基于主动判别函数的手写体识别*

孙广玲[1+], 刘家锋[1], 唐降龙[1], 石大明[2], 赵 巍[1]

[1](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[2](School of Computer Engineering, Nanyang Technological University, Singapore)

Active Discriminant Function for Handwriting Recognition

SUN Guang-Ling[1+], LIU Jia-Feng[1], TANG Xiang-Long[1], SHI Da-Ming[2], ZHAO Wei[1]

[1](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](School of Computer Engineering, Nanyang Technologicl University, Singapore)

+ Corresponding author: Phn: +86-451-86417602, E-mail: sunguangling@hit.edu.cn

Received 2003-12-29; Accepted 2004-11-03

Sun GL, Liu JF, Tang XL, Shi DM, Zhao W. Active discriminant function for handwriting recognition. *Journal of Software*, 2005,16(4):523−532. DOI: 10.1360/jos160523

Abstract: A novel recognition method called Active Discriminant Function (ADF) for handwriting recognition is presented. First, statistical feature based Active Prototype Model (APM) in the principal subspace is proposed and an optimal APM corresponding to an unknown pattern is obtained. Second, ADF that is a weighted summation of two distances is proposed. One measures the distance between an unknown pattern and the principal subspace; the other measures the distance between an unknown pattern and the minor subspace. Third, as parameters of ADF, constraints for APM are optimized by applying Minimum Classification Error (MCE) criterion. The optimal constraints help to improve recognition accuracy of ADF. Finally, experiments are conducted on handwritten financial Chinese characters used in bank bill, and empirical results demonstrate that ADF is fairly promising for handwriting recognition.

Key words: handwriting recognition; active discriminant function; active prototype model; principal component analysis; minimum classification error criterion

* Supported by the Foundation of Harbin City for the Intellectual, Heilongjiang Province of China under Grant No.2004AFXXJ053 (哈尔滨市后备人才基金项目)

SUN Guang-Ling was born in 1973. She is a Ph.D. and a lecturer at School of Computer Science and Technology, Harbin Institute of Technology. Her research interests include character recognition, image processing and analysis, pattern recognition. **LIU Jia-Feng** was born in 1968. He is a Ph.D. and an assistant professor at School of Computer Science and Technology, Harbin Institute of Technology. His research areas are image processing and analysis, pattern recognition and artificial intelligence. **TANG Xiang-Long** was born in 1960. He is a professor and doctoral supervisor at School of Computer Science and Technology, Harbin Institute of Technology. His research areas are pattern recognition and intelligence system. **SHI Da-Ming** was born in 1971. He is a Ph.D. and an assistant professor at School of Computer Engineering, Nanyang Technological University, Singapore and a senior member of the IEEE. His research areas are image processing and analysis, machine learning and pattern recognition. **ZHAO Wei** was born in 1972. She is a Ph.D. candidate and a lecturer at School of Computer Science and Technology, Harbin Institute of Technology. Her research areas are character recognition and pattern recognition.

摘 要: 提出了一种新的基于主动判别函数的手写体识别方法.首先,提出了主子空间中基于统计特征的主动原型模板并给出了对应于待识模式的最优主动原型模板.然后,提出了主动判别函数,它是两个加权距离之和;其中一个是待识模式与主子空间的距离,另一个是待识模式与次子空间的距离.其次,作为主动判别函数的参数,主动原型模板的约束可应用最小分类错误准则进行优化,这一最优约束有助于提高主动判别函数的识别正确率.最后,在银行票据中使用的手写金融汉字样本库上进行实验.实验结果表明,主动判别函数在手写体识别方面是非常有前景的.

关键词: 手写体识别;主动判别函数;主动原型模板;主成分分析;最小分类错误准则

中图法分类号: TP18 文献标识码: A

1 Introduction

Handwriting recognition is a well known important but difficult problem because of great variations involved in handwriting. A number of research efforts have been made in this area[1−5]. Of the numerous schemes and strategies proposed, deformable model is a distinctive and effective technique. In our previous work, Active Handwriting Model (AHM), as a type of deformable model, has been proposed to deal with linear and nonlinear variations involved in handwritten Chinese characters and good performances have been achieved[6−8]. However, there is a notable drawback in AHM: landmark points labeling problem. In AHM, landmark points need to be labeled manually. Unfortunately, the manual labeling landmark points is a fairly difficult and boring task especially for complex characters, such as Chinese characters. Moreover, manual labeling will definitely bring into additional errors due to some uncontrollable factors. To overcome the drawback, we propose a novel statistical feature based deformable model, which is called Active Prototype Model (APM). Examples of statistical features are peripheral shape feature, stroke density feature, weighted directional code feature, directional element feature and so on[9−11]. It is obvious that statistical features can be obtained much easier than landmark points. We apply Principal Component Analysis (PCA) technique[12] to get principal subspace for each category by learning samples of each category. After that, APM is generated in a similar manner to AHM by adjusting parameters corresponding to the principal modes of variation. To avoid deviating too much from the original model, APM must be generated under certain constraints. Then, an optimal APM in the principal subspace corresponding to an unknown pattern is generated.

The distance between an unknown pattern and the corresponding optimal APM is defined as a distance between the unknown pattern and principal subspace. If the number of principal components is less than the dimension of the original feature space, to fully utilize the information involved in minor subspace, we also define a distance between an unknown pattern and the minor subspace. This distance is based on the residual in minor subspace.

Active Discriminant Function (ADF) that is a weighted summation of the above two distances is proposed. Further, as parameters of ADF, constraints for APM can be optimized using certain optimal criterions. We utilize Minimum Classification Error (MCE) criterion[13,14] to search the optimal constraints. MCE is one of the optimal criterions in discriminant learning and is directly related with classification error; therefore, optimal constraints help to improve recognition accuracy of ADF.

The remainder of the paper is organized as follows: In Section 2, APM in the principal subspace is proposed and an optimal APM is obtained. In Section 3, two distances are calculated, one is the distance between an unknown pattern and the principal subspace, and the other is the distance between an unknown pattern and the minor subspace. Then, ADF that calculates a weighted summation of the two distances is presented and constrains for APM are optimized based on MCE criterion learning. Empirical results are exhibited in Section 4. Conclusions are given in Section 5.

2 Active Prototype Model and Optimal Active Prototype Model

In this section, we first review AHM. Then, APM is defined and an optimal APM is obtained.

2.1 Active handwriting model

The first step of AHM is to find a principal subspace that is spanned by the principal eigenvectors obtained by learning samples using PCA. Then, through adjusting parameter "*b*", which is corresponding to the principal components of variation, AHM "Γ" can be produced:

$$\begin{aligned} &\Gamma = \psi + U \cdot b, \quad U = [u_1, u_2, ..., u_k], \quad \lambda_1 \geq \lambda_2, ..., \geq \lambda_k \\ &\Gamma \in R^{d\times 1}, \psi \in R^{d\times 1}, U \in R^{d\times k}, b \in R^{k\times 1} \\ &1 \leq k \leq d \end{aligned} \tag{1}$$

Where ψ denotes the mean vector of a category in the original feature space, λ_j and u_j, $1 \leq j \leq k$, denote eigenvalues and eigenvectors of the covariance matrix of the category. The eigenvalues are sorted in a decreasing order and the eigenvectors are sorted correspondingly. d denotes the dimension of the original feature space, and k denotes the number of the principal components.

2.2 Active prototype model

Note that Γ is a vector that belongs to $R^{d\times 1}$. This means Γ is generated in the original space. In fact, a deformable model can also be generated in principal subspace. When Γ is projected into the principal subspace, a deformable model in the principal subspace "Γ'" is produced. It can be implemented by multiplying U^T at the two sides of (1):

$$U^T \cdot \Gamma = U^T \cdot \psi + U^T \cdot U \cdot b \tag{2}$$

Accordingly, the following result is derived:

$$\begin{aligned} &\Gamma' = \psi' + b \\ &\Gamma' = U^T \cdot \Gamma, \quad \psi' = U^T \cdot \psi, \quad U^T \cdot U = I_{k\times k} \\ &\Gamma' \in R^{k\times 1}, \quad \psi' \in R^{k\times 1}, \quad b \in R^{k\times 1} \end{aligned} \tag{3}$$

Γ' is named as Active Prototype Model.

2.3 Optimal active prototype model

Assume $x \in R^{d\times 1}$ denotes a statistical feature vector and it can be projected into the principal subspace:

$$\begin{aligned} &x' = U^T \cdot x \\ &x' \in R^{k\times 1} \end{aligned} \tag{4}$$

We adopt *power norm* as a distance metric between x' and Γ':

$$\left\| x' - \Gamma' \right\|_{p/r} = \left(\sum_{j=1}^{k} \left| x'_j - \Gamma'_j \right|^p \right)^{\frac{1}{r}} \tag{5}$$

According to (3), (4) and (5), the following formulation can be obtained:

$$\begin{aligned} &\left\| x' - \Gamma' \right\|_{p/r} = \left\| \overline{x}' - b \right\|_{p/r} = \left(\sum_{j=1}^{k} \left| \overline{x}'_j - b_j \right|^p \right)^{\frac{1}{r}} \\ &\overline{x}' = U^T \cdot (x - \psi), \ \overline{x}'_j = (x - \psi)^T \cdot u_j, \ \overline{x}' \in R^{k\times 1} \end{aligned} \tag{6}$$

Since b can be adjusted, $\left\| \overline{x}' - b \right\|_{p/r}$ is in fact a dynamic distance. Our target is to minimize the distance by

searching b in its searching range. For b_j, j=1,2, k are uncorrelated, the optimal b_j can be searched independently. Let b_j^* denote an optimal b_j. Given constraints of the searching range, it is not difficult to find b_j^* as follows:

$$b_j^* = \begin{cases} \overline{x}_j, & \text{if } -\theta_j \le \overline{x}_j \le \theta_j \\ -\theta_j, & \text{if } \overline{x}_j < -\theta_j \\ \theta_j, & \text{if } \overline{x}_j > \theta_j \end{cases}, \quad 1 \le j \le k \tag{7}$$
$$\theta_j > 0$$

where $-\theta_j$ and θ_j are left searching boundary and right searching boundary of the j-th dimension, respectively. All b_j^* compose b*. By bringing b* into (3), an optimal APM "Γ'" corresponding to x is obtained accordingly:

$$\Gamma'^* = \psi' + b^* \tag{8}$$

3 Active Discriminant Function

Based on Γ'*, the distance between x and Γ'* is derived as below:

$$\| x' - \Gamma'^* \|_{p/r} = \left[\sum_{j=1}^{k} | \overline{x}_j - b_j^* |^p \right]^{\frac{1}{r}} = \left(\sum_{j=1}^{k} f_p(x;\psi,u_j,\theta_j) \right)^{\frac{1}{r}} \tag{9}$$

where

$$f_p(x;\psi,u_j,\theta_j) = \begin{cases} (\overline{x}_j - \theta_j)^p, \overline{x}_j > \theta_j \\ (-\overline{x}_j - \theta_j)^p, \overline{x}_j < -\theta_j, \quad 1 \le j \le k \\ 0, -\theta_j \le \overline{x}' \le \theta_j \end{cases} \tag{10}$$

Naturally, we can take the distance as a distance between x and the principal subspace. If we use $D_{pri}(x)$ to represent the distance between them, it will be as follows:

$$D_{pri}(x) = \left(\sum_{j=1}^{k} f_p(x;\psi,u_j,\theta_j) \right)^{\frac{1}{r}}, 1 \le j \le k \tag{11}$$

However, $D_{pri}(x)$ only measures the distance between x and the principal subspace. As stated in introduction, when the number of principal components is less than the dimension of the original feature space, information involved in the minor subspace should also be considered. Therefore, we define another distance which measures the distance between x and the minor subspace. Let $\overline{\varepsilon}^2(x)$ denote an average residual, it is to be calculated as below:

$$\overline{\varepsilon}^2(x) = [\| x - \psi \|_2^2 - \sum_{j=1}^{k} (\overline{x}_j)^2] / (d-k) \tag{12}$$

where $\| \bullet \|_2$ denotes the 2-*norm* of a vector.

$$\hat{\overline{x}}' = | \overline{\varepsilon}(x) | \tag{13}$$

$\hat{\overline{x}}'$ is an approximation of $| \overline{x}_j |, k+1 \le j \le d$. If we use $D_{minor}(x)$ to represent the distance between x and the minor subspace, it will be as follows:

$$D_{\min or}(x) = \left((d-k) \cdot (\hat{\overline{x}}')^p \right)^{\frac{1}{r}} = \left((d-k)^{1-\frac{p}{2}} \cdot \left[\| x - \psi \|_2^2 - \sum_{j=1}^{k} (\overline{x}_j)^2 \right]^{\frac{p}{2}} \right)^{\frac{1}{r}} \tag{14}$$

The weighted summation of the two distances is defined as ADF. If we use Λ to represent a parameter set of

ADF, the following form can be obtained:

$$\begin{aligned} &g_{ADF}(x;\Lambda) = (1-a)\cdot D_{pri}(x) + a\cdot D_{\min or}(x) \\ &\Lambda = \{\psi, u_j, \theta_j, k,\ 1\le j\le k\},\quad 0\le a\le 1 \end{aligned} \tag{15}$$

where a denotes the weight.

With no loss of generality, we choose p=r=1 to implement a specific ADF of the i-th class:

$$g^*_{ADF,i}(x;\Lambda_i) = (1-a)\cdot\sum_{j=1}^{k} f_1(x;\psi_i,u_{i,j},\theta_{i,j}) + a\cdot\{(d-k)\cdot\left[\|x-\psi_i\|_2^2 - \sum_{j=1}^{k}(\vec{x}'_{i,j})^2\right]\}^{\frac{1}{2}} \tag{16}$$

where $\Lambda_i = \{\psi_i, u_{i,j}, \theta_{i,j}, k_i, 1\le j\le k_i\}$. For simplicity, we set $k_1 = k_2 = ... = k_C = k$. $\Lambda_{tot} = \{\Lambda_1, \Lambda_2, ..., \Lambda_C\}$ represents the parameter sets of all categories and C is the number of the categories. From (16), it can be seen that if k equals to d, ADF will just be the distance between x and the principal subspace. Considering a simple expression, $g^*_{ADF,i}(x;\Lambda_i)$ is replaced with a concise form $g^*_i(x;\Lambda_i)$ later.

As stated in introduction, to avoid deviating too much from the original model, the adjusted parameters b must be constrained in a range. θ is just such a range. We have defined ADF in which θ is one of its parameters. Therefore, ADF is possible to be learned based on certain optimal criterion, and the corresponding optimal θ can be obtained. MCE criterion which is directly related with classification error is adopted, so that the optimal θ is helpful for improving recognition accuracy of ADF. MCE criterion is one of the criterions in discriminative learning and its original formulation has been provided in Ref.[13]. MCE criterion based learning algorithm is implemented by using stochastic approximation to minimize an objective function. In our problem, the learning algorithm of θ based on MCE criterion is outlined in the following.

1) Define misclassification measure function $d^*_i(x;\Lambda_i,\Lambda_o)$:

$$\begin{aligned} &d^*_i(x;\Lambda_i,\Lambda_o) = -g^*_i(x;\Lambda_i) + g^*_o(x;\Lambda_o),\ \ x\in\omega_i \\ &o = \arg\min_{1\le j\le C, j\ne i} g^*_j(x;\Lambda_j) \end{aligned} \tag{17}$$

2) Define a loss function $l^*_i(x;\Lambda_i,\Lambda_o)$ associated with a pattern:

$$l^*_i(x;\Lambda_i,\Lambda_o) = \frac{1}{1+e^{\zeta\cdot[d^*_i(x;\Lambda_i,\Lambda_o)+\alpha]}} \tag{18}$$

where ζ is a positive constant and α is an arbitrary constant.

3) Define an objective function $L^*(X;\Lambda)$ associated with all patterns:

$$L^*(X;\Lambda) = \sum_{n=1}^{N}\sum_{i=1}^{C} l^*_i(x_n;\Lambda_i,\Lambda_o)I(x_n\in\omega_i) \tag{19}$$

where N is the number of samples and $I(z)$ is an indicator function:

$$I(z) = \begin{cases} 1, & z \text{ is true} \\ 0, & z \text{ is false} \end{cases} \tag{20}$$

where N is the number of samples.

4) Minimize $L^*(X;\Lambda)$ using stochastic approximation:

$$\theta_{l,j,t+1} = \theta_{l,j,t} - \varepsilon_t\cdot\sum_{i=1}^{C}\frac{\partial l^*_i(x_n;\Lambda_i,\Lambda_o)I(x_n\in\omega_i)}{\partial\theta_{l,j}},\ \ 1\le l\le C, 1\le j\le k \tag{21}$$

where ε_t is an iterative learning rate, which is a small positive value satisfying: $\sum_{t=1}^{\infty}\varepsilon_t \to \infty$, $\sum_{t=1}^{\infty}\varepsilon_t^2 < \infty$. As far as the computation of $\sum_{i=1}^{C}\frac{\partial l_i^*(x_n;\Lambda_i,\Lambda_o)I(x_n\in\omega_i)}{\partial\theta_{l,j}}$ is concerned, see the appendix.

4 Experiments and Results

In this section, we first give a brief introduction to the experimental database, then gradient feature is extracted and empirical results are exhibited and analyzed.

4.1 An introduction to experimental database

In bank bill recognition system, the variations of handwritten financial Chinese characters such as "零壹贰……"(zero, one, two…) are so great that they are very meaningful to test the performance of ADF. Hence, we carry on experiments on these handwritten financial Chinese characters. There is a total of 21 categories in this problem and they are "零壹贰叁肆伍陆柒捌玖拾佰仟万亿元圆角分整正"(zero, one,…, ten, hundred,…,billion, yuan, jiao, fen, zheng, zheng), respectively. Samples are collected from real bank bill and part of the samples is shown in Fig.1. It can be seen that these samples have severe variations and connections between strokes.

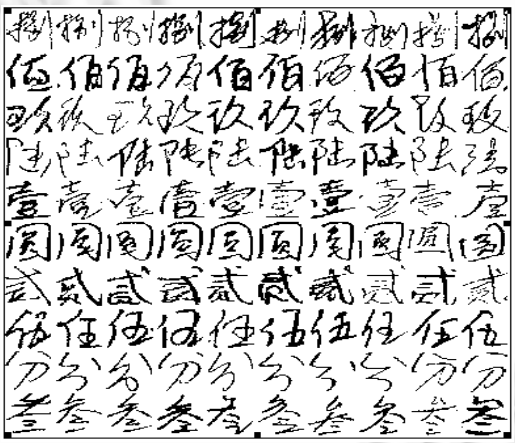

Fig.1 Part of samples of handwritten financial Chinese character

Each category includes 1 200 samples. These samples are collected from bank bill scanned by 300dpi. They are transformed into binary images and have arbitrary sizes. All samples of each category are partitioned randomly into training samples and testing samples. The number of training sample is 900 and the remaining 300 samples are used as testing samples.

4.2 Feature extraction and transformation

First, original image of a character is to be normalized into a uniform size. A nonlinear normalization algorithm proposed in Ref.[15] is adopted, and the normalization size is 64×64 pixels. During normalization, bilinear interpolation is conducted to enhance the smoothness and continuity of the normalized image; accordingly, a gray-scale between 0 and 255 is produced. That is to say, the normalized image is a gray-scale image.

Second, feature is to be extracted from the normalized image. Directional feature has been proved very effective in handwriting recognition such as weighted directional code feature[10] and directional element feature[11]. However, they are only suitable for binary image. Here, we extract gradient features from gray-scale normalized

image and the procedure is given as follows:

1) Make an image gray-scale transformation so that mean gray-scale and minimum gray-scale are 0 and −1, respectively.

2) Calculate $g(x,y)$ of a pixel positioned at (x,y) by adopting Sobel edge detection operator.

$$\begin{aligned} g(x,y) &= [g_x(x,y), g_y(x,y)]^T \\ g_x(x,y) &= h(x+1,y-1) + 2\cdot h(x+1,y) + h(x+1,y+1) \\ &\quad - h(x-1,y+1) - 2\cdot h(x-1,y) - h(x-1,y-1) \\ g_y(x,y) &= h(x-1,y-1) + 2\cdot h(x,y-1) + h(x+1,y-1) \\ &\quad - h(x+1,y+1) - 2\cdot h(x,y+1) - h(x-1,y+1) \end{aligned} \tag{22}$$

where $h(x,y)$ denotes the gray-scale of pixel positioned at (x,y). Accordingly, gradient strength $r(x,y)$ and orientation $\theta(x,y)$ are computed as follows:

$$r(x,y) = \sqrt{[g_x(x,y)]^2 + [g_y(x,y)]^2} \tag{23}$$

$$\theta(x,y) = \arctan\frac{g_y(x,y)}{g_x(x,y)}, -\frac{\pi}{2} \le \theta(x,y) < \frac{\pi}{2} \tag{24}$$

3) Distribute each pixel into one of the 8 quantized orientations according to its orientation $\theta(x,y)$.

4) Distribute each pixel in each orientation set into a subarea according to its coordinate (x,y). The subarea dividing is the same as the proposed in Ref.[11] such that the number of subarea is $7\times 7 = 49$.

5) Accumulate gradient strength $r(x,y)$ of pixels in each orientation and each subarea, so that a feature vector of dimension 392 is generated.

6) Apply Box-Cox transformation $x^{0.5}$ to each component of the feature vector such that the feature vector distribution is Gaussian-like[16].

Third, to improve computation efficiency and save store cost, the feature transformation technique is performed such that the dimension is reduced from 392 to 196[12].

4.3 Empirical results

All empirical results are tested on testing set.

(1) First experiment

The purpose of this experiment is to check the functions of $D_{minor}(\boldsymbol{x})$ and weight "a". Four curves of accuracy versus weight have been plotted in Fig.2. The four curves are obtained when the number of principal components "k" is 32, 64, 96 and 128, respectively. $\boldsymbol{\psi}_i$ and $\boldsymbol{u}_{i,j}$, $i=1,2,...,21$, $j=1,2,...,k$ are estimated by Maximum Likelihood estimation. $\theta_{i,j}, i=1,2,...,21, j=1,2,...,k$ are initialized based on eigenvalues of each category: $\theta^0_{i,j} = sqrt(\lambda_{i,j})$, $i=1,2,...,21$, $j=1,2,...,k$. The setting of a has been shown in Table 1.

Table 1 Setting of a

$a(0)$	$a(t)$	$a(19)$
0	$a(t) = a(t-1) + 0.05$	1

Table 2 Accuracy of variable k

K	32	64	96	128
a	0	0	0	0
Accuracy (%)	91.54	97.03	98.09	98.46
a*	0.60	0.40	0.45	0.45
Highest accuracy (%)	98.62	98.68	98.70	98.76

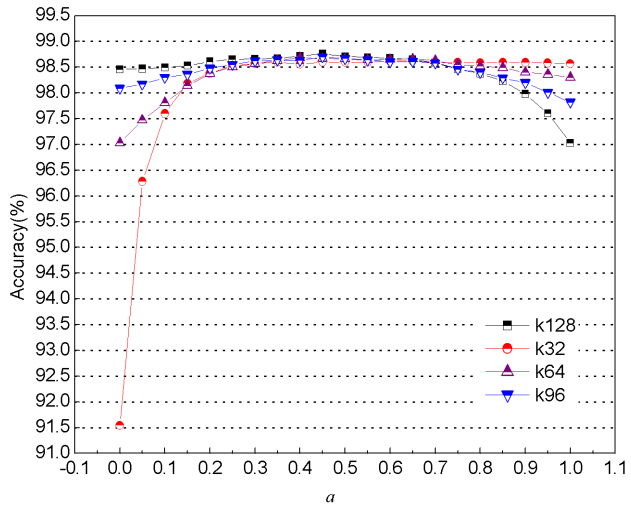

Fig.2 An illustration of accuracy versus a

a^* in Table 2 denotes the value corresponding to the highest accuracy under each number of the principal components. From Fig.2 and the testing results shown in Table 2, it can be seen that if a is set a proper value, the accuracy will be higher than that if a is set 0. Obviously, if a is set 0, it just means only $D_{pri}(x)$ is considered. Moreover, for improving recognition accuracy, the meaning of a proper setting of a will become more and more significant with the number of the principal components decreasing. For example, when the number of the principal components is 32, an increasing of 7.08 percent is obtained compared to the case that only $D_{pri}(x)$ is considered. However, when the number of the principal components is 128, only an increasing of 0.3 percent is obtained compared to the case that only $D_{pri}(x)$ is considered. The results demonstrate the important roles of $D_{\min or}(x)$ and weight "a", especially when the number of the principal components is small enough.

(2) Second experiment

The purpose of this experiment is to test the performance of θ learning algorithm based on MCE criterion.

ε_t is chosen as $\varepsilon_t = \varepsilon_0 \cdot \frac{1}{t}, t \geq 1$, which satisfies the requirements for ε_t, ε_0 is set 0.08, ζ and α are set 0.35 and 0. Iteration number is set 20; a is set a^* for each number of the principal components. The adjusting formulas have been provided in appendix.

Table 3 Accuracy of variable k after θ is learned

k	32	64	96	128
Accuracy (%)	98.80	98.83	98.84	98.86

It can be seen that for each number of the principal components, the accuracy has been improved after θ is learned. This result demonstrates the effectiveness of the learning algorithm of θ based on MCE criterion. Furthermore, the accuracy differences under variable numbers of the principal components have reduced after learning.

(3) Third experiment

The purpose of this experiment is to compare the performance of ADF with other recognition methods. Besides ADF, the performances of four recognition methods have been compared. They are Modified Quadratic Discriminant Functions (MQDF)[5], MultiLayer Perceptron (MLP)[12], Support Vector Machine (SVM)[3] and AHM[8].

To test the performance of AHM, some radicals are defined in advance. MQDF, MLP and SVM use the same features as ADF, and parameter setting of these methods is explained in Table 4. The accuracy of AHM is the lowest because of severe connections between strokes even radicals; the accuracy of ADF is only a bit lower than that of SVM in all tested methods but in computation efficiency and store cost aspects, ADF is superior to SVM. Considering a general performance of accuracy, computation efficiency and store cost, ADF is still fairly satisfied.

Table 4 Accuracy comparison of five recognition methods

Method	MQDF k=32	MLP (2-layer, 392 hidden units)	SVM (Polynomial kernel, 3 degrees)	AHM	ADF
Accuracy (%)	98.56	98.78	98.92	98.30	98.86

5 Conclusions

In this paper, ADF for handwriting recognition is proposed. First, APM in the principal subspace is presented and an optimal APM corresponding to an unknown pattern is obtained. Then, ADF that is a weighted summation of two distances is proposed. One is the distance between an unknown pattern and the principal subspace and the other is the distance between an unknown pattern and the minor subspace. Last, optimal constraints of APM are searched by applying MCE criterion. Experiments are conducted on 21 categories of the handwritten financial Chinese characters used in bank bill, and a good performance of ADF has been achieved.

As a matter of fact, there commonly exist nonlinear variations in handwriting. However, PCA is only effective to cope with linear variations, so future study will focus on how to use nonlinear PCA technique to obtain nonlinear ADF.

Acknowledgement The authors would like to thank Dr. Shi Daming at School of Computer Engineering, Nanyang Technological University for discussing some valuable issues about this paper.

References:

[1] Plamondon R, Srihari SN. On-Line and off-line handwriting recognition: a comprehensive survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(1):63−84.

[2] Liu CL, Nakashima K, Sako H, Hiromichi F. Handwritten digit recognition: Benchmarking of state of the art techniques. Pattern Recognition, 2003,36:2271−2285.

[3] Gao X, Jin LW, Yin JX, Huang JC. A new SVM-based handwritten Chinese character recognition. Chinese Journal of Electronics, 2002,30(5):651−654 (in Chinese with English abstract).

[4] Kim IJ, Kim JH. Statistical character structural modeling and its application to handwritten Chinese character recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(11):1422−1436.

[5] Kimura F, Takashina K, Tsuruoka S, *et al*. Modified quadratic discriminant functions and the applications to Chinese character recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1987,9(1):149−153.

[6] Shi D, Gunn SR, Damper RI. Handwritten Chinese radical recognition using nonlinear active shape models. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(2):277−280.

[7] Shi D, Gunn SR, Damper RI. Handwritten Chinese character recognition using nonlinear active shape models and viterbi algorithm. Pattern Recognition Letters, 2002,23:1853−1862.

[8] Shi D, Gunn SR, Damper RI. A radical approach to handwritten Chinese character recognition using active handwriting models. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. Kauai, Hawaii, 2001. 670−675.

[9] Tang YY, Tu LT, Liu JM, Lee SW, Lin WW, Shyu IS. Offline recognition of Chinese handwriting by multi-feature and multi-level classification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998,20(5):556−561.

[10] Kimura F, Wakabayashi T, Tsuruoka S, Miyake Y. Improvement of handwritten Japanese character recognition using weighted direction code histogram. Pattern Recognition, 1997,30(8):1329−1337.

[11] Kato N, Suzuki M, Omachi S, Aso H, Nemoto Y. A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1999,21(3):258−262.

[12] Duda R, Hart P, Stork DG. Pattern Classification. 2nd ed., New York: John Wiley and Sons Inc, 2001.

[13] Juang BH, Katagiri S. Discriminative learning for minimum error classification. IEEE Trans. on Signal Processing, 1992,40(12): 3043−3054.

[14] Katagiri S, Juang BH, Lee CH. Pattern recognition using a family of design algorithm based upon the generalized probabilistic descent method. Proceedings of the IEEE, 1998,86(11):2345−2372.

[15] Lee SW, Park JS. Nonlinear shape normalization methods for the recognition of large-set handwritten characters. Pattern Recognition, 1994,27(7):895−902.

[16] Wakabayashi T, Tsuruoka S, Kimura F, Miyake Y. On the size and variable transformation of feature vector for handwritten character recognition. Trans IEICE Japan, 1996,J76-D-II(12):2495−2503.

附中文参考文献:

[3] 高学,金连文,尹俊勋,黄建成.一种基于支持向量机的手写汉字识别方法.电子学报,2002,30(5):651−654.

Appendix:

$$(x \in \omega_q) = \text{true}, \quad o = \arg\min_{1 \le i \le C, i \ne q} (g_i^*(x; \Lambda_i)) \tag{A.1}$$

$$\sum_{i=1}^{C} \frac{\partial l_i^*(x_n; \Lambda_i, \Lambda_o) I(x_n \in \omega_i)}{\partial \theta_{l,j}} = \left. \frac{\partial l_q^*(x; \Lambda_q, \Lambda_o)}{\partial \theta_{l,j}} \right|_{l=q} = [\Lambda_q(x; \Lambda_q, \Lambda_o) \cdot (-1)] \cdot \frac{\partial g_l^*(x; \Lambda_l)}{\partial \theta_{l,j}}, \quad 1 \le j \le k \tag{A.2}$$

$$\sum_{i=1}^{C} \frac{\partial l_i^*(x_n; \Lambda_i, \Lambda_o) I(x_n \in \omega_i)}{\partial \theta_{l,j}} = \left. \frac{\partial l_q^*(x; \Lambda_q, \Lambda_o)}{\partial \theta_{l,j}} \right|_{l=o} = \Lambda_q(x; \Lambda_q, \Lambda_o) \cdot \frac{\partial g_l^*(x; \Lambda_l)}{\partial \theta_{l,j}}, \quad 1 \le j \le k \tag{A.3}$$

$$\Lambda_q(x; \Lambda_q, \Lambda_o) = \frac{-\zeta \cdot e^{\zeta \cdot [d_q^*(x; \Lambda_q, \Lambda_o) + \alpha]}}{\left(1 + e^{\zeta \cdot [d_q^*(x; \Lambda_q, \Lambda_o) + \alpha]}\right)^2} \tag{A.4}$$

$$\frac{\partial g_l^*(x; \Lambda_l)}{\partial \theta_{l,j}} = \begin{cases} a - 1, & \theta_{l,j} < |\vec{x}_{l,j}| \\ 0, & |\vec{x}_{l,j}| \le \theta_{l,j} \end{cases}, \quad l = q, o, \quad 1 \le j \le k \tag{A.5}$$

Transformation of θ

It has been proved that Λ_{tot} will converge to a local optimum with probability 1 without constraints on Λ_{tot}. However, in our problem, it is required $\theta_{i,j} > 0,\ 1 \le i \le C,\ 1 \le j \le k$. To tackle this problem, we introduce a new variable $\tilde{\theta}_{i,j}$ that has such a relationship with $\theta_{i,j}$ as follows:

$$\theta_{i,j} = \exp(\tilde{\theta}_{i,j}),\ 1 \le i \le C,\ 1 \le j \le k \tag{A.6}$$

Then parameter set Λ_i will become $\tilde{\Lambda}_i = \{\psi_i, u_{i,j}, \tilde{\theta}_{i,j}, k, 1 \le j \le k\}$. $\frac{\partial l_q^*(x; \tilde{\Lambda}_q, \tilde{\Lambda}_o)}{\partial \tilde{\theta}_{i,j}}$ is to be calculated as follows correspondingly:

$$\frac{\partial l_q^*(x; \tilde{\Lambda}_q, \tilde{\Lambda}_o)}{\partial \tilde{\theta}_{i,j}} = \frac{\partial l_q^*(x; \Lambda_q, \Lambda_o)}{\partial \theta_{i,j}} \cdot \exp(\tilde{\theta}_{i,j}) \tag{A.7}$$