preserving projections. *Advances in Neural Information Processing Systems*, pages 153–160, 2003.

[6] M. Loog, R. P. W. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7):762–766, 2001.

[7] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[8] S. Srinivasan and K. L. Boyer. Head pose estimation using view based eigenspaces. *Proc. International Conference on Pattern Recognition*, 4:302–305, 2002.

[9] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine learning research*, 8:1027–1061, 2007.

[10] J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[11] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[12] X. Wang and X. Tang. Unified subspace analysis for face recognition. *Proceedings Ninth IEEE International Conference on Computer Vision*, 1:679–686, 2003.

[13] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems*, 17:1601–1608, 2004.