Peripheral Contour Feature Based On-line Handwritten Uyghur Character Recognition

Zulpiya KAHAR[1], Mayire IBRAYIM[2], Dilmurat TURSUN[3] and Askar HAMDULLA[4,*]

[1] Institute of Information Science and Engineering, Xinjiang University
Xinjiang, 830046, P.R. China

[2] School of Electronics Information, Wuhan University
Wuhan, 430072, P.R. China

[3] Institute of Information Science and Engineering, Xinjiang University
Xinjiang, 830046, P.R. China

[4] College of Software, Xinjiang University
Xinjiang, 830046, P.R. China

Abstract

In this paper, we conduct a deep research on the grid direction feature, peripheral contour feature, stroke number feature and additional part's location feature for recognition of 32 independent forms of Uyghur characters through fully considering the unique shapes and writing styles of Uyghur characters. We divide the whole characters to subclasses by applying stroke number feature and additional part's location feature of individual characters, and then conducted several recognition experiments using the above features alone and various combinations. The experimental results show that peripheral contour feature is the best choice. The design of the character recognition in this article is an important part of the string recognition / word recognition, and laid the foundation for the implementation of Uyghur character handwritten input method in various platforms.

Keywords: *Uyghur Characters, On-line Handwriting, Character Recognition, Peripheral Contour Feature, Grid Direction Feature.*

1. Introduction

Handwritten character recognition is not only an important branch of pattern recognition, but also a comprehensive technology. In recent years, on-line handwriting has been highly valued and widely used in daily life as a natural and convenient input method. On-line handwritten recognition technologies for ethnic group languages of China have become the focus of research in recent years. Tsinghua University and Xinjiang University are representative researchers [1] among them. Due to accuracy of Uyghur script recognition suffers from various factors such input tools, unique structures, many personalized writing styles etc., so it is a thorny problem. As a kind of popular language in minority areas of west china, research on Uyghur handwriting recognition method is very helpful to promote the speed of informatization and development of science at those areas.

Handwritten character recognition technology can be divided into off-line and on-line [2]. Off-line handwritten recognition no need to conduct real-time interaction with users; On-line handwritten recognition technology get the handwriting path information through the writing pad and trajectory capture device, convert the handwriting scripture into text that it is recognizable to the computers [3]. On-line handwritten character recognition method can be divided into two categories, one is based on the whole word recognition method, and another is based on segmentation method. Most online handwriting recognition are all based on the latter method, due to the fact that character is easy to identify, easy to classify, therefore, the word recognition problem is also easy to solve. In general, character recognition process is divided into handwritten character data acquisition and pre-processing, feature extraction, feature classification and post-processing steps. The key steps in character recognition are the normalization method in pre-processing; the feature extraction method and the classifier [4].Therefore, find the appropriate normalization and feature extraction method and effective classifier are very important issues.

Uyghur belongs to the Altaic Turkic language, and the writing style is very different from Chinese and English characters. Character is the basic component of Uyghur word structure, consists of one main stroke and one or more additional strokes [5].When writing, write the main stroke first and additional parts last[6]. Additional part includes different number and position of point and ا , ۋ , ۇ , ې four basic shapes. Structure of Uyghur character is simple, writing is random, much deformation, easy to produce similar characters, these characteristics bring a lot of difficulties to Uyghur word recognition. Reference [7] applies different structure and statistical features to 128 different writing forms of 32 characters, get better recognition result. This paper carries analysis on 32 independent forms of Uyghur characters, study a method based on peripheral contour feature. After verification by experiment, this method obviously improves Uyghur handwritten character recognition accuracy.

2. On-line Handwritten Character Recognition System

Structure of on-line handwritten character recognition system is as shown in Fig.1. The whole recognition system consists of five function module, they are: pre-processing, feature extraction, establishment of template library, classification, recognition results output. Main work of training part is to study the training sample set, and then get template library; Main work of testing part is to identify the characters to be inputted to identification, and output the results. Below are the details about each functional module.

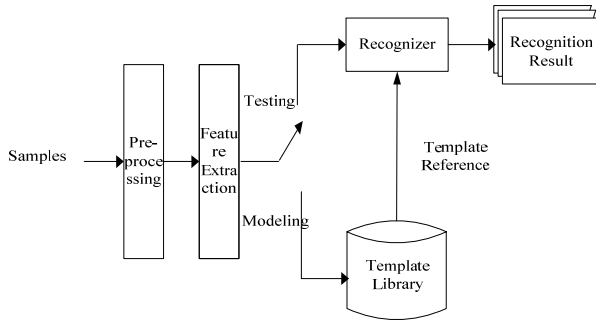


Fig.1 Block diagram of on-line handwritten character recognition system

2.1 Data Acquisition and Pre-processing

First step in on-line handwritten character recognition is data collection, to obtain the raw materials. When it is writing, the pen's trajectory is recorded. Pen trajectory is usually described by its board of (x, y) coordinates and the states of pen's up and down movement. Pen up state divides the writing trajectory into the strokes that the trajectory is the length from the state of pen down to the state of pen up. Therefore, a script of Uyghur text is expressed by its set of strokes, and each stroke is shown by its point sequence. Stroke points depend on the stroke length and sampling frequency, each point coordinates (x, y) value relies on the position of pen points on the pad and the pad resolution ratio.

On-line handwritten recognition pre-processing procedure includes two steps: noise elimination and standardization [8]. In order to convenient for feature extraction, reduce the effects of character deformation and distortion brought by normalization procedure, and to improve the reliability of the character recognition, this paper applies the linear normalization and non-linear normalization combined method based on point density. That is to find out the external rectangular of character point sequence, if external rectangular is not square, so to be going through linear normalization for character first, and then conduct non-linear normalization based on point density. If the external rectangular is square, so conduct linear normalization only. This paper considers the capriciousness of handwritten characters, and all the characters are normalized to 96 * 96 lattices.

2.2 Feature Extraction Algorithm

Feature extraction is one of the cores of pattern recognition, and feature selection is the key to pattern recognition. However, there is lack of effective feature extraction method. This paper considers Uyghur characters' characteristics and writing rules, according to the principle of feature extraction, choose the peripheral contour feature, grid direction feature, stroke number feature and additional part's location feature, and find out the feature extraction method for each one. Below the three features and their feature extraction methods are introduced respectively.

2.2.1 Peripheral Contour Feature

The grid feature is the combination of structural feature and the statistical feature. Character image is evenly or unevenly divided into several areas, called the "grid". Various features searched within each grid, statistics of features conducted by the grid unit, to enhance anti-jamming of features. This paper extracts the peripheral contour feature, and uses the center of gravity division method. Division method is as follows:

First of all, from every character point sequence calculating barycentric coordinates. Center of gravity ($CentX$, $CentY$) will divide the character's external rectangular shapes into four parts, see Fig.2 (a), then the four parts are separately divided into eight parts as a sub-graph; see Fig.2 (b), so one character is divided into 32 grids.

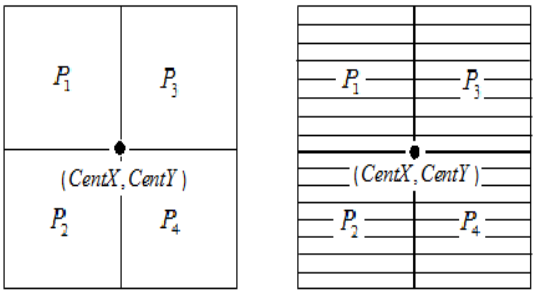


Fig.2 Grid partition chart a) Grid center of gravity b) Grid partition

According to the above center of gravity grid partition method, we will divide the character into grids as shown in Fig.3, recorded as $S_L^1, S_L^2,..., S_L^{16}; S_R^1, S_R^2,..., S_R^{16}; H_L, H_R$ in the diagram indicate that the width of the left and right i grid.

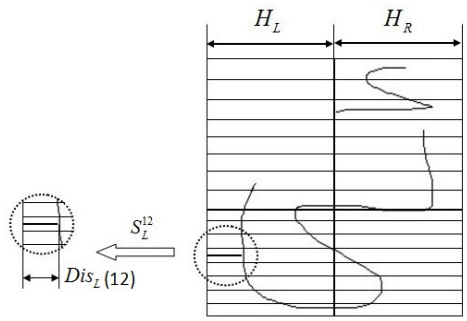


Fig.3 Peripheral contour feature extraction

In each block, we conduct transverse projection for character pen section, and get $Dis_L(12)$ depicted in S_L^{12}, the distance from character rectangular edge to periphery of character. So this paper gets 32 dimensional peripheral contour features.

2.2.2 Stroke Number Feature and Additional Part's Location Feature

Stroke number feature refers to the number of strokes that form the handwritten character trajectory. Stroke number feature is recorded by the number of pen up or loose of mouse button. Such as, stroke number feature of character ﺍ is 2.

In Uygur characters, in addition to some of the characters only consisted of main body parts, most of the characters consisted of main body part and additional parts, and there are three positional relationship between main body part and additional part, namely additional part is on the top of the main body part, additional part is below of the main body part, and additional part is in the middle of the main body part. In this paper, using 0 means no additional parts, 1 represents additional part is on the above of the main body part, use 2 represents additional part is on below of the main body part, with 3 represents additional part is in the middle of the main body part. According to Uyghur writing rules, write the main body part of the character first, and then write the additional part. So the positional relationship between main body part and additional part is actually the relationship between first stroke and other strokes of one character. This paper extracts additional part's location feature through compare the top and the bottom of first stroke with other strokes.

2.2.3 Grid Direction Feature

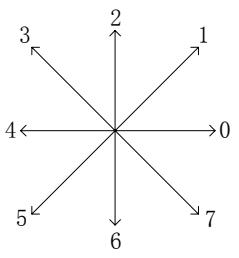
Divide each normalized character lattice (96 * 96 lattices) into16 grids. Independently calculate the features in each grid sector. All pen section in one grid to be quantified as the up, down, right, left and two diagonal eight directions, and compute pen length in each direction, thus forms a 128 (16 * 8) dimensional feature vector, as shown in Fig.4.

Block direction feature extraction method is as follows:

Step1: 96 * 96 normalized character is divided into 16 blocks in 24 * 24 sizes.

Step2: For handwriting point in each block extract eight direction features.

Step3: Calculate the handwriting point total in each block for each direction, and consider the total point as the feature of this block.



0:
0 0 6 1 0 0 0 0 0 0 1 1 0 0 6 0 0 0
2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 5
5 0 0 0 0 0 0 0 3 7 1 0 0 0 1 3 0 0
0 0 1 0 0 0 3 7 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 3 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 6 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0

Fig.4 Eight direction decomposition and directional feature of letter ﺏ

2.3Building the Template Library and Classification

There are the most popular two methods to build the template library in on-line handwritten character recognition: first one is to build one template for each character, second is to use preliminary classification algorithm separate into several sub-classes according to each character writing form, and then build a template for each sub-class, that is to say, to build multiple templates for each character. This paper used the two methods for test. The first kind is the traditional method that is to use a peripheral contour feature on each character to build one template; the second is the preliminary classification method which this paper based on. In the first step, the characters are divided into subclasses according to number of stroke feature and additional part's location feature of a character. Then for each sub-class build one template according to peripheral contour feature of characters within the sub-class. In this paper we take the average value of feature vectors (peripheral contour feature or grid direction feature) from each sub-class that participates in training as sub-class feature vector. Each template contains the sub-class category (alpha code), number of stroke feature, additional part's location feature, and the information such as representative feature vector.

After selecting effective feature extraction method, find out the appropriate classification method is very important. The traditional method is to use extracted character feature as input feature, and then calculate the distance with each template, finally output the recognition result. Flow chart of traditional classification method is as shown in Fig.5.

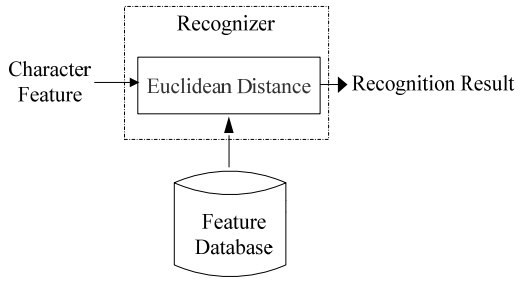


Fig.5 Traditional classification algorithm flow chart

This paper reduces the dimension of feature space by using number of stroke feature and additional part's location feature. From the high capacity feature library, we screen out the characters that have the same features which are number of stroke feature and additional part's location feature, so as to get low capacity feature database. We get the recognition results through calculating the distance between the features (including high dimensional features such as peripheral counter feature and grid direction feature) and each template with lower capacity feature database. This kind of quick search method based on the preliminary classification [9] scheme is as shown in Fig.6.

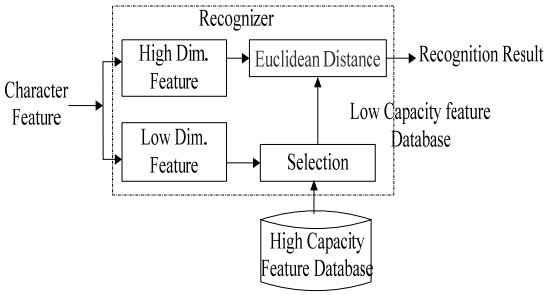


Fig.6 Quick search and comparing scheme

This paper compares number of stroke feature and additional part's location feature of each sample with each template respectively. If we get the similarity between the features of sample and template, then we take comparison of the high dimensional feature vector with the template representative feature vector. That is according to the minimum distance classifier calculate Euclidean distance, and output the alpha code in template which has the minimum distance with sample as recognition result.

Minimum distance classifier is to find the distance from unknown class vector to the center of training set all class representative feature vector, a kind of image classification method that belongs the unknown class vector to a class which have minimum distance [10].Its realization thought is: assume c class representative mode feature vector expressed with $R_1, \cdots, R_c$; among $R_i = \{R_{i1}, R_{i2}, \cdots, R_{id}\}$, d is feature vector dimension. Be identified pattern feature vector expressed with $X = \{x_1, x_2, \cdots, x_d\}$, $d(X, R_i)$ is the distance between X and $R_i (i = 1,2,\cdots,c)$ .if $d(X, R_i)$ has a minimum value, then X belongs to class i. This paper used Euclidean distance measurement method to calculate the similarity between samples. Euclidian distance between feature vector of to identify sample X with feature vector of template R_i is:

$$d(X, R_i) = \sqrt{\sum_{j=1}^{d} (x_j - R_{ij})^2} \quad (1)$$

X_j , R_{ij} is j component of corresponding feature vector.

3. Experimental Results and Analysis

This experiment collects Uyghur handwriting character samples with writing pad. Handwriting process requires each person to write as usual as own way. We collect 400 different people's 32 character samples; a total of 400 * 32 = 12800 samples, Fig.7 is for part of samples. We use the first 280 people's 8960 writing sample as training sample, and 120 people's 3840 writing sample as test sample.

Fig.7 Example of on-line handwritten Uyghur character samples

This paper studies peripheral contour feature, divide the characters into 16 * 2 pieces, get 32 dimensional distance feature. Make the statistics of character stroke number, make judgment to the location of additional part of the character, and get 2 dimensional features. Put the character into 16 grids, for each grid point conduct directional decomposing, and get 128 dimensional directional features, and then applies minimum distance classifier for test. This paper conduct test evaluation for peripheral contour feature, stroke number feature, additional part's location feature, grid direction feature and their different combinations respectively. We take the stroke number feature and additional part's location feature as low dimensional feature in test. Table 1 gives test accuracy of different combinations of features and different classification scheme, peripheral counter feature is expressed with P, grid direction feature is expressed with G in table.

Table 1: Recognition accuracy of different combinations of features and different classification scheme (%)

Different scheme / Character writing	Classification by traditional method			Classification by this paper proposed method		
	P	G	P + G	Low Dim. feature + P	Low Dim. feature + G	Low Dim. feature + P + G
ب	86.7	37.5	82.5	87.5	49.2	76.7
ت	60.8	30.8	51.7	69.2	61.7	61.7
ج	37.5	27.5	10.8	60.8	44.2	25.8
خ	78.3	24.2	28.3	79.2	22.5	57.5
د	35.8	60.8	50.8	57.5	60.8	52.5
ر	48.3	28.3	26.7	54.2	19.2	40
ز	48.3	34.2	59.2	50.8	22.5	67.5
س	47.5	45.8	42.5	57.5	30.8	31.7
ش	67.5	29.2	45	78.3	35.8	53.3
غ	90.8	5.8	68.3	88.3	19.2	63.3
ف	73.3	38.3	45	85	45	60.8
ق	76.7	16.7	45.8	85.8	51.7	67.5
ك	48.3	53.3	47.5	46.7	59.2	51.7
ل	67.5	26.7	8.3	65.8	24.2	36.7
م	92.5	68.3	73.3	94.2	62.5	79.2
ن	45	38.3	58.3	57.5	42.5	60
ه	20	24.2	54.2	28.3	35.8	58.3
ي	65	72.5	25	69.2	65.8	32.5
ئا	39.2	78.3	42.5	46.7	76.7	40
ئه	60	11.7	23.3	72.5	25.8	46.7
پ	35	39.2	35	60	60	48.3
چ	47.5	38.3	25	62.5	55.8	54.2
ژ	48.3	5	39.2	54.2	36.7	57.5
گ	51.7	67.5	28.3	52.5	70.8	41.7
ڭ	50.8	49.2	26.7	62.5	71.7	50.8
ئو	53.3	35	44.2	67.5	19.2	44.2
ۇ	46.7	36.7	11.7	49.2	28.3	14.2
ۆ	39.2	50	28.3	55	42.5	30.8
ئۇ	78.3	30.8	15	79.2	31.7	16.7
ۋ	63.3	45.8	40	74.2	54.2	58.3
ئې	64.2	45.8	50.8	83.3	70.8	74.2
ئى	65	38.3	60	78.3	53.3	65.8

This paper calculates the recognition accuracy for each character. Table1 shows, one feature has a better recognition effect for one character, and to another one is not very good. We can see that recognition effect of the combined feature based on peripheral contour feature, stroke number feature and additional part's location feature is relatively good. Overall, ف,ئې,ل,م,غ have higher recognition accuracy, and ز,ر,د,ه,ئا,ك,ۇ,ئۇ characters' recognition rates are not ideal. This caused by the character similarity, writing speed and writing randomness. Average test accuracy of different combinations of features and different classification scheme is as shown in table 2:

Table 2: Average test accuracy of different scheme (%)

Features:	P	G	P + G	Low Dim. feature + P	Low Dim. feature + G	Low Dim. feature + P + G
Average:	57.27	38.57	40.42	**66.04**	45.31	50.63

In experiment, through different classification scheme for different combination of features conduct test evaluation. Table 2 shows, combination of peripheral contour feature, stroke number feature and additional part's location feature gets the highest recognition rate, reached 66.04%.It is clear that, peripheral contour feature can better represent detailed information of a character, depict more subtle differences between similar characters, and stroke number feature and additional part's location feature played a major role in experiment. Identification mistakes are due to the natural shape similarity, writing randomness of Uyghur characters and etc.

4. Conclusion

Through analysis of writing style and shape of Uyghur characters, this paper conducts a deep research on on-line handwritten Uyghur character recognition scheme based on peripheral contour feature; and gets a higher recognition rate applying effective feature extraction method and quick search method based on preliminary classification. In experiment, we get 66.04% recognition accuracy using peripheral contour feature, stroke number feature, additional part's location feature and the minimum distance classifier. Due to the character recognition is an important part of string identification and word recognition, and due to the recognition error caused by natural shape similarity, writing randomness of Uyghur characters, which are the next step to research content that need to solve.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61263038) and Program for New Century Excellent Talents of the ministry of education (NCET-10-0969), and Key Technologies R&D Program of China (2009BAH41B03).

References

[1]H. Linfeng ,and Z. Hui, On-line handwriting Uyghur character recognition based on support vector machine (SVM),Computer Applications and Software, Vol. 36, No. 3,2012,pp.179-182.

[2]W. Shangqing, and Z. Fenghao Off-line handwriting Chinese character recognition based on Bayesian grid, Computer-Aided Engineering, Vol. 15, No. 3, 2006,pp.72-74.

[3]M. Ammar, and R. A. Majed, Online handwritten recognition for the Arabic letter set. World Scientific and Engineering Academy and Society (WSEAS), Vol. 11, No. 5, 2011,pp.42 – 49.

[4] C. Mohamed, and K. Nawwaf, Character Recognition Systems: A Guide for Students and Practitioners, New York, JOHN WILEY & SONS, 2007.

[5] Y. Baoshe, and I. Hoxur, A handwriting Uyghur character recognition algorithm, Computer Engineering, Vol. 36, No. 2, 2010,pp.186-188.

[6] I. Mayire, Research on the technology of online handwritten Uyghur characters recognition, M.S. thesis, Xinjiang University, Urumqi, China, 2009.

[7] I. Mayire, H. Askar, Design and Implementation of Prototype System for Online Handwritten Uyghur Character Recognition, Wuhan University Journal of Natural Sciences, Vol. 17, No. 2,2012,pp.131-136.

[8] Z. Meng, and Y. Zhongqiu, Image preprocessing research in Handwriting numeral recognition. Micro-Computer Information, 2006,Vol. 22, No. 6, 2006, pp.256-258.

[9]I. Mayire, and M. Kamil, Multi-classifier combination scheme for recognition of online handwritten Uyghur characters, Computer Engineering and Applications, Vol. 48, No. 31,2012,pp.140-145.

[10]D. Ranagul, Research on the key technologies of online handwritten Uyghur word recognition, M.S. thesis, Xinjiang University, Urumqi, China, 2011.

[11]C. Huiming, Application study of image Euclidean distance in face recognition, Computer Engineering and Design, Vol. 29, No. 14, 2008,pp.3735-3737.

Zulpiya KAHAR, M.S. Student. She was born in Xinjiang in 1987, China. She received the bachelor degree from Xinjiang University, China in 2010. She is currently working toward M.S. degree in Xinjiang University, China. Her research interest is character recognition.

Mayire IBRAYIM, She was born in Xinjiang in 1981, China. She received the M. S. degree in 2009 from Xinjiang University, China. She is currently working toward PhD. degree in Wuhan University, China. Her research interests are handwriting recognition and text image recognition

Dilmurat Tursun received B.E. in 1983 in Electrical Engineering, from Xinjiang University of China. Currently, he is a professor in the Institute of Information Science and Engineering of Xinjiang University. His research interests include image processing and natural language processing

Askar Hamdulla received B.E. in 1996, M.E. in 1999, and Ph.D. in 2003, all in Information Science and Engineering, from University of Electronic Science and Technology of China. In 2010, he was a visiting scholar at Center for Signal and Image Processing, Georgia Institute of Technology, GA, USA, tutored by Professor Biing-Hwang(Fred) Juang. Currently, he is a professor in the School of Software Engineering, Xinjiang University. He has published more than 80 technical papers on speech synthesis, natural language processing and image processing. He is an affiliate member of IEEE.