A New Fast and Effective Recognition Method based Construction Shape for Printed Digital

Guoli Zhang† and Guibo Ni††

† School of Mathematics and Physics, North China Electric Power University, Baoding 071003, China

† † School of Computer Science and Technology, North China Electric Power University, Baoding 071003, China

Summary
This paper proposes a structural analysis approach for Printed digital, through analyzing the structure shape of the printed digital. This method does not need to complex thin to the digital picture, reducing the erroneous question which is brought by thinning, so the recognition speed is quickly. The experiments are performed to test the efficiency of this proposed structural analysis approach.

Key words:
printed digital, features extraction, upside-horizontal line length, downside-horizontal line length, closed circle.

1. Introduction

The optical character recognition (OCR) has come into application times from experimental study for dozens years development. The printed digital recognition is a branch of OCR, having the enormous practical value. In the past dozens of years, the researchers proposed many recognition methods. These methods can be approximately divided into two kinds: one is based on statistical characteristic method, the other is based on structure characteristic method[1]. Generally speaking, the statistics characteristic classification is easy to train, and can also obtain a high recognition rate in training set. The structure characteristic method is able to describe the character structure of the recognition object, can effectively unify the geometry and the structure knowledge in the recognition process, therefore it can obtain the high recognition result. In fact, the printed digital recognition is correspondingly simple compared with the Chinese character recognition or other languages character recognition. First, the printed digital is few, only ten digitals are needed to identify, second, stroke is few, and the structure is also simple. Because the printed digital has above characters, we can recognize these digital by analyzing and using the especial structure of the digital, give up the old recognition strategy, such as binaryzation-denoising-regularization-thinning-feature extraction-classification. The combination of horizontal line character and the through-1ine numbers to recognize printed digital is given in literature[2-3], but this method may mistake 3 and 8 etc.. The method based on convex-concave character can recognize digital well, but some digital has its unique structure so that we need not extract convex-concave character for all digital. We only need to extract convex-concave character of the digital which is difficult to recognize. This paper combine two methods above to form a new recognition method for printed digital, the detail is described in section 2 and section 3.

Manuscript received May 5, 2008
Manuscript revised May 20, 2008

2. Character extraction

Because of the special structure shape of the printed digital, we can extract four kinds character: height-width rate character, upper-horizontal line length character, lower-horizontal line length character, closed circle character.

2.1 Height-width rate character

It is easy to see that the width of 1 is the smallest, In other words its height-width rate is biggest. In fact, height-width rate is most effective character of the digital 1 for recognition. In recognizing, firstly extract the height-width rate of the digital given, if this value is bigger than some given value, then the digital is recognized as 1.

2.2 Upper-horizontal line length character

There is upper- horizontal line in some digital, such as 5 and 7. We use the continuously pixel points in the upper of the digital to describe the upper-horizontal line length. As fig.1 shows:

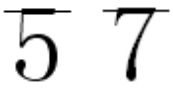


Fig1: Upper-horizontal line

The upper-horizontal line length of the digital 5 and 7 are biggest, which are almost equal to their width. We can distinguish them from other digitals by using the upper-horizontal line length.

2.3 Lower-horizontal line length character

There is lower-horizontal line in some digital, such as 2 and 4. We use the continuously pixel points in the lower of the digital to describe the lower-horizontal line length. As fig.2 shows:

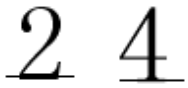


Fig2: Lower-horizontal line

The lower-horizontal line length of the digital 2 occupy almost whole digital width, the lower-horizontal line length of the digital 4 is relatively greater, so the digital 2 and 4 can be discriminated from other digitals by the lower-horizontal length..

2.4 Closed circle character

Closed circle character is to see that whether there is the closed circle in some digital. In general, we use the following method to estimate whether there is closed circle in the digital given, i.e. firstly, launch beam to top, foot, left, right, left-upside, left-downside, light-upside and right-downside from all back points, then compute the intersection numbers of the beam and the digital for every back point. The intersection numbers is called the value of the back point, written as VBP. If VBP is 8 for some back points, then there is a closed circle at least in the digital. For increasing recognition speed, need not compute VBP of all back points. Because there are the different structure shapes of the different digitals, so we can select to several special back points. We select six points from the position as fig.3 show and compute the VBP of 6 back points respectively. If the VBP is 8 for some back points, and then there is a closed circle at least in the digital. Furthermore, we use the following method to judge the position and numbers of the closed circle. We launch beam downward from upper three points to 3/4 lower position of the digital. If VBP is 8 for one of three points, then the upper part in the digital has a closed circle. Likewise when we launch beam upward from lower three points to 3/4 upper position of the digital. If VBP is 8 for one of three points, then the lower part in the digital has a closed circle. If both upper part and lower part in the digital have closed circle, then the digital has two closed circles, and the digital is 8. If there is no closed circle in upper part and lower part of the digital, we launch beam to top, foot, left, right, left-upside, light-upside, left-downside, and right-downside from the six points in order to examine whether the digital has closed circle, if VBP is 8 for one of six points, then the digital has one closed circle, and the digital is 0, otherwise the digital is 3.

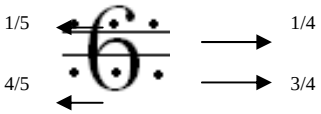
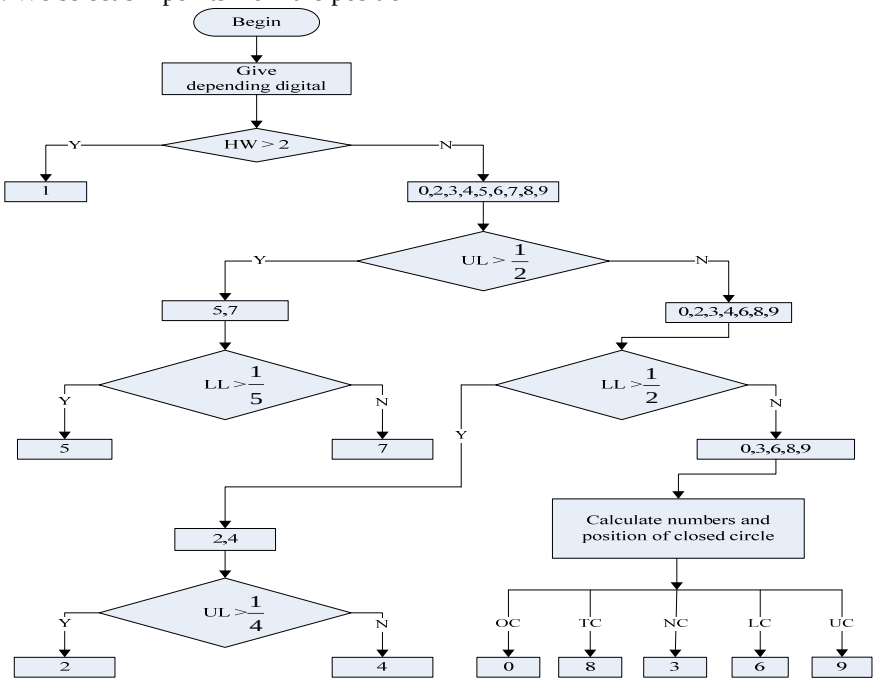


Fig3: Special back points

For example, digital 0 has one closed circle, digital 6 has a upper closed circle, digital 9 has a lower closed circle, digital 8 has two closed circles, digital 3 has not closed circle.



3. Recognition

In recognition, Firstly do pretreatment for the digital image, include doing denoising, binaryzation and regularization for pending digital image, then implement to recognize according to the following steps:

Step1 extracting height-width rate, written as HW, if HW is greater than the threshold value, the digital is recognized as 1. The threshold value is set as 2 in later experiment.

Step2 Extract upper-horizontal line length, written as UL, if UL is greater than 1/2 of the whole width of the digital, this digital is 5 or 7. The whole width of the digital is denoted by WWD.

Step3 Compare the lower-horizontal line length, written as LL, if LL is greater than 1/5 of WWD, this digital is 5, otherwise 7.

Step4 If UL is not greater than 1/2 of WWD, and then extracts LL. If LL is greater than 1/2 of WWD, this digital is 4 or 2,

Step5 Compare UL of them, if UL is greater than 1/4 of WWD, this digital is 2, otherwise 4.

Step6 If both UL and LL are smaller than 1/2 of WWD, this digital must be is one of 0, 3, 6, 8 and 9.

Step7 Extract closed circle of the digital, if a digital has a closed circle, it is 0. If a digital has two closed circle, it is 8. If the upper part of a digital has a closed circle, it is 9. If the lower part of a digital has a closed circle, it is 6. If a digital has no closed circle, it is 3.

Recognition flow as fig.4 shows.

Where OC represents one circle, TC represents two circles, NC represents no circle, LC represents lower circle, UC represents upper circle.

4. Experiment result

We use Matlab6.5 to recognize to two hundreds digital in Windows XP platform, the recognition rate can reach 100%. Experiment result shows that this method has such advances as simple, quick speed recognition, high precision and strong anti-jamming.

Acknowledgments

This work is supported by the homecoming Fund (20514001) of North China Electric Power University.

References

[1] Liu Huichun, Ma Shuyuan, Wu Pingdong, Li Xiaomei. Research 0n Handwritten Digits Recognition[J]. Computer Engineering. 2003,29(4).-24-25,61.

[2] Yan Guofi, Huang Shan, Li Daizhang, Shang Jianhua. Application 0f a Quick Recognition Algorithm for Figures in Num ber Recognition 0f ID Card Image[J]. Computer Engineering. 2003,29(1).-178-179.

[3] Xing Ruyi, XiaBing, Yuan Hongyun. The fast method on printing digital recognition[J]. Journal of Hebei Institute of Architectural Science and Technology. 2004,21(3). -65-67.

[4] Luo Jia, Wang Ling. A New Method for the Of-Line Recognition of Handwritten Digits Based on Convex-Concave Coding[J]. Computer Engineering & Science. 2007,29(5). 69-70,108.

[5] Ye Fei, Li Feng. Fast hand—written digital character recognition based on global feature[J]. Computer Engineering and Design. 2006,27 (22).-4347-4348,4352.