http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (19): 4932-4936, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4932.4936

Recognition of Uighur New Writing Case-same Shape Letters

[1]Zhao Fei, [1]Halmurat Mamat and [2]Liu Xi-Jun
[1]School of Information Science and Engineering,
XinJiang University, Urumqi Xinjiang, 830046 China
[2]Xinjiang Laboratory of Multi-Language Information Technology,
Urumqi Xinjiang, 830046, China

Abstract: Traditional template matching recognition method can be easily confused with the recognition of the Uighur New writing case-same shape, Overcome the traditional template matching's shortcomings and improves the recognition rate which needs to do some further research on the basis of template matching to the identification of the case-same shape. Method is mainly based on the judgment that capital letters outlines are relatively big and small letters outlines are relatively small for the same form letters and then differentiate them, namely according to ratio of letter's outline size and the full text average of letters outlines size to judge the case of the letters. Results show that, this method significantly improves the recognition rate of the characters.

Key words: Uighur new writing, case-same shape letters, recognize, template matching,

INTRODUCTION

The printing of Uighur new writing has been widely applied in Xinjiang from the 1960s to the 1980s. However, much valuable information have been only saved on papers due to the limited technology of that day. Now they should be organized and electronical which contain the recognition of the documents written by the new writing. The recognition of printed Uighur new writing (Chuan and Zhu, 1964) occupies an important position in the recognition of national languages and has been paid attention and researched. In the recognition system, the wrong recognition of case-same letters has a great influence on the rate of overall recognition, that is why we need to carry out research on the recognition of case-same letters.

Research on the recognition of case-same letters is rare at present. There is not previous and targeted research on the recognition of the case-same letters of Uighur new writing yet. Layout analysis method (Abdelwahab and Rolf, 1998) is applied as the main method in the recognition of case-same letters in English which exploits the layout of letters to determine the case of English letters in the post-treatment (Wang and Halmurat, 2011). The judgment of case-same letters in the Uighur new writing mainly remains in the stage of the extraction of features (Feng and Tang, 2010; Alimjan and Halmurat, 2010; Pang and Jin, 2007). Since the extraction of features is carried out after the normalization of characters and case-same letters are hardly identified in shape, it is very difficult to use feature extraction to distinguish the case of letters which have the same shape.

Initial research on the recognition of Uighur new writing is relatively late. According to the available information, it was originated from 'Research and development of a printed Uighur new writing character recognition system' (Zou *et al*., 2012) by Xinjiang University in recent years and has not yet been mature with rare information. The technology of recognition of printed English letters (Yin *et al*., 2008) and Chinese letters (Wang and Feng, 2010) are very mature. So the recognition of Uighur new writing can use related result of Chinese and English literature (Su and Shr, 1993; Wu and Ding, 2001) for reference. However, compared with Chinese and English, it also has its own characteristics:

- Uighur new writing letters are similar to English letters but it has eleven more letters and five more case-same letters than English
- Uppercase and lowercase letters alternate frequently in the essay of Uighur new writing
- Uighur new writing has more similar letters than English
- There is a great difference between Uighur new writing and Chinese characters. Chinese characters is Chinese character while Uighur new writing is made up of letters

Corresponding Author: School of Information Science and Engineering, Xinjiang University, Urumqi, 830046, China

After learning from recognition of Chinese and English letters, Recognition of Uighur new writing has applied template match recognition (Bian and Zhang, 2007). Based on its own features which have more case-same letters, research and distinguish of case-same letters become a significant part of new writing recognition system. Recognition of case-same letters is to record its average size (that is the sum of outline length and width) before normalization of recognized letters, to calculate the ratio between outline size and average outline size of all characters. The ratio is a relative size and should be recorded. After template matching recognition, find out case-same letters from the result of recognition one by one and judge the case of each letters according to its relative size. If the relative size is larger than one certain threshold, the letter is a uppercase letter, otherwise a lowercase letter.

RECOGNITION OF CASE-SAME LETTERS

The recognition of case-same letters is based on template match recognition. The following text will introduce the characteristic of Uighur new writing and its recognition system.

CHARACTERISTICS OF UIGHUR NEW WRITING

Uighur new writing consists of the Latin letters. It has 37 letters Which include 26 Latin letters, 7 extended Latin letters (Ⱪ, H, ɵ, θ, Ü) and 4 two-letter (ZH, CH, SH, NG).

Besides the above features, Uighur new writing letters are divided into uppercase and lowercase letters which has 14 case-same letters (As shown in Table 1). It is difficult to recognize many case-same letters. How to distinguish the case of letters with the same shape is the main content which needs to be researched.

All letters of Uighur new writing:

Uppercase: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, Ƣ,Ⱨ,Ⱪ,Ə,Ɵ,Ʊ,Ƶ, NG, Zh, CH, Sh. Lowercase:a, b, c, d, e, f, g, h, I, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, ƣ,h,ⱪ,ə,ɵ,ü,ƶ, ng, zh, ch, h

UIGHUR NEW WRITING RECOGNITION SYSTEM

The Uighur new writing recognition system is based on Research and development of a printed Uighur new writing character recognition system, the object is printed Uighur new writing, The modules of recognition system are pretreatment module, feature extraction module, template library module, match recognition module, uppercase and lowercase judgment module.

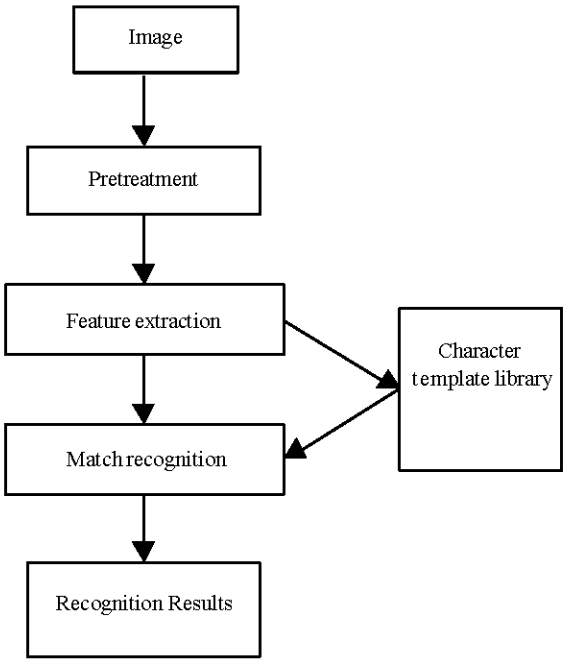

Fig. 1: Architecture printed Uighur new writing template matching recognition system

Table 1: Case-same shape letters in Uighur new writing

Uppercase	C K O P S V W X Z Ƣ Ⱪ Ə Ɵ Ƶ
Lowercase	c k o p s v w x z ƣ k ə e z

Presentation of template match recognition: As shown in Fig. 1, In the process of recognition. Firstly, input the scanned images of printed Uighur new writing, then get the final result through preprocessing, feature extraction and matching recognition.

- The preprocess of image is to transfer the 256 bmp figure to grayscale figure, binarization, denoise, tilt correction (Tang *et al*., 2013), character segmentation, character normalization and other treatments
- Feature extraction is to extract the corresponding features of characters from the preprocessed image. Then apply the better extracted features to distinguish different characters
- Match recognition is to compare the feature of unrecognized characters with the features of characters in the template library. Then obtain the most similar character as the result of recognition according to the match results

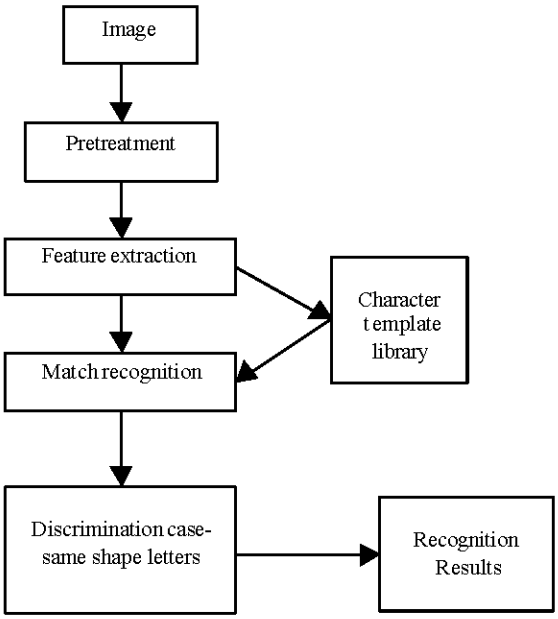

Fig. 2: Architecture of added discrimination case-same shape letters template matching module recognition system

Table 2: Ratio of Case letters and the corresponding sum of length and width

j	Upper	Lower	RLW1[j]	RLW2[j]	Var[j]
0	C	c	145	95	25
1	K	k	159	95	32
2	O	o	159	95	32
3	P	p	136	124	6
4	S	s	131	80	25
5	V	v	152	95	28
6	W	w	158	109	24
7	X	x	159	93	33
8	Z	z	134	88	23
9			163	138	12
10			160	139	10
11			138	88	25
12			156	95	30
13			158	113	22

- Character template is a library which consists of all characters and their features. After the feature extraction, add corresponding characters to each feature to form a template library

Problems in the template match recognition: The Uighur new writing recognition system applies the template match recognition. Template match recognition is to normalize the character first, Then distinguish the characters based on the characteristics of the characters. As case-same letters almost have the same characteristics, it's difficult to tell whether the letters are uppercase letters or lowercase. In the Uighur new writing, too many case-same letters which have the same shape lead to the wrong recognition easily. Therefore, the recognition of case-same shape letters is the key to improve the rate of right recognition. Case-same shape letters which can lead to wrong recognition easily are in Table 2.

Uighur new writing recognition system: Compare Fig. 1 with Fig. 2, As can be seen that the Uighur new writing recognition system is based on template match recognition system while the latter has no uppercase letter and lowercase letter judgment module. The uppercase letter and lowercase letter judgment module showed in Fig. 2 is the main part need to be researched as followed. It means the recognition of case-same shape letters.

CASE-SAME SHAPE LETTERS RECOGNITION METHOD

For an article to be recognized, the significant difference between the uppercase letters and lowercase letters is that uppercase letters are larger than lowercase letters. For the case-same shape letters, They can be recognized as long as we mark the larger letters as uppercase letters and the relatively smaller letters as lowercase letters. After the unrecognized image is pretreated and split (Fig. 3), The recognition system records the length and width of each split characters. After the recognition, distinguish uppercase letters and lowercase letters according to the size of the length and width of the letters if the letters are case-same shape letters.

Average sum of length and width of the character: For an article to be handled as following:

Step 1: After the pretreatment and split, record the length of the ith character as Length[i] until you have recorded the length of each characters; then record the width as Width[i]. Such as the letter? in Fig. 2, Length[i] is the length of side ab in the quadrilateral abcd and Width[i] is ad

Calculate the summation of Length[i] and Width[i]:

$$LWth[i] = Length [i]+width [i] \quad (1)$$

Step 2: Calculate the average of the sum of length and width of the character

Calculate ith character's sum of length and width as LWth[i], Until calculate the sum of length and width of each character in the unrecognized article. The average

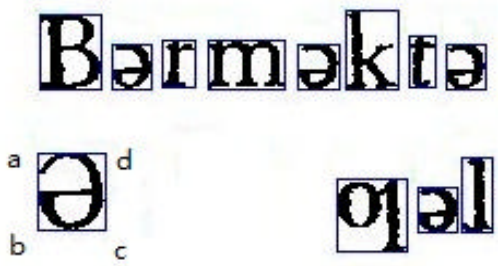

Fig. 3: Image segmentation

Table 3: Contrast of the experimental result

Order	num1	num2	num3	num4	num5 (%)	num6 (%)
1	857	93	42	41	89.15	93.93
2	671	52	21	19	92.25	95.08
3	894	56	15	15	93.73	95.52
4	741	64	32	30	91.36	95.41
5	792	59	20	19	92.55	94.95
6	570	65	32	31	88.60	94.04
Sum	4525	389	162	155	91.40	94.83

value aver equals the sum of Length and width of all characters divided by the total number of characters in the article.

$$aver = (\sum_{i=0}^{n-1} LWth[i]) / n \quad (2)$$

where, n is the total number of characters in the article).

Ratio of sum of length and width of the character to aver: Calculate the ratio of sum of length and width of the character to aver and mark as RLW[i], namely:

$$RLW [i] = LWth[i]/aver \quad (3)$$

Judgment of uppercase letters and lowercase letters:

- Judge the uppercase letters and lowercase letters if the result of the recognition system is any letter in Table 2

The rule of judgment: If the result of recognition of ith character is any letter in Table 2, you must make the following judgment (Eq. 4):

$$RLW[I]> (RLW2[j]+Var[j] \quad (4)$$

If the above condition holds, it is a uppercase letter, otherwise a lowercase letter, until the judgment of all the letters.

- The computing method of RLW1 [j], RLW2 [j] and Var [j]

RLW1 [j] is the corresponding ratio of uppercase letters in Table 2. RLW2 [j] is the corresponding ratio of lowercase letters. The value of j determines the corresponding value in Table 2 of RLW1 [j], RLW2 [j] and Var [j].

For example: The RLW1[0] of letter C equals to 145, the RLW2[0] of letter c equals to 95, then Var[0], the corresponding value of letter C or c, equals to 25.

Still take the letter C as an example:

Get 10 statistics of the average value of sum of length and width of uppercase C namely LengthWidth_C and the average value of sum of length and width of lowercase c namely LengthWidth_c from different articles.

Calculate RLW1 [j], RLW2 [j]:

$$RLW[0] = Length\ Width\_C/aver \quad (5)$$

$$RLW2[0] = Length\ Width\_C/aver \quad (6)$$

The remaining letters'RLW1[j] and RLW2[j] are calculated with the same method, the range of j is (0, 1, 2,....., 13).

Calculate Var [j]

$$Var[j] = RLW[j]-RLW2[j]/2 \quad (7)$$

Through statistics about different size of images of articles written in Uighur new writing need to be recognized, we can find that RLW1 [j], RLW2 [j] and Var[j] are stable. So the value of RLW1[j], RLW2[j] and Var[j] need to be calculated once and can be applied to all the articles written in Uighur new writing. For the articles written in the other languages, the value of RLW1[j], RLW2[j] and Var[j] need to be calculated again in order to improve the accuracy.

EXPERIMENTAL RESULTS AND ANALYSIS

Experimental images are scanned from "Xinjiang art" published in 1978 whose resolution are 300dpi and 256 gray levels. As most of the published books are old with poor quality of papers and printing, the quality of scanned images have a certain effect on the result of recognition. The experiment puts six random images through two kinds of recognition. Firstly, template match recognition which has no the module of the distinguish of uppercase letters and lowercase letters. Secondly, template match recognition which has the module of the distinguish of uppercase letters and lowercase letters.

The result of the experiment is shown in Table 3 in which num1 represents the total number of in the article, num2 represents the number of characters that are

mistakenly recognized before the involving of the module of the distinguish of uppercase letters and lowercase letters, num3 represents the number of characters that are mistakenly recognized because the case-same letters that have the same shape before the involving of the module of the distinguish of uppercase letters and lowercase letters, num4 represents the number of corrected characters after the involving of the module mentioned above, num5 represents the rate of successful recognition before the involving of the module, num6 represents the rate of successful recognition after the involving of the module.

From above Table 3, the average rate of successful recognition increases by 3.43% with the involving of the distinguish of uppercase letters and lowercase letters. The result of recognition is better on the whole.

Advantages: According to the experimental results, the algorithm of the distinguish of uppercase letters and lowercase letters can distinguish the uppercase letters and lowercase letters well and improve the rate of successful recognition significantly. Besides, the algorithm can be applied to distinguish the uppercase and lowercase letters which have the similar shape. Such as U and u, Ü and ü and so on.

Disadvantages: The result is not so well when applies the algorithm to some uppercase and lowercase letters whose sum of length and width have no significant difference. Such as P and p.

Besides, statistical information about the relatively outline size of the case-same letters should be obtained before the system is applied. The process of statistics is quite cumbersome.

CONCLUSION

Recognition methods of case-same letters has not only enhanced deficiencies of traditional template matching recognition but also greatly improved system recognition rate. However, there are limitations such as: (1) Poor distinguish of P and p, complicated task for computers to calculate coincident ratio of case-same letters, etc. (2) Articles ready to be recognized should be in the same letter size which has big effects for the recognition rate. This method will have good practicability since most of articles can meet this requirement.

According to average calculation mode, characters with different size will affect the stability of average value. Method for recognition of characters with different size should be strengthened. Removing statistical judgement for big size characters (Such as: Head with relatively big size) and small size characters (Such as: Punctuation and small characters, etc.) is the solution. Taking character recognition only for characters with the same size can enhance stability of average value which adapts to articles with different character size and this is what we need to verify. It is not enough for just correcting case problems in the recognition system, for example: false recognition of l, 1 and I. This kind of errors can be improved by applying dictionary methods (Chen *et al*., 2006) which will enhance system recognition rate.

REFERENCES

Abdelwahab, Z. and I. Rolf, 1998. Optical font recognition using typographical feature. IEEE Trans. Pattern Anal. Mach. Intell., 20: 877-882.

Alimjan, Y. and M. Halmurat, 2010. Preprocessing and feature extraction of on_line Uyghur handwriting recognition. J. Xinjiang Univ., 27: 232-237.

Bian, Z.Q. and X.G. Zhang, 2007. Pattern Recognition. Tsinghua University Press, China, pp: 136-156.

Chuan, M.J. and Z.N. Zhu, 1964. Uygur, Kazak language reform program and the Uighur new writing program. Language Reform, 12: 1-3.

Chen, G.P., M.X. Zhang and Y.W. Fu, 2006. Implementation of high performance multi-font printed english character recognition system. Comput. Eng. Appl., 12: 183-186.

Feng, W.X. and M. Tang, 2010. Detailed Image Pattern Recognition Programming using Visual C++. Machinery Industry Press, China, pp: 204-205.

Pang, D.H. and W.J. Jin, 2007. English character feature extraction. Comput. Simulation, 24: 208-210.

Su, L. and D.M. Shr, 1993. Efficient algorithms for segmentation and recognition of printed characters in document processing. IEEE Pac Rim, 5: 240-243.

Tang, Q.Q., M. Halmurat and A. Saypidin, 2013. Skew detection and correction of Uighur character scanned page. Appl. Res. Comput., 5: 1551-1557.

Wang, J. and M. Halmurat, 2011. Printed uygur character recognition post-processing. J. Xinjiang Univ., 28: 248-252.

Wang, K.J. and W.X. Feng, 2010. Chinese Printed Document Recognition Technology. Science Press, China, pp: 52-71.

Wu, Z.J. and X.Q. Ding, 2001. Implementation of robust multi-font printed english character recognition system. Comput. Eng. Appl., 20: 120-122.

Yin, F., W.B. Wang and D.Y. Chen, 2008. Architecture and implementation of the printed english document recognition system. J. Harbin Univ. Sci. Technol., 13: 9-12.

Zou, X., M. Halmurat and S. Arkin, 2012. Research and development of a printed Uighur new writing character recognition system. J. Xinjiang Univ., 29: 223-228.