A Novel Approach for Online Handwriting Recognition of Tibetan Characters

Jianjun Qian, Weilan Wang, Daohui Wang

***Abstract*—A new method is proposed for online handwriting recognition of Tibetan characters. At first, input pattern is preprocessing. Then, extracting direction feature matrix and edge feature matrix of Tibetan character respectively, they are together formed original feature matrix. It is compressed into final feature matrix with IMLDA (image matrix liner discriminate analysis) technique. Finally, the pattern of character is recognized by the SMQDF (second modified quadratic discriminate function) classifier based on prior training. The effectiveness of the proposed method was visualized by the experiments. The overall recognition rate is 93.72 percent, which confirms the effectiveness of the proposed methods.**

***Index Terms*— IMLDA, Online handwriting recognition of Tibetan characters, SMQDF classifier, Transformation matrix.**

I. Introduction

Handwriting recognition is the process that transfers writer's handwriting from dimensional expression to semantic signal expression. Compared with the offline handwriting recognition, online handwriting recognition requires that handwriting process and conversion is simultaneously. The online handwriting recognition has great academic and practical value because of it is more suitable for the natural handwriting, therefore, it is got more comprehensive attention from the researcher and developed pretty quickly [1, 2]. In recent years, Thierry Artieres etc have presented a hierarchical HMM approach to online handwriting shape recognition [3]; In-Jung Kim etc proposed systematic character structure modeling method for online handwriting Chinese characters recognition [4]; In addition, support vector machine method has already been successfully applied to online handwriting recognition [5].

The methods of online handwriting recognition can be roughly divided into three categories: statistical methods, structure methods and hybrid statistical-structure methods. It can descript two dimension spatial structure of character with strong ability based on the structure methods, but it also has its own weakness, for example, weak anti-interference ability and unstable. It is difficult to adapt to variability of characters. On the contrary, it has stronger robustness and higher stability based on the statistical method. However, curve feature matching, Markov model, time delay neural network and so on. All of them have their shortness. For example, because of large amount of computation of curve feature matching method, it is not suitable for large character set recognition. HMM (Hidden Markov Model) is the most popular techniques for speech recognition and, has been successfully applied to online handwriting recognition [6]. Although it needs less parameter, it can't solve the dynamic length sequence because of fixed time window, based on the back-propagation algorithm of neural network.

Tibetan is one of the most excellent characters in the world. In order to use Tibetan in computer, the key problem is how to effectively enter Tibetan into the computer. In the past few years, we have successfully developed keyboard Tibetan intelligent input system [7, 8]. The research of printed Tibetan recognition has achieved satisfactory result and completed Tibetan Printed OCR system [9, 10]. Meanwhile, we also received some preliminary results on online handwriting Tibetan characters recognition [11, 12]. Here, we explore a new approach to training and recognition based on SMQDF classifier. Experiment result showing positive effect.

The rest of this paper is organized as follows: Section 2 describes feature extraction. The dimensionality reduction of feature matrix is expressed in section 3. The training of classifier is introduced in section 4. In section 5, an online handwriting Tibetan character recognition system is introduced. Section 6 provides experiment results. Finally, the conclusions are drawn in section 7.

II. Feature Extraction

Better features should not only have high stability and be undisturbed of noise, but also it can adapt to different font. Meanwhile, what features we choose should have strong ability of distinction, which can make the distance during different classes as long as possible and similar samples can gather together.

On the same principle as above, preprocessing [11] is the first step. Then, character image is evenly divided into 36 sub-blocks according to structure feature of Tibetan character. And extracting direction features of each sub-block's pen-trajectory points [13]. Meanwhile, it is also evenly divided into 24 sub-blocks according to edge feature of Tibetan character. Extracting edge features of each

Manuscript received December 8, 2009. This work was supported in part by The Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Personnel of People's Republic of China, and the Tackle Key Problems in Science and Technology Foundation of Gansu Province under Grant (NO: 2GS064-A52–035-04)

Jianjun Qian was with Information Technology Institute, Northwest University for Nationalities, Lanzhou, Gansu, Privince,730030 China. (e-mail: qjjtx@126.com)

Weilan Wang was with School of Computer Science and Information Engineering, Northwest University for Nationalities, Lanzhou, Gansu Privince,730030 China; (corresponding author to provide e-mail: weilanchina@gmail.com)

Daohui Wang was with Information Technology Institute, Northwest University for Nationalities, Lanzhou, Gansu, Privince,730030 China. e-mail: (crystal008@163.com)

sub-block's pen-trajectory points [13]. Original Tibetan character feature is composed of direction feature and edge feature.

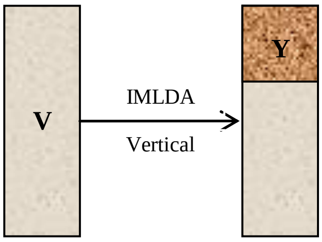

Fig 1 Illustration of IMLDA transform

III. FEATURE TRANSFORMATION

If the original feature matrix is used to train and classify directly, it will increase computation complexity and reduce classifier performance, when samples are not enough. Therefore, we must compress high dimension feature matrix into low one. This paper applies IMLDA technology in vertical direction to complete feature transformation [14, 15]. Assuming that $\{\{V_i^{(j)}, 1 \le i \le N_j\}, 1 \le j \le C\}$ is original feature set, where $V_i^{(j)}$ represents ith sample's original feature matrix of jth class, N_j is the sample number of jth class, C is class number. Each class represents a character of modern Tibetan (not include Sanskrit) set. Each class's mean value and all classes' value will be calculated by formula (1) and (2), respectively.

$$\mu_j = \frac{1}{N_j}\sum_{i=1}^{N_j} V_i^{(j)} \tag{1}$$

$$\mu = \frac{1}{C}\sum_{j=1}^{C} \mu_j \tag{2}$$

Then, the within-class scatter matrix S_w and between-class scatter matrix S_b can be calculated:

$$S_w = \frac{1}{C}\sum_{j=1}^{C}\left(\frac{1}{N}\sum_{i=1}^{N_j}(V_i^{(j)} - \mu_j)(V_i^{(j)} - \mu_j)^T\right) \tag{3}$$

$$S_b = \frac{1}{C}\sum_{j=1}^{C}(\mu_j - \mu)(\mu_j - \mu)^T \tag{4}$$

Here, we choose $\left|(S_b + S_w)/S_w\right|$ as optimization criterion, namely find liner transformation matrix A, which can make formula $\left|\frac{A^T(S_b+S_w)A}{A^T S_w A}\right|$ get the max value. A is $n \times m$ matrix, n is line number of original feature matrix, meanwhile, assuming m is line number of feature matrix after transformation. We can use the following method to achieve transformation matrix: firstly, computing eigenvalue and eigenvector of $S_w^{-1}(S_b + S_w)$; secondly, eigenvalues are ordered by descending, eigenvectors' order also make corresponding changes; finally, the top m eigenvector form liner transformation matrix $A = [\xi_1, \xi_2, \ldots, \xi_m]$. Here, m is 12.

Feature transformation formula is:

$$Y = A^T \cdot V \tag{5}$$

Where, V is original feature matrix, Y is final feature matrix. The process is illustrated in Fig 1.

IV. TRAINING CLASSIFIER

A. *MQDF Classifier*

MQDF classifier's discriminate function [16, 17]:

$$g_j(Y) = \sum_{i=1}^{k}\frac{((Y-\mu_j)^T \zeta_i^{(j)})^2}{\lambda_i^{(j)}} + \sum_{i=k+1}^{m}\frac{((Y-\mu_j)^T \zeta_i^{(j)})^2}{\lambda} + \sum_{i=1}^{k}\log\lambda_i^{(j)} + \sum_{i=k+1}^{m}\log\lambda \tag{6}$$

$$j = 1.2......, C$$

Where, Y is input feature vector, m is line number of feature matrix, k is constant integer less than m, k and λ, that are experience parameters, are decided by experiment. μ_j represents the jth class's mean vector, $\zeta_i^{(j)}$ is the ith eigenvector of the jth class's covariance matrix, $\lambda_i^{(j)}$ is the ith eigenvalue of the jth class's covariance matrix. We can obtain the classified result based on the following criterion:

If $g_i(Y) = \min_{1 \le j \le C} g_j(Y)$ (C is class number), then we believe that input pattern Y belongs to the ith class.

B. *SMQDF Classifier*

MQDF classifier is widely used in the area of character recognition. However, it only applies to feature vector, and it is not appropriate for feature matrix. For this reason, we proposed SMQDF classifier in this paper, its discriminate function as shown in the follow formula:

$$g_j(Y) = \sum_{i=1}^{k}\frac{((Y-\mu_j)^T \zeta_i^{(j)})^T((Y-\mu_j)^T \zeta_i^{(j)})}{\lambda_i^{(j)}} + \sum_{i=k+1}^{m}\frac{((Y-\mu_j)^T \zeta_i^{(j)})^T((Y-\mu_j)^T \zeta_i^{(j)})}{\lambda} + \sum_{i=1}^{k}\log\lambda_i^{(j)} + \sum_{i=k+1}^{m}\log\lambda \tag{7}$$

$$j = 1.2......, C$$

Where, Y is feature matrix, m is positive integer, λ is experience parameter. When classifies, Y belongs to the class whose $g_i(y)$ is minimum.

C. *Generating Recognition Base*

Computing mean value and covariance matrix of each class by the following formulas:

$$\mu_j = \frac{1}{N_j}\sum_{i=1}^{N_j} Y_i^{(j)}, \tag{8}$$

$$\Sigma_j = \frac{1}{N_j}\sum_{i=1}^{N_j}(Y_i^{(j)} - \mu_j)(Y_i^{(j)} - \mu_j)^T) \tag{9}$$

Where, $Y_i^{(j)}$ represents the ith sample's feature matrix of the jth class, N_j is the number of the jth class, μ_j is the jth class's mean value, $\sum_j$ represents the jth class's covariance matrix.

The eigenvalue and eigenvector of each class's covariance are calculated, then, eigenvalues are ordered by descending, eigenvectors' order also makes corresponding changes.

$\lambda_i^{(j)}$ is the ith eigenvalue of $\sum_j$, $\zeta_i^{(j)}$ is the ith eigenvector of $\sum_j$

We use the formula (10) to compute the parameter λ of SMQDF classifier, namely small feature substitution value.

$$\lambda = \frac{1}{C}\sum_{j=1}^{C}\lambda_m^{(j)} \tag{10}$$

Saving all the parameters $\lambda_i^{(j)}$, $j=1,2,\cdots,C$, $i=1,2,\cdots,k$ 、 $\zeta_i^{(j)}, j=1,2,\cdots,C, i=1,2,\cdots,m$ 、 $\mu_j, j=1,2,\cdots,C,\lambda$ to recognition base. Training phase is over

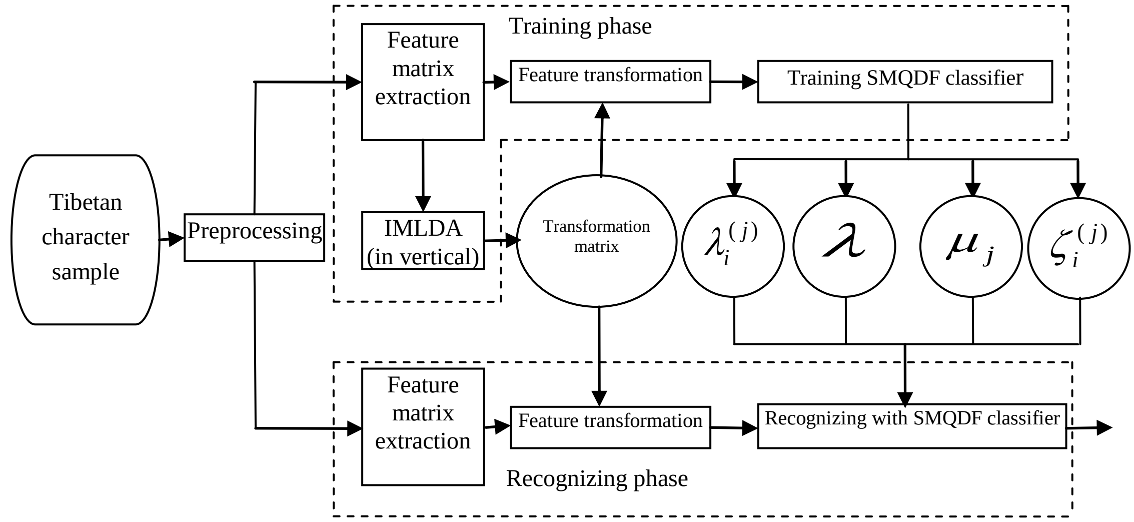

Fig 2. Recognition system frame

.

Table 1.Radical characters' recognition rate

C	Top 1	Top 3	Top 5	Top 10	C	Top 1	Top 3	Top 5	Top 10
ཀ	67.18%	91.44%	93.51%	96.72%	ཁ	78.68%	86.89%	91.80%	95.08%
ག	81.96%	91.80%	91.80%	95.08%	ང	75.13%	86.61%	86.61%	90.16%
ཅ	62.45%	70.49%	76.08%	88.52%	ཆ	91.80%	96.72%	96.72%	96.72%
ཇ	63.97%	85.39%	88.21%	93.44%	ཉ	77.05%	90.16%	93.44%	96.72%
ཏ	65.49%	85.21%	89.44%	95.08%	ཐ	65.68%	84.01%	90.16%	95.08%
ད	90.16%	91.80%	93.44%	93.44%	ན	68.30%	83.80%	88.85%	91.56%
པ	68.85%	85.25%	90.16%	93.44%	ཕ	62.29%	77.04%	91.80%	95.08%
བ	63.93%	77.05%	81.96%	85.25%	མ	79.87%	89.73%	91.14%	93.66%
ཙ	66.38%	83.28%	88.52%	91.25%	ཚ	78.68%	93.44%	93.44%	95.08%
ཛ	65.52%	80.33%	86.89%	93.44%	ཝ	40.32%	75.41%	88.52%	95.08%
ཞ	83.61%	91.80%	95.08%	95.08%	ཟ	72.13%	86.88%	88.52%	90.16%
འ	78.87%	88.03%	90.24%	91.25%	ཡ	75.40%	86.89%	93.44%	96.72%
ར	70.49%	85.25%	93.44%	98.36%	ལ	60.66%	88.52%	93.44%	93.44%
ཤ	72.13%	85.24%	86.88%	90.16%	ས	68.31%	78.69%	86.89%	91.80%
ཧ	65.68%	78.68%	81.97%	91.80%	ཨ	80.33%	93.44%	93.44%	96.72%

V. RECOGNITION SYSTEM

Recognition process as shown in Fig 2, it is similar with training process. Thus, it also needs preprocessing, then extracting original feature matrix V . In feature transformation phase, we can compute final feature matrix Y with transformation matrix A (As shown in formula (5)), which is generated in training process phase. All the related parameters of classifier were supported by recognition base, when recognizing with SMQDF classifier.

Finally, we only need compute $g_j(Y)$ of each class with formula (7).Y belongs to the ith class when $g_i(Y)$ is equal to $\min\limits_{1\le j\le C} g_j(Y)$.

VI. EXPERIMENT RESULT

In this system, character set, containing 562 characters, are frequently used characters of modern Tibetan. There are 203 sets samples total, training samples are 142 sets and testing samples are 61 sets. The average recognition rates achieve 93.72% for Top 10, 55.56% for Top 1, with online handwriting Tibetan character recognition method this paper proposed. We will analyses all those recognition results in detail in this section.

A. The results of radical characters recognition

The recognition rate of 30 radical characters are shown in Table 1, the columns are C (character), Top 1, Top 3, Top 5 and Top 10, respectively. For Top 10, the maximal recognition rate achieve 98.36%, the minimum recognition rate is 85.25%, and average recognition rate reach 93.51%.

B. Best recognized case and worst recognized case

The best recognized cases in our test achieve high recognition rate of 97% to 98%, those characters include several radicals and several characters without vowel. And ten worst recognized cases are shown in Table 2, we can find those characters with vowel that is above vowel or below vowel. The top six lists for the table have above vowel and followed four lists have below vowel.

Table 2. Ten worst recognized cases

C	Top 1	Top 3	Top 5	Top 10
[illegible]	36.07%	50.82%	59.02%	65.57%
[illegible]	18.03%	44.26%	57.38%	68.85%
[illegible]	26.23%	42.62%	44.26%	70.49%
[illegible]	24.59%	36.07%	59.02%	70.49%
[illegible]	11.48%	32.79%	54.98%	70.49%
[illegible]	24.59%	42.62%	55.74%	70.49%
[illegible]	22.73%	45.46%	54.55%	70.49%
[illegible]	26.23%	52.46%	57.38%	73.77%
[illegible]	39.34%	60.66%	62.30%	73.77%
[illegible]	18.03%	42.62%	59.02%	73.77%

C. Recognition rate of character without vowel

There are 151 characters without vowel in 562 character sets. the recognition rate of top 1, top 3, top 5, top 10 are 68.63%, 84.52%, 90.55% and 95.01% respectively, as Table 3 shown. Compare with total character recognition, that character's recognition rate is better obviously.

Table 3. Recognition rate of character without vowel

	Top 1	Top 3	Top 5	Top 10
Recognition rates	68.63%	84.52%	90.55%	95.01%

D. Recognition rate contrast on the character without above vowel and with above vowel

Experimentation finds that the character's recognition rates have bigger difference for without above vowel and with above vowel. Table 4 shows some examples, three columns of 1, 2 and 3, list characters (C) and their recognition rates (Top 10) respectively. Those illustrations the recognition rates are high for the character without vowel, once the character take a vowel, its recognition rate descend immediately. Actually, the distinction of a lot of characters is whether with vowel or which vowel. We only list 7 teams in Tibetan character set, examples of this sort there are 103 teams.

Table 4. Contrast character recognition rate without vowel and with vowel

2		1		3	
C	Top 10	C	Top 10	C	Top 10
སྒི	85.25%	སྒ	98.36%	སྒེ	80.33%
ཁི	93.44%	ཁ	95.08%	ཁེ	88.52%
དི	81.97%	ད	95.08%	དེ	80.33%
སྡི	77.23%	སྡ	86.89%	སྡེ	78.69%
ཀིུ	86.36%	ཀུ	95.08%	ཀེུ	78.69%
ཟླི	90.91%	ཟླ	95.08%	ཟླེ	86.36%
སྟི	86.36%	སྟ	93.44%	སྟེ	81.82%

E. Confusable character recognition

Table 5. Part confusable characters

Order	C	Top 10	Confusable characters
1	སྒི	85.25%	[illegible]…
2	སྒེ	80.33%	[illegible]…
3	སྡི	75.41%	[illegible]…
4	སྡེ	78.69%	[illegible]…
5	དི	81.97%	[illegible]…
6	དེ	80.33%	[illegible]…
7	སྟི	73.77%	[illegible]…
8	སྟེ	78.69%	[illegible]…
9	ཀིུ	86.36%	[illegible]…
10	ཀེུ	78.69%	[illegible]…

Commonly, in our system, low-recognition rate characters are essentially with vowel. Table 5 shows part such characters; four columns indicate order, characters (C), Top 10 and confusable characters. There are more than 20 characters for each character's confusable characters, Table 5 only list 9 confusable characters by similar degree from high to low. Pay attention to the characters of order 1 and 2, 3 and 4, 5 and 6, 7 and 8, 9 and 10, each pair characters only have different vowel, this two vowels have small difference, and there is the other in they each other's confusable characters (indicate by italic text and bold). On the other hand, due to the similar of radical characters form some characters' similar. Such as "ྒ" "ྦ" , the difference is very small, make"སྒི" and "སྦི" very similar; similar characters: “ད” “ཧ” “ཏ” ,lead very similar characters: “ལྡི” “ལྷི” and “ལྷེ”; and similar characters:“ལ” “ས”, lead very similar characters: “ལྡི” and “སྡི”, etc. “ཀྱུ” and “ཀྱེ” only is vowel’s similar make they similar in order 9 and order 10, of course, like “ག” and “ཀ” similar results in “གྱུ” and “ཀྱུ”, “གྱེ” and “ཀྱེ” similar, and one is similar to other, etc. Much structurally similar character exists in Tibetan character set, so that makes online handwriting Tibetan character recognition very difficult.

VII. CONCLUSIONS

In this paper, we proposed a novel approach for online handwriting Tibetan character recognition. Feature extraction is the first step, we can obtain original Tibetan character feature matrix through preprocessing, dividing character into several blocks and extracting their direction features and edge features. The second step, we complete feature transformation with IMLDA in vertical direction. At last, we can apply SMQDF classifier to train and recognize.

The overall recognition rate is 93.72 percent. However, it also exists some problems. First, total recognition rate is not high; second, the first character recognition rate is relative lower; third, it is not ideal of similar Tibetan character recognition result. All these remain challenging tasks for future.

REFERENCES

[1] Plamondon R, Srihari S N Online and offline handwriting recognition: a comp rehensive survey[J]. IEEE Transactions on Pattern Analysis andMachine Intelligence, 2000, 22 (1) : 63- 84.

[2] Cheng-Lin Liu, Stafan Jaeger and Masaki Nakagawa. Online Recognition of Chinese Characters: The State-of-the-Art. IEEE Transactions on Pattern Analysis and Machine Intelligence, February 2004. 26(2): 198-213

[3] Thierry Artieres, Sanparith Marukatat, and Patrick Gallinari. Online Handwriting Shape Recognition Using Segmental Hidden Markov Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, February 2007. 29(2): 299-310.

[4] In-Jung Kim ,Jin-Hyung Kim. Statistical Character Structure Modeling and Its Application to Handwritten Chinese Character Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, November 2003.25(11):1422-1436

[5] C. Bahlmann, B. Haasdonk, and H. Burkhardt. On-line handwriting recognition with support vector machines a kernel approach. In Int. Workshop on Frontiers in Handwriting Recognition (IWFHR), Niagara-on-the-Lake, August 2002.

[6] J. Hu, M. K. Brown, and W. Turin. “HMM Based On-Line Handwriting Recognition”. IEEE Trans. Pattern Analysis and Machine Intelligence, Oct. 1996, 18(10): 1039-1045.

[7] Weilan Wang. Intelligent Input Software of Tibetan. Computer Standards & Interfaces. 2007 (29),pp462–466

[8] Weilan Wang, Lingwang Kun. A Fast Input Method for Tibetan Based on Word in Unicode. International MultiConference of Engineers and Computer Scientists 2008 Hong Kong, 19-21 March.pp374-377.

[9] Ding Xiaoqing, Wang Hua. Multi-Font Printed Tibetan Character Recognition. JOURNAL OF CHINESE INFORMATION PROCESSING. Year:2003 Issue:06 Volume: 17

[10] Wang Weilan，Ding Xiaoqing Chen Li, Wang Hua. Study on Printed Tibetn Character Recognition. Computer Engineering. 2003. Vol. 29. pp37-38,94

[11] Liu Hongyi, Wang Weilan. Nonlinear Shape Normalization Methods for On-line Recognition of Handwritten Tibetan Characters (In Chinese). Application Research of Computers. 2006(9).pp179-181.

[12] Wang Weilan, Chen Wan-jun. MCLRNN Model for Online Handwritten Tibetan Character Recognition Based on Stroke Characteristics (In Chinese). Computer Engineering and Applications. 2008.Vol.26, No.14.pp 91-93,194

[13] T rier O D, J ain A K, Taxt T. Feature extraction methods for character recognition-a survey [J]. Pattern Recognition, 1996, 29 (4) : 641 662.

[14] J. Yang, J.-y. Yang, A.F. Frangi, D. Zhang, Uncorrelated projection discriminant analysis and its application to face image feature extraction, International Journal Pattern Recognition Artificial Intelligence. 17 (8) (2003) 1325–1347

[15] J. Yang, D. Zhang, Y. Xu, J.-y. Yang, Two-dimensional discriminant transform of face recognition [J]. Pattern Recognition, 2005, 38, 1125-1129.

[16] Kimura F, Shridhar M. Handwritten numerical recognition based on multiple algorithms [J]. Pattern Recognition,1991, 24 (10) : 969-983

[17] Kimura F, Takash ina K, T suruoka S, et al. Modified quadratic discriminant functions and its application to Chinese character recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1987, 9 (1) : 149 - 153.