Confused Distance Maximization for Large Category Dimensionality Reduction

Xu-Yao Zhang Cheng-Lin Liu
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences, Beijing, China
xyz@nlpr.ia.ac.cn liucl@nlpr.ia.ac.cn

Abstract

The Fisher linear discriminant analysis (FDA) is the most well-known supervised dimensionality reduction model. However, when the number of classes is much larger than the reduced dimensionality, FDA suffers from the class separation problem in that it will preserve the distances of the already well-separated classes and cause a large overlap of neighboring classes. To cope with this problem, we propose a new model called confused distance maximization (CDM). The objective of CDM is to maximize the distance of the most confusable classes, according to the confusion matrix estimated from the training data with a pre-learned classifier. Compared with FDA that maximizes the sum of the distances of all class pairs, CDM is more relevant to the classification accuracy by weighting the pairwise distance according to the confusion matrix. Furthermore, CDM is computationally inexpensive which makes it indeed efficient and effective for large category problems. Experiments on two large-scale 3,755-class Chinese handwriting databases (offline and online) demonstrate that CDM can achieve the best performance compared with FDA and other competitive weighting based criteria.

1 Introduction

When solving a pattern classification problem, it is common to apply a feature extraction method as a preprocessing technique, not only to reduce the computation complexity, but also to obtain better generalization performance, by reducing irrelevant and redundant information in the data, and overcoming the estimation problems in statistical classifier learning.

The most well-known technique for linear dimensionality reduction is the Fisher linear discriminant analysis (FDA) [5], which learns a linear transformation matrix $W \in \mathbb{R}^{d\times d'}$ to transform the feature from $\mathbb{R}^d$ into a low-dimensional space $\mathbb{R}^{d'}$. The objective of FDA is to maximize the between-class distance as well as minimize the within-class distance. When the class-conditional distribution is Gaussian with equal covariance matrix for all the classes (homoscedastic), and the reduced dimensionality is $K-1$ (K is the number of classes), FDA is the optimal model for linear dimensionality reduction.

However, for the large category problems where $K >> d > d'$, FDA is only a suboptimal model which suffers from the class separation problem. The objective of FDA can be formulated as maximizing the sum of all the pairwise distances between different classes, which will overemphasize the large distance of the already well-separated classes, and confuse the small distance classes that are close in the original feature space. Many models have been proposed to solve the class separation problem of FDA. Loog et al. [10] proposed the approximate pairwise accuracy criterion (aPAC), which uses a weighting function to emphasize the close class pairs in the between-class scatter matrix. Lotlikar and Kothari [11] developed the fractional-step FDA, which essentially is also a weighting approach but selects a subspace through fractional steps. By proving that FDA is equivalent to maximizing the arithmetic mean of all pairwise distances, Tao et al. [12] proposed to use the geometric mean, while Bian and Tao [2] proposed to use the harmonic mean, to replace the arithmetic mean used in FDA. Recently, the idea of maximizing the minimal pairwise distance was proposed by many authors to solve the class separation problem [15] [13] [14] [3]. Simultaneously maximizing all the pairwise distances was also proposed as a multi-objective optimization problem [1] to handle the class separation problem.

In this paper, we propose a new model of confused distance maximization (CDM) to solve the class separation problem. The objective of CDM is to maximize the distance of the most confusable classes, by weighting the pairwise distance according to the confusion ma-

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.196

trix estimated from the training data with a pre-learned classifier. Compared with the weighting criteria used in [10] and [11], CDM is more relevant to the classification accuracy since the weights are defined as the confusion probability learned from the training data. Furthermore, while the above mentioned models usually need a complex iterative optimization to solve the model, the computation of CDM is still an eigen-decomposition problem. This makes CDM indeed efficient and effective for large category problems. Experiments on two large-scale 3,755-class Chinese handwriting databases demonstrated that CDM can get the best performance compared with FDA and other competitive extensions.

The rest of the paper is organized as following: Section 2 gives an introduction of FDA and the class separation problem; Section 3 presents the proposed model of confused distance maximization (CDM); Section 4 reports the experimental results; and Section 5 draws the concluding remarks.

2 FDA and the Class Separation Problem

Let $\mu_k \in \mathbb{R}^d$ be the mean vector, and $\Sigma_k \in \mathbb{R}^{d\times d}$ be the covariance matrix for class k, where $k = 1 \cdots K$. The within-class and between-class scatter matrices are defined as:

$$S_w = \sum_{k=1}^{K} p_k \Sigma_k \,, \tag{1}$$

$$S_b = \frac{1}{2} \sum_{i,j=1}^{K} p_i p_j (\mu_i - \mu_j)(\mu_i - \mu_j)^\top \,, \tag{2}$$

where $p_k = N_k/N$, $N = \sum_{k=1}^{K} N_k$ (N_k is the number of samples in class k). The objective of FDA is to learn a transformation matrix $W \in \mathbb{R}^{d\times d'}$ to minimize the within-class variance and as well as maximize the between-class variance. There are many formulations of FDA. Two typical criteria are given in the following [5]:

$$\max \ \mathrm{tr}\left\{ \left(W^\top S_w W\right)^{-1} \left(W^\top S_b W\right) \right\} \,, \tag{3}$$

$$\max \ \ln\left|W^\top S_b W\right| - \ln\left|W^\top S_w W\right| \,, \tag{4}$$

which are equivalent to a constrained problem:

$$\max_{W\in\mathbb{R}^{d\times d'}} \quad \mathrm{tr}\left(W^\top S_b W\right) \quad \text{s.t. } W^\top S_w W = I \,, \tag{5}$$

where I is the identity matrix. Usually, this model is solved by a two-step approach. The first step is the whitening:

$$W_{\mathrm{whiten}} = P\Lambda^{-1/2} \in \mathbb{R}^{d\times d} \,, \tag{6}$$

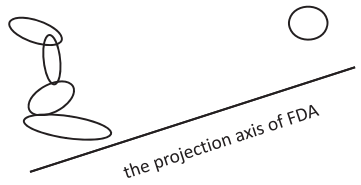


Figure 1. An illustration of the class separation problem.

where P is the eigenvector matrix and Λ is the diagonal eigenvalue matrix[1] of the within-class scatter matrix:

$$S_w = P\Lambda P^\top \,. \tag{7}$$

The whitening transformation satisfy

$$W_{\mathrm{whiten}}^\top S_w W_{\mathrm{whiten}} = I \,. \tag{8}$$

Let $W_{\mathrm{FDA}} = W_{\mathrm{whiten}} W$, we can rewrite FDA of (5) as:

$$\begin{aligned} \max_{W\in\mathbb{R}^{d\times d'}} \quad & \mathrm{tr}\left(W^\top W_{\mathrm{whiten}}^\top S_b W_{\mathrm{whiten}} W\right) \,, \\ \text{s.t.} \quad & W^\top W = I \,. \end{aligned} \tag{9}$$

Hence the second step of FDA is to solve (9). This is exactly the principal component analysis (PCA) among $W_{\mathrm{whiten}}^\top \mu_1, \cdots, W_{\mathrm{whiten}}^\top \mu_K$. Let $\triangle_{ij}$ be the distance of class i and j in the transformed subspace

$$\triangle_{ij} = \left\| W^\top W_{\mathrm{whiten}}^\top (\mu_i - \mu_j) \right\|_2^2 \,. \tag{10}$$

The model of (9) is equivalent to:

$$\max_{W\in\mathbb{R}^{d\times d'}} \quad \sum_{i,j=1}^{K} p_i p_j \triangle_{ij} \quad \text{s.t. } W^\top W = I \,. \tag{11}$$

The first step of FDA is to learn a suitable distance metric: in the whitened space, the Euclidean distance become the optimal measurement. The dimensionality reduction is actually implemented in the second step of model (11). Because (11) is to maximize the sum of all the pairwise distances, it will cause the class separation problem [10]. To illustrate this, consider that one class is located remotely from the other classes and can be considered as an outlier (Figure 1). In this case, by optimizing (11), the projection axis of FDA is the one that separates the outlier from the remaining classes as much as possible. The pairs of large-distance classes completely dominate the solution of (11). As a consequence, there is a large overlap among the remaining classes, leading to an overall low and suboptimal classification performance.

[1] The diagonal zero values in Λ are set to be a small constant.

To solve the class separation problem, Tao et al. [12] proposed to maximize the geometric mean $\{\max \sum_{i\neq j} p_i p_j \log \triangle_{ij}\}$. Bian and Tao [2] further proposed to maximize the harmonic mean $\{\max -\sum_{i\neq j} p_i p_j \triangle_{ij}^{-1}\}$. Recently, many authors have proposed to maximize the minimal distance $\{\max \min_{i\neq j} \triangle_{ij}\}$ [15] [13] [14] [3]. Abou-Moustafa et al. [1] further proposed to maximize all the pairwise distances simultaneously. Although these methods have reported improved performance, they are all based on some complex iterative optimization procedures to solve the models, which makes them not scalable for large category (e.g. thousands of classes) problems.

3 Confused Distance Maximization

The proposed confused distance maximization (CDM) model is aimed to solve the class separation problem of FDA, and is very efficient and effective for large category problems.

3.1 CDM

Instead of maximizing the sum of all the pairwise distances, we focus on maximizing the distance of the most confusable classes. To do so, (11) is generalized by introducing a weighting function:

$$\max_{W\in\mathbb{R}^{d\times d'}} \sum_{i,j=1}^{K} f_{ij} p_i p_j \triangle_{ij} \quad \text{s.t. } W^\top W = I\,, \quad (12)$$

where $f_{ij} \geq 0$ is a weighting function that depends on the probability of confusion between class i and class j. The model of (12) is equivalent to:

$$\max_{W\in\mathbb{R}^{d\times d'}} \operatorname{tr}\left(W^\top \widehat{S_b} W\right) \quad \text{s.t. } W^\top W = I\,, \quad (13)$$

where

$$\widehat{S_b} = \sum_{i,j=1}^{K} f_{ij} p_i p_j (\widehat{\mu_i} - \widehat{\mu_j})(\widehat{\mu_i} - \widehat{\mu_j})^\top\,, \quad (14)$$

and $\widehat{\mu_i} = W_{\text{whiten}}^\top \mu_i$ is the whitened class-mean. The model of (13) can be solved by taking the columns of the $d \times d'$ matrix W to be the d' eigenvectors corresponding to the d' largest eigenvalues of $\widehat{S_b}$. The final transformation matrix of CDM is then defined as:

$$W_{\text{CDM}} = W_{\text{whiten}} W \in \mathbb{R}^{d\times d'}\,. \quad (15)$$

Clearly, choosing f_{ij} to be the constant function will reduce CDM to the ordinary FDA. The complete procedure of CDM is shown in Algorithm 1, which is very efficient for large category problems.

Algorithm 1 Confused Distance Maximization

Input:

mean and covariance: $\mu_k, \Sigma_k, k = 1\cdots K$
prior probabilities: $p_k, k = 1\cdots K$
confusion matrix: $F = \{f_{ij}\} \in \mathbb{R}^{K\times K}$

1: $S_w = \sum_{k=1}^{K} p_k \Sigma_k$
2: $W_{\text{whiten}} = P\Lambda^{-1/2}$ where $S_w = P\Lambda P^\top$
3: $\widehat{\mu_i} = W_{\text{whiten}}^\top \mu_i\,, \quad i = 1, \cdots, K$
4: $\widehat{S_b} = \sum_{i,j=1}^{K} f_{ij} p_i p_j (\widehat{\mu_i} - \widehat{\mu_j})(\widehat{\mu_i} - \widehat{\mu_j})^\top$
5: W be the d' eigenvectors of $\widehat{S_b}$ corresponding to the d' largest eigenvalues

Output: $W_{\text{CDM}} = W_{\text{whiten}} W \in \mathbb{R}^{d\times d'}$

3.2 Confusion Matrix

The key problem of CDM is how to define the confusion matrix $F = \{f_{ij}\} \in \mathbb{R}^{K\times K}$. We estimate the confusion matrix from the training data with a pre-learned classifier.

$$f_{ij} = \begin{cases} \dfrac{N_{i\rightsquigarrow j}}{N_i}, & i \neq j \\ \\ 0, & i = j \end{cases} \quad (16)$$

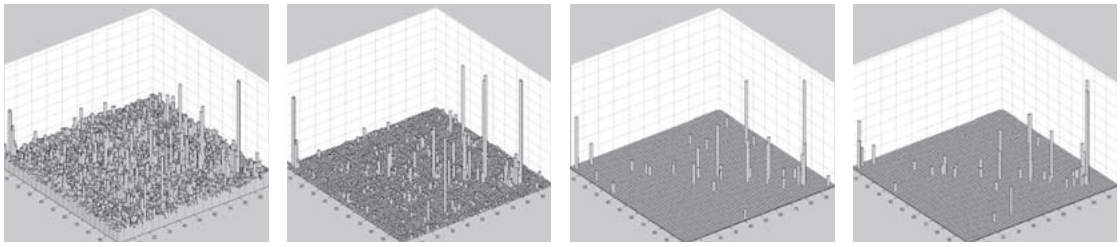
where N_i is the number of samples in class i, and $N_{i\rightsquigarrow j}$ is the number of samples that come from class i but classified into class j by a specific classifier. In real applications, it is a small probability event to get a zero confusion matrix ($f_{ij} = 0, \forall i, j$), because the classification accuracy can seldom be 100%.[2]

The confusion matrix is classifier-specific. Different classifiers will have their own dimensionality reduction matrices, which are estimated based on their own confusion information. This makes CDM more relevant to the classification accuracy. The classes are weighted according to their confusion probabilities: if one class is very likely to be confused with another class, the distance between them are much more important in the dimensionality reduction model (12).

To get better generalization performance, the confusion matrix should be estimated from a dataset which is different from the dataset used to train the basic classifier. In our experiments, we randomly partition the training set into two subsets[3]: using $3/4$ for training the basic classifier, and $1/4$ for estimating the confusion matrix. After that the CDM model is trained on the whole training dataset, and the classifier is also retrained on the whole training dataset in the reduced space.

[2] In practice, we can also define a modified confusion matrix as $f'_{ij} = (1-\delta) f_{ij} + \delta$, where $0 \leq \delta \leq 1$ is a tradeoff parameter to balance between CDM and FDA.

[3] When the number of training samples is small, the cross-validation may be better to estimate a more precise confusion matrix.



(a) aPAC (b) POWER (c) CDM (Euclidean) (d) CDM (MQDF)

Figure 2. The 100×100 weighting matrix of different methods for the first 100 classes of the $3,755$-class problem.

The confusion matrix can be defined either in the original feature space or the reduced low-dimensional space. Since we want to reduce the dimensionality from $\mathbb{R}^d$ to $\mathbb{R}^{d'}$, we can estimate a confusion matrix in the $\mathbb{R}^d$ space, and then use the confusion matrix to reduce the dimensionality based on Algorithm 1. We denote this method CDM1. However, the confusion matrix in the original space may be different from that in the low-dimensional space, and will not reflect the real confusion information. To solve this problem, we can first use FDA to reduce the dimensionality to $\mathbb{R}^{d'}$, and then estimate the confusion matrix in this low-dimensional space. After that, the confusion matrix is incorporated into the CDM algorithm to learn the dimensionality reduction matrix again. We call this model CDM2. Because of the more accurate confusion information, CDM2 is hoped to give better performance than CDM1.

3.3 Comparing Other Weighting Methods

The idea of weighting the pairwise distance was also proposed by other authors [10] [11]. The approximate pairwise accuracy criterion (aPAC) proposed by [10] used a weighting function as following

$$f_{ij} = \frac{1}{2d_{ij}^2} \operatorname{erf}\left(\frac{d_{ij}}{2\sqrt{2}}\right) , \tag{17}$$

where $\operatorname{erf}(\cdot)$ is the error function[4] which results in $[-1, 1]$, and d_{ij} is the distance of the whitened-mean between class i and class j:

$$d_{ij} = \left\| W_{\text{whiten}}^{\top}(\mu_i - \mu_j) \right\|_2 . \tag{18}$$

Lotlikar and Kothari [11] proposed a weighting function as:

$$f_{ij} = d_{ij}^{-m} , \tag{19}$$

where f_{ij} should drop fast than d_{ij} increasing, then m is suggested to be $m \geq 3$. Since Eq. (19) is a power function, we denote this method POWER. The idea of reducing the dimensionality by a fractional-step proposed by [11], can be actually used for all the weighing functions (CDM, aPAC and POWER), therefore is not the focus of this paper.

The weighting matrices of aPAC and POWER are based on the pairwise distance d_{ij}, while the confusion matrix used in CDM is based on the classification results. This makes CDM more relevant to the classification accuracy. We show the weighting matrices of different methods for a 3,755-class online handwriting problem in Figure 2. The weighting matrix is symmetrical for aPAC and POWER, but unsymmetrical for CDM. Furthermore, the weighting matrix of CDM is much more sparse than aPAC and POWER, because each class is only confused with a small number of classes. We can find that the weighting matrix of POWER ($m = 8$) in Figure 2(b) is much like the confusion matrix of CDM, but there are many additional small noises (small non-zero elements). The locality property makes CDM more relevant to the classification accuracy by focusing on the most confusable classes. Moreover, the confusion matrix used in CDM is classifier-specific. We show the confusion matrices for two different classifiers (Section 4.2) in Figure 2(c) and Figure 2(d). We can find that the confusion matrices are different for different classifiers. This indicates that CDM is more relevant to a particular classifier, since the weighting matrices of aPAC and POWER are independent with the classifier.

4 Experiments

4.1 Database

We evaluate the performance of different models on two $3,755$-class Chinese handwriting databases [8]: the offline handwriting database CASIA-HWDB1.1 and the online handwriting database CASIA-OLHWDB1.1. Both of them contain handwritten Chinese characters from 300 writers (240 for training and 60 for testing). Each writer has about $3,755$ characters (one for each class). For representing a offline character sample, we extract features from gray-scale character images (back-

[4] http://wikipedia.org/wiki/Error_function

Table 2. The classification accuracy (%) of different models on the offline handwriting database.

$d = 512$	Euclidean					MQDF				
d'	FDA	aPAC	POWER	CDM1	CDM2	FDA	aPAC	POWER	CDM1	CDM2
60	78.80	78.84	79.23	79.19	**79.69**	86.35	86.33	86.55	86.80	**87.04**
80	80.56	80.58	80.86	80.83	**81.00**	88.14	88.13	88.29	88.44	**88.49**
100	81.43	81.41	81.55	81.54	**81.67**	88.87	88.87	89.00	89.13	**89.22**
120	81.88	81.89	82.00	81.89	**82.02**	89.26	89.26	89.34	89.40	**89.41**
140	82.09	82.05	**82.13**	82.03	82.10	89.47	89.46	89.47	89.55	**89.58**
160	82.13	82.13	**82.16**	82.12	82.10	89.53	89.49	**89.54**	89.49	**89.54**
180	**82.19**	82.16	82.13	82.13	82.13	89.51	89.50	**89.53**	89.47	89.52
200	**82.15**	82.13	82.10	82.07	82.06	89.39	89.32	89.37	**89.40**	89.38
Average	81.40	81.40	81.52	81.48	**81.60**	88.82	88.80	88.89	88.96	**89.02**

	#class	#dim	#train	#test
offline	$3,755$	512	$897,758$	$223,991$
online	$3,755$	512	$898,573$	$224,559$

Table 1. The database information.

ground eliminated) using the normalization-cooperated gradient feature (NCGF) method [7]. For representing a online character sample, we use a benchmark feature extraction method [9]: 8-direction histogram feature extraction combined with pseudo 2D bi-moment normalization (P2DBMN). We also add the direction values of off-strokes (pen lifts) to real strokes with a weight of 0.5 [4]. The feature dimensionality is 512 for both the offline and the online handwriting samples. We summarize the complete information of the two databases in Table 1 .

4.2 Experimental Setting

Two classifiers are used to evaluate the classification performance in the reduced low-dimensional space. For large-category problems, the nearest neighbor classifier is too expensive, we only consider the nearest class-mean classifier: $x \in \arg\min_{k=1}^{K} d(x, \mu_k)$, where $d(x, \mu_k)$ is the distance between x and μ_k, which can be either the Euclidean distance:

$$d_e(x, \mu_k) = \|x - \mu_k\|_2^2 \ , \tag{20}$$

or the Mahalanobis distance[5]:

$$d_m(x, \mu_k) = (x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log|\Sigma_k| \ . \tag{21}$$

To efficiently compute Σ_k^{-1}, we use the modified quadratic discriminant function (MQDF) [6] method, which replace the minor eigenvalues of Σ_k with a constant to stabilize the generalization performance. The minor eigenvalues are set to be common for all the classes and selected with cross-validation on the training dataset. The number of principal components used in MQDF is set to be 50 for all the methods.

The FDA (Section 2), CDM1, CDM2 (Section 3.2) and aPAC, POWER (Section 3.3) are compared according to the classification accuracy in the reduced low-dimensional space ($d = 512 \rightarrow d' = 60, 80, \cdots, 200$). For the POWER method, we change $m = 3, 4, \cdots, 12$ for Eq. (19) and report the best performance on the test dataset.

4.3 Experimental Results

The experimental results are shown in Table 2 (offline) and Table 3 (online). We can see that the classification accuracies of aPAC and FDA are nearly the same. For the POWER method, the best results are achieved when $m = 9$ for the offline database and $m = 8$ for the online database. From the results we can conclude that: (1) The CDM can get consistently better performance than the other methods, especially when the reduced dimensionality is small. This indicates that the confusion matrix (CDM) is more suitable than the distance based weighting (aPAC, POWER), to measure the importance of the pairwise distance in dimensionality reduction. (2) CDM2 can achieve higher classification accuracy than CDM1, which implies that the confusion matrix estimated in the low-dimensional space (CDM2) is more accurate than that estimated in the original space (CDM1), because the final classifier is directly evaluated in the reduced space. (3) Compared with the baseline model FDA, the CDM algorithms can improve the classification accuracy consistently, for both Euclidean distance based classifier and MQDF, on either the offline or the online handwriting database. The computation cost of CDM is nearly the same with FDA, except the process of confusion matrix estimation. This makes CDM an efficient and effective candidate for large category dimensionality reduction.

[5]This is actually the quadratic discriminant function (QDF), derived from the Bayes decision theory, under the assumption of Gaussian class-conditional distribution.

Table 3. The classification accuracy (%) of different models on the online handwriting database.

$d = 512$	Euclidean					MQDF				
d'	FDA	aPAC	POWER	CDM1	CDM2	FDA	aPAC	POWER	CDM1	CDM2
60	86.12	86.20	86.46	86.58	**86.85**	91.30	91.32	91.53	91.72	**91.77**
80	87.32	87.33	87.65	87.61	**87.79**	92.40	92.39	92.58	**92.66**	**92.66**
100	87.85	87.87	88.04	88.09	**88.20**	92.85	92.84	93.05	**93.09**	**93.09**
120	88.15	88.15	88.28	88.29	**88.37**	93.11	93.12	93.18	**93.29**	93.25
140	88.22	88.22	88.39	88.35	**88.41**	93.20	93.18	93.29	93.33	**93.37**
160	88.24	88.24	88.38	88.37	**88.44**	93.22	93.18	93.23	93.29	**93.32**
180	88.20	88.20	88.30	88.30	**88.34**	93.19	93.17	93.19	**93.21**	**93.21**
200	88.15	88.16	88.25	88.23	**88.30**	93.09	93.06	93.13	**93.15**	**93.15**
Average	87.78	87.80	87.97	87.98	**88.09**	92.80	92.78	92.90	92.97	**92.98**

5 Conclusion

In this paper, the confused distance maximization (CDM) is proposed to solve the class separation problem for large category dimensionality reduction. The objective of CDM is to maximize the distance of the most confusable classes, by weighting the pairwise distance according to a confusion matrix estimated from the training data with a pre-learned classifier. Compared with other weighting based methods, CDM is more relevant to the classification accuracy. Moreover, the computation of CDM is still an eigen-decomposition problem, which makes CDM efficient for large scale applications. Experiments on two 3,755-class Chinese handwriting databases demonstrated that CDM can achieve the best performance compared with FDA and other weighting based extensions. Our future work involves the integration of the fractional-step [11] and CDM for further boosting the performance. Because the confusion matrix is highly dependent on the particular classifier, the joint learning of the confusion matrix, classifier and the dimensionality reduction matrix is also an interesting topic.

Acknowledgements

This work was supported by the National Basic Research Program of China (973 Program) Grant 2012CB316302, the Strategic Priority Research Program of the CAS (Grant XDA06030300), the National Natural Science Foundation of China (NSFC) Grants 60825301 and 60933010.

References

[1] K. Abou-Moustafa, F. De La Torre, and F. Ferrie. Pareto discriminant analysis. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2010.

[2] W. Bian and D. Tao. Harmonic mean for subspace selection. *Proc. Int'l Conf. Pattern Recognition*, 2008.

[3] W. Bian and D. Tao. Max-Min distance analysis by using sequential SDP relaxation for dimension reduction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2011.

[4] K. Ding, G. Deng, and L. Jin. An investigation of imaginary stroke techinique for cursive online handwriting Chinese character recognition. *Proc. Int'l Conf. Document Analysis and Recognition*, 2009.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[6] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake. Modified quadratic discriminant functions and the application to Chinese character recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1987.

[7] C.-L. Liu. Normalization-cooperated gradient feature extraction for handwritten character recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2007.

[8] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang. CASIA online and offline Chinese handwriting databases. *Proc. Int'l Conf. Document Analysis and Recognition*, 2011.

[9] C.-L. Liu and X.-D. Zhou. Online Japanese character recognition using trajectory-based normalization and direction feature extraction. *Proc. Int'l Workshop Frontiers in Handwriting Recognition*, 2006.

[10] M. Loog, R. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2001.

[11] R. Lotlikar and R. Kothari. Fractional-step dimensionality reduction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2000.

[12] D. Tao, X. Li, X. Wu, and S. Maybank. Geometric mean for subspace selection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2009.

[13] B. Xu, K. Huang, and C.-L. Liu. Dimensionality reduction by minimal distance maximization. *Proc. Int'l Conf. Pattern Recognition*, 2010.

[14] Y. Yu, J. Jiang, and L. Zhang. Distance metric learning by minimal distance maximization. *Pattern Recognition*, 2011.

[15] Y. Zhang and D. Yeung. Worst-case linear discriminant analysis. *Advances in Neural Information Processing Systems*, 2010.