A PROTOTYE OF MULTI-FONT PRINTED CHINESE CHARACTER READER

Shu Wenhao Guo Dong-min

Harbin Institute of Technology

China

ABSTRACT

An approach to multi-font printed Chinese character recognition is proposed in this paper. The problems of inputting image of characters, preprocessing, character segmentation,feature extraction as well as character classification have been discussed. According to the characteristics of multi-font printed Chinese characters,the number of cutting across strokes, the external and internal areas within a character are used as features for multi-stage classification. Experiments show that this method has the advantages of overcoming noises, displacements of characters and it can be achieved identical recognition result when a character appears in different fonts. On this basis, a multi-font printed Chinese character recognition system has been built, which can recognize 3755 most frequently used printed Chinese characters with Song and Bold Face fonts mixed in one page of document by using a same dictionary. The average recognition rate is more than 99%.

I.Introduction

The computer-based Chinese information processing systems are getting to be widely used in China in recent years. The great power of computing machines is gradually recognized by the broad masses of people. The execution time of modern digital computer is measured in nanoseconds. The laser printer can print texts with a speed of several thousands lines per minute. So the problems of information processing and information output may be considered to be solved. The only question left with the Chinese information processing system is the input of Chinese characters. Till now people input Chinese characters into computer by striking keys on a keyboard, using a specially designed code, which is difficult to be accepted by Chinese people and the input speed is apparently slower than those of computer and laser printer. So the Chinese character input is usually considered as a "bottle-neck" of the whole Chinese information processing system.

The pattern recognition technique is widely accepted as a promising tool to be used to break this "bottle-neck". This is why Chinese character recognition becomes attractive to many researchers in China. In recent years there has been tremendous progress in research on Chinese character recognition. Quite a few experimental single-font printed Chinese character recognition systems have passed their examination tests and very good results have been reached. From the practical point of view it is necessary to solve the problems associated with multi-font printed Chinese character recognition, because that in practical case a page of Chinese document is usually consisted of many Chinese characters printed in different fonts. To meet these practical requirements we are proposing a prototype of multi-font printed Chinese character reader in this paper.

II.Hardware configuration

The proposed system is consisted of a facsimile RICOH FX-120 connected through a special interface to a 16-bit microcomputer IBM PC/XT. The block diagram is shown in Fig.1

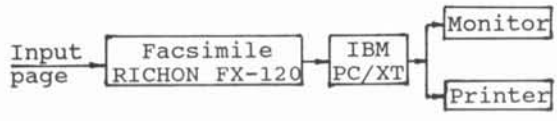

Fig.1

The facsimile is used as image input device.The main functions of the interface are :

1. Serial to parallel conversion.

The signal from CCD scanner is in serial format. It is necessary to convert it into parallel before being read by computer.

2. Synchronization.

Special measure has been taken to

guarantee the synchronization between the sending and recieving of image data.In our case WAIT instruction and TEST port were used.

3. Automatic feeding of pages.

Through connecting the control line to a certain port of computer it can be achieved programmable control of page feeding.

III.Preprocessing.

One page of scanned document image has to be segmented into lines and single characters before being recognized. A knowledge-based segmenting method was proposed in our research. From statistics of many times experiments we discovered that there are very few cases with continuous three black dots appearing in non-character regions. This kind of knowledge has been used as criteria for deciding line and character segmentation. We stored the maximum, minimum and normal widths of Chinese characters with different sizes, as well as threshold values of distances between characters in the computer as parameters and rules for segmentation. Taking the horizontal and vertical projections of current line of character image and comparing the peaks and valleys of these projections with those stored parameters the segmentation can be achieved quite correctly.

IV.Feature extraction.

Three kinds of parameters have been taken as features :

1. Numbers of cutting across times.

While scanning the character image from left to right , the scanning line will intersect with strokes many times. the number of cutting across times will reflect,to a certain extent,the number of strokes and the structure of Chinese character skeletons. Dividing a Chinese character into several stripe regions, taking the cutting across numbers within each region and normalizing it, we will get 6 features from both X and Y directions, see Fig.2

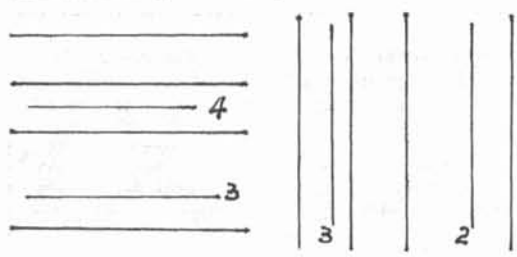

Fig.2 Cutting across feature

2. The external areas

The total number of white dots between the outmost edges of a character frame and the strokes which were first met from left to right or top down scanning is called as external areas. Dividing the frame of a Chinese character into several equally divided stripe regions, taking the external areas within each region and normalizing it, 16 features may be obtained from X and Y directions, see Fig. 3.

3. Internal areas

In order to distinguish Chinese characters with similar figures it is necessary to take interior features into accounts, so the internal areas are used as features to reflect the interior structure of a Chinese character. Scanning from left (and right) to the other side,accumulating the total numbers of white dots between changing point from black to white to changing point from white to black we get the internal area of a Chinese character in X direction. In the same way while scanning from top and down we may get the internal area of a Chinese character in Y direction.

Dividing the character frame into several equally divided stripe regions, taking the internal area and normalizing it we may get 16 features from both X and Y directions, see Fig.4

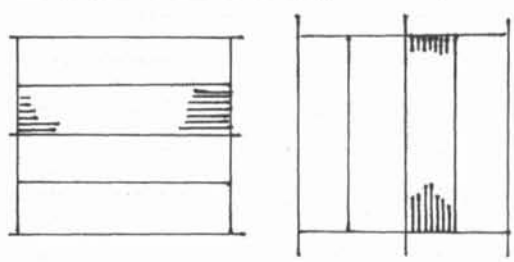

Fig.3 External area feature of Chinese character

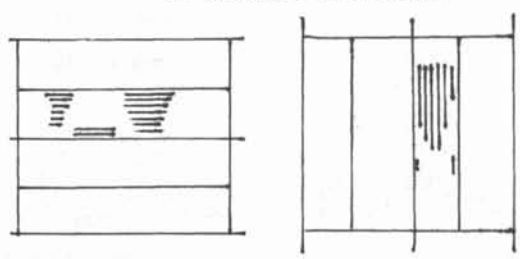

Fig.4 Internal area feature of Chinese charater

V.Feature selection.

As we can see, there are altogether 38 features obtained from the calculation in section IV. Obviously,it is

necessary to set up some criteria for selecting proper features among them. Here we calculate two kinds of quantities :

1.Global difference in various fonts of a same character :

$$YC_j = \sum_{i=1}^{3755} | tH_{ji} - tS_{ji} | \quad (j=1,2,...,38)$$

where, tH_{ji} represents the jth feature component of characters with bold face font; tS_{ji} represents the jth component of the same character with Song font.

The value of YC_j shows the difference between various fonts of a same character.So it is preferable to select those features whose values of YC_j is minimum. It will be of great advantage to the classification of characters in different fonts.

2. Global difference between different characters of a same font.

$$TH_j = \sum_{i=1}^{3755} \sum_{k=1}^{3756} | tH_{ji} - tH_{jk} | \quad (j=1,2,..,38)$$

$$TS_j = \sum_{i=1}^{3755} \sum_{k=1}^{3755} | tS_{ji} - tS_{jk} | \quad (j=1,2,..,38)$$

It can be seen that the larger values of TH_j and TS_j the more significant the distance between characters within one font. This is useful for the classification of single font characters.

According to the principles mentioned above,a parameter is defined as follows

$$\Delta T_j = TH_j + TS_j - YC_j \quad (j=1,2,...,38)$$

Alighing the ΔT_j in a descending order we will get :

$$\Delta T'_1 > \Delta T'_2 > \cdots > \Delta T'_{38}$$

According to $\Delta T'_i$ the order of jth vector can be determined. In our case we chose the first 32 feature vectors for classification. Experiments show that these features are capable of resisting noises and displacements as well as good for classification.

VI.Training and character recognition

In order to get a well-defined feature dictionary for recognition, it is necessary to do some training before loading it into computer. For this purpose an averaging algorithm is defined as follows :

$$M = (1/(n+m)) * (TH_1+TH_2+\cdots+TH_n+TS_1+TS_2+\cdots+TS_m)$$

where, TH_i , TS_i (i=1,2,..,n;j=1,2,..,m) are features of Bold Face and Song fonts respectively ; i, j represent ith and jth input; n,m denote the total input times of characters with Song and Bold Face fonts respectivelty.

Allocating the feature vectors and the corresponding national standard code for information exchange of every Chinese character in order, a final feature dictionary for recognition has been established. Altogether 3755 most frequently used modern Chinese characters have been included in this dictionary.

The character recognition procedure is:

1. Scanning the printed document through facsimile and transforming the character image into dot matrix;

2. Line and character segmentation;

3. Obtaining the feature vectors for every input character to be recognized through feature extraction calculations

4. Matching the feature vector with those in feature dictionary, find out the most similar one and make final decision.

The decision making criterion is as follows :

$$D(T_k,P) = \sum_{i=1}^{N} (w_i * | t_i - P_i |)$$

where w_i weighted factor i=1,2,..,N
t_i ith feature vector of input character
P_i ith fature vector in dictionary

VI. Concluding remark.

Based on the method proposed in this paper a prototype of multi-font printed Chinese character reader has been built up. It can read printed Chinese document mixed with characters in Bold Face and Song fonts. The documents being recognized are shown in Fig.5 and Fig.6. The recognized result is shown in Fig.7. The average recognition rate is more than 99%.

References

[1] Kenichi Mori and Isao Masuda, "Advances in recognition of Chinese charater", Proc. of 5th Int.Conf.on Pattern Recognition, pp. 692 702, 1980

[2] H. Fujisawa , Y. Nakano etc., "Development of a KANJI OCR : An optical Chinese character reader" pp. 816 819, 1978

[3] Wu Yu-pu, Zhao Jin-tai,"A summary of approaches to printed Chinese character recognition",1986

[4] Electronic Engineering Dept. of Qinghua university,"Proceedings on printed Chinese character recognition", 1986,10

[5] Zhong Cai-jei etc.,"A summary of OCR", Automation abroad,1983.3

[6] Standards of People's Republic of China,"Chinese character code for information exchangebasic set (GB 2312 80),Chinese standards publish company,1981
[7] Zhang Shou-shun,"Computer Processing on Chinese information", Aeronautic publish comp.,1984
[8] Guo Pin-xin," Chinese information processing technique", Renmin Yudien publish comp.,1985
[9] Guo Dong-min," The establishment of templates in corner method of Chinese character recognition ", Workshop on natural language processing, Beijing, 1985,11
[10] Shu Wenhao, " The improvements of stroke matching method in Chinese character recognition ", Academic Journal of Harbin Institute of Technology, 1986,6
[11] Shu Wenhao,"An approach to printed Chinese character recognition" 1982 Int. Conf. on Chinese Computing",Washinton D.C.,Sept. 1982
[12] Computer and system science dept. of Nankai university, " Technique report on Chinese character reader", 1987.12
[13] " Proceedings of 2nd national conference on Chinese character and speech recognition", Dailien, 1987.8
[14] Shu Wenhao, " The state of art in research on Chinese character recognition", Computer world, 1987.11.8

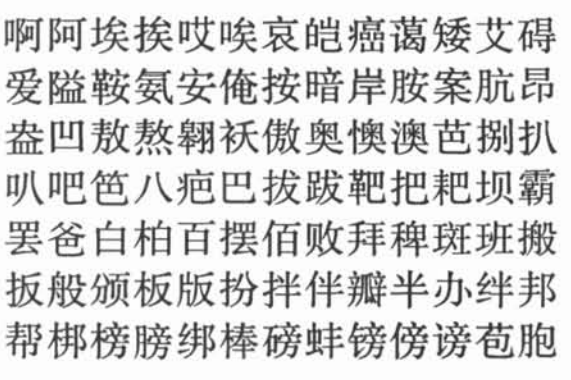

Fig.5 Printed sample for recognition (Song)

啊阿埃挨哎唉哀皑癌蔼矮艾碍爱隘鞍氨安俺按暗岸胺案肮昂盎凹敖熬翱袄傲奥懊澳芭捌扒叭吧笆八疤巴拔跋靶把耙坝霸罢爸白柏百摆佰败拜稗斑班搬扳般颁板版扮拌伴瓣半办绊邦帮梆榜膀绑棒磅蚌镑傍谤苞胞

Fig.7 Recognized reasult

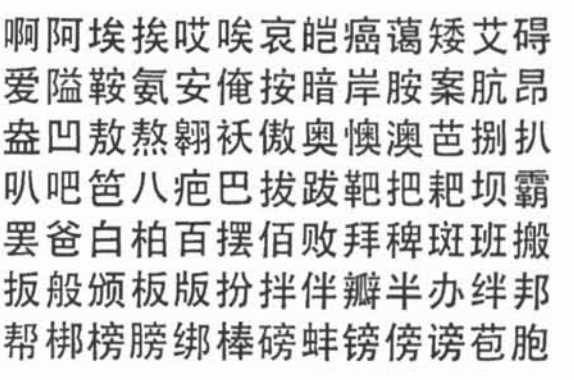

Fig.6 Printed sample for recognition (Blod Face)

啊阿埃挨哎唉哀皑癌蔼矮艾碍爱隘鞍氨安俺按暗岸胺案肮昂盎凹敖熬翱袄傲奥懊澳芭捌扒叭吧笆八疤巴拔跋靶把耙坝霸罢爸白柏百摆佰败拜稗斑班搬扳般颁板版扮拌伴瓣半办绊邦帮梆榜膀绑棒磅蚌镑傍谤苞胞

Fig.7 Recognized reasult