Extracting Characters From Books Based On The OCR Technology

Mingkai Zhang[1, a], Xiaoyi Bao[1, b] ,Xin Wang[1, c], Jifeng Ding[1, d] and Xiaoke Xu[1, e]

[1]School of Dalian,Nationalities University, Liaoning 116600, China;

[a]974569651@qq.com, [b]934952212@qq.com, [c]1134724200@qq.com, [d] djf@dlnu.edu.cn

Keywords: image pre-processing, character segmentation, feature extraction.

Abstract. Since the 1990s,information based on image, video and audio frequency become popular trends during daily information exchange. Information identification about images and videos which based on the Optical Character Recognition(OCR)have become more and more popular in various enterprise. Compared with traditional manual entry, OCR is not only save manpower cost but also optimize the resource configurations.

1. Introduction

With the development of communication technology, multimedia technological such as image, video and audio frequency is now coming into every field of our life. More and more people try to recognize text information from these multimedia and show them out to copy, store and export. Nowadays we plan to achieve the method to extract information in order to the next step for audio processing. So how to extract these information is a vital technology for this.

In the digital image processing field, text image has abundant edges and grains. There are both advantages and disadvantages between the two characteristics. As edge feature, the algorithm speed is faster but the complex background make the final result worse. As texture feature, universal property is widely but it has a higher complexity of algorithm. This article in predecessor’s research foundation, finds a algorithm which aims to scripts format and direction of rules to extract the information we need in text.

2.Overall design

The hardware of the system is collected by the image acquisition circuit, then analyze, process, recognize and understand the image by main chip, then display by the display circuit. The image acquisition circuit is a OV2640 camera, The camera has high sensitivity, high flexibility, support JPEG output, etc.. Main chip is STM32F407,it is a chip that the most advanced in 32 series,It has a DSP that the highest frequency band is168M. Display circuit is a 4.3 inch color screen,this module uses the TFTLCD panel that can display 16 colors true color image,software is processing the image though Image gray, Binarization, Image denoising, Character cutting

3.Image pre-processing

When the system test text image, it will locate and extract the text area of text image. Firstly we need to do is image pre-processing. Pre-processing can be divided into image gray, binarization, denoisingand character segmentation.

Image gray. Before processing the image, we need to transform the color image collected by camera into gray scale consist of 256 colors. In this way, computer can process the image better. In RGB model, when R=G=B, color express one of gray color and the number R=G=B called gray level, therefore, each pixel only need to one byte to put gray level(intensity or brightness value)which range form 0-255.

Binarization. Binarization image is gray, the pixels set to a value of 0 (black) or 255 (white), Text i mage binarization is separated from the background of strokes. Omitting irrelevant information for f acilitate subsequent processing. Then the binarization threshold is very important, The larger thresh old will retain more information,but a lot of useless information may be retained to cause interferen ce; The smaller threshold will lost a lot of useful information,cause the information is not complete when extract the text.So how to decide the binarization threshold is the important of the algorithm. Binarization algorithm has many kinds ,the flow diagram is as follows.

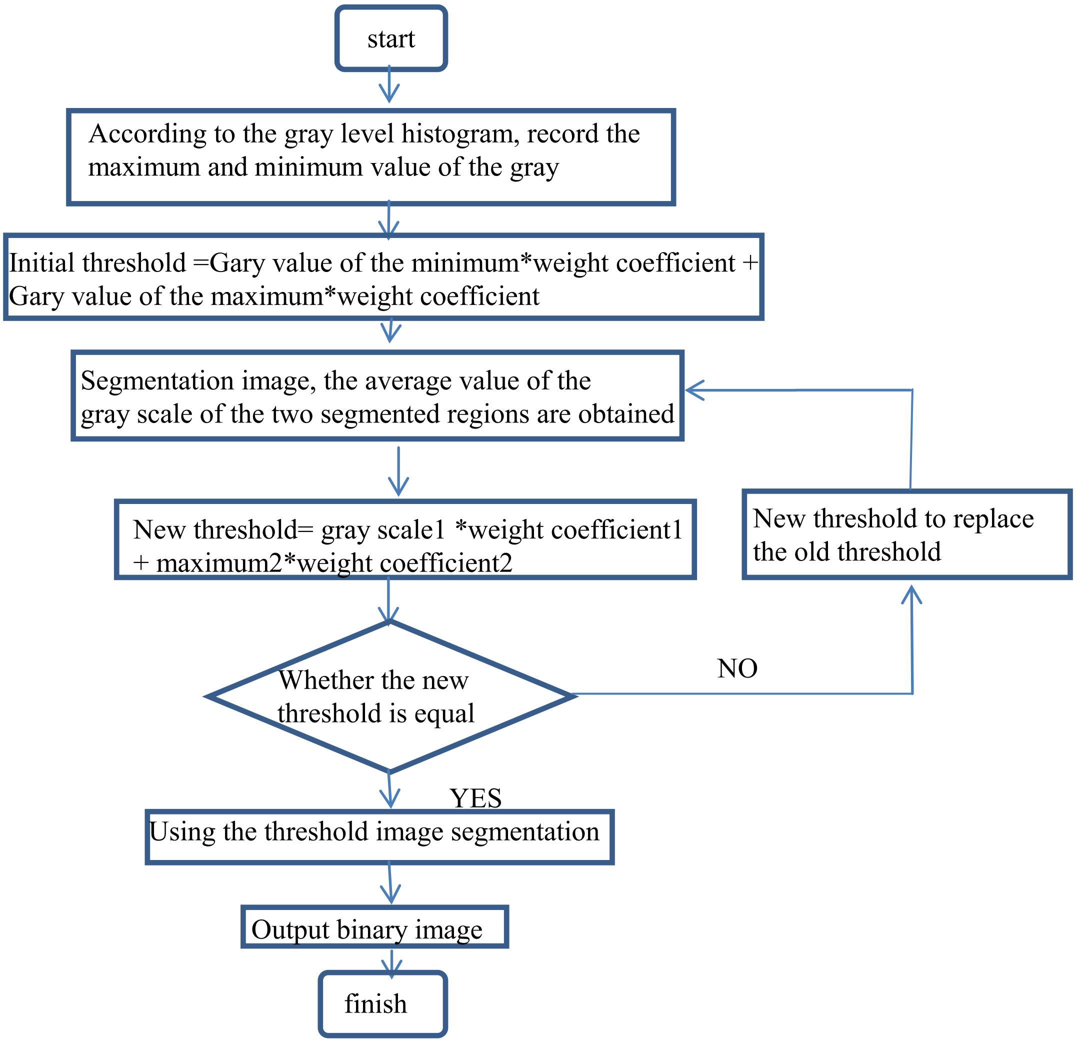
start
According to the gray level histogram, record the maximum and minimum value of the gray
Initial threshold =Gary value of the minimum*weight coefficient + Gary value of the maximum*weight coefficient
Segmentation image, the average value of the gray scale of the two segmented regions are obtained
New threshold= gray scale1 *weight coefficient1 + maximum2*weight coefficient2
New threshold to replace the old threshold
Whether the new threshold is equal
NO
YES
Using the threshold image segmentation
Output binary image
finish

Figure 1. Binarization flow diagram

Image denoising. The factors which hinder the information acquisition called pattern noise. These noises can be brought during spread or quantization. For collecting text image's system, the noise is more brought by the collecting process, which called multiplicative noise. What image denoising do is eliminating disturb factors that infect the right information. The figure diagram is as flows.

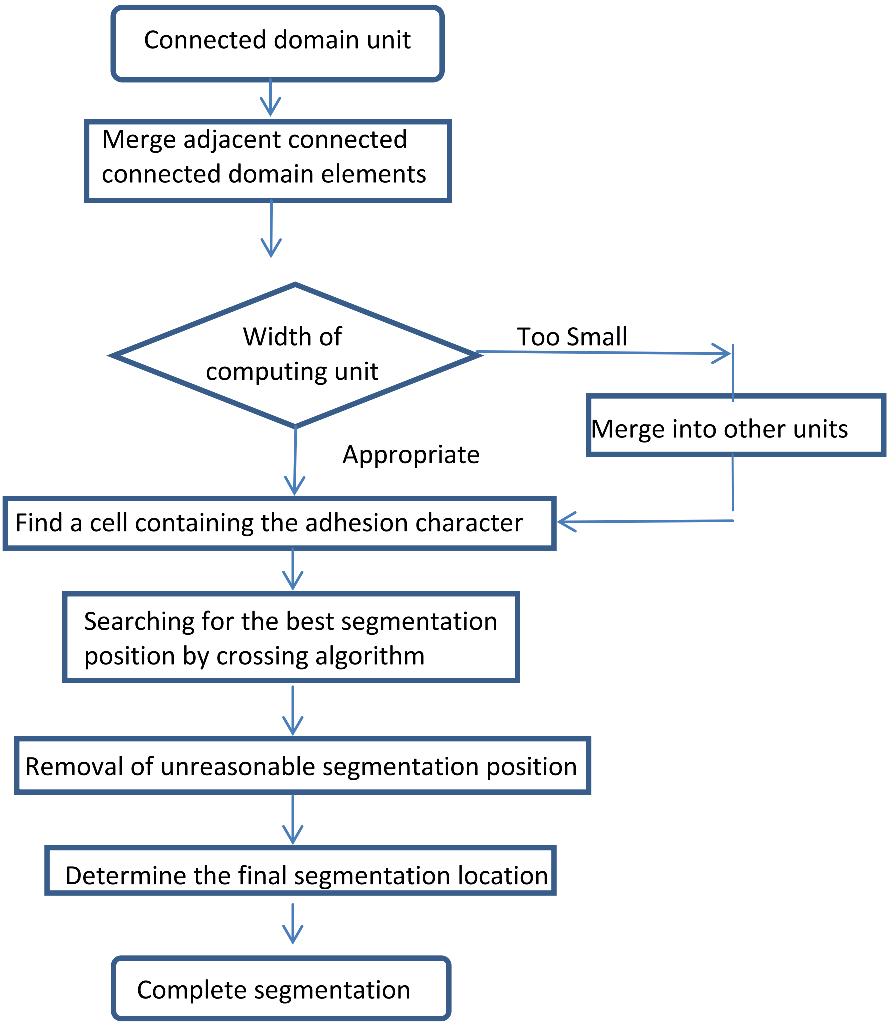
Connected domain unit
Merge adjacent connected connected domain elements
Width of computing unit
Too Small
Merge into other units
Appropriate
Find a cell containing the adhesion character
Searching for the best segmentation position by crossing algorithm
Removal of unreasonable segmentation position
Determine the final segmentation location
Complete segmentation

Figure 2. Denoising flow diagram

Character cutting. The contents of text image are not single character but combine with many characters include: Chinese characters, punctuation and charts. Character cutting is significant part of paper reading system because the direction test of text image is achieves by using a single character direction. If character in text can regard as the basis, we must divide whole text image into single character area. So the text requires character segmentation.

There are two methods of character cutting. One of it is based on characters' architectural feature. This feature use characters' height, width pitch and stroke structure feature to cut the character. Another method is based on character recognition, this method can recognize character while cutting. And confirm the final position according to the result of recognition. The second method need combine with recognition algorithm so we adopt the first algorithm.

4.Extract of Chinese character features

Every Chinese character feature's has different applicable environment. We should choose different Chinese character feature for different systems to reach the best result.

Thirteen feature point extraction method . Thirteen feature point extraction is a method that extracting thirteen feature points from each character. This algorithm has better adaptation.

First step is to divide texts into eight areas randomly and make statistic of the black pixel number, then we can obtain eight proper values.

Second step is to record the middle of two row and the middle of the two row, regard them as four feature vectors. It means draw two lines at horizontal's 1/3 and vertical 2/3 and record the number of black pixels of the two lines and the image of the text to get four character values.

The last feature value is the sum of all black pixels of the text image. We get thirteen feature point, that is the thirteen feature point extraction method.

5.Recognition algorithm

Correlation matching method. Correlation matching method is a statistical recognition method. The advantage of statistical method is that it is easy to build the feature extraction and template, strong anti-interference capability, it makes that it insensitive to local noise. The disadvantage is that the ability to distinguish similar words is weak, It is sensitive to the change of the style of writing. The figure diagram is as flows.

6.Systematic measurement and analysis

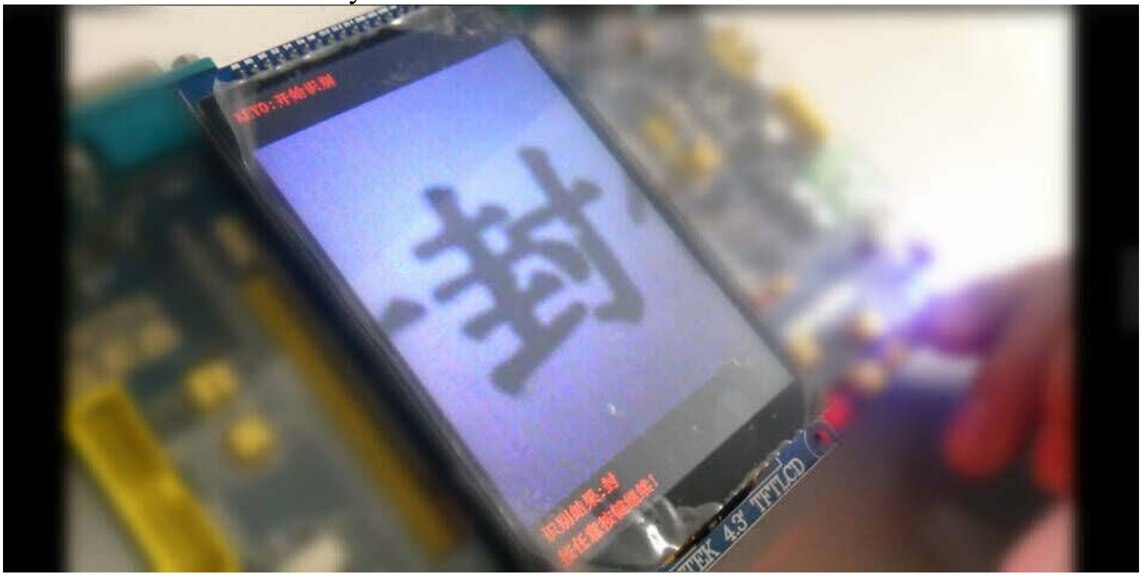

Figure 3. Test result

The system has the highest recognition rate for the training template.it can reach 100%, For any interception of the better quality of the text image, Recognition rate is also good, but the recognition rate is low for the text image collected by the camera

7. Summary

With the development of information society, Chinese character recognition technology is becoming more and more important. We designed this system on the basis of summarizing the previous work done by our predecessors. The system has achieved the expected requirements in recognition rate and recognition effect. The problem that needs to be solved is to improve the pretreatment technology of Chinese character recognition system.

References

[1]Takashi M.A survey of image Processing and computer vision software systems on work stations. IAPR/TCS,1992.1~35

[2]KitehenL,Rosenfeld A.Gray-levele comer detection [J]. Pattem Recognition Letters, 1982,l:95~102

[3] R.Stefaneli.Some Parallel thinning algorithms for digital Pictures. J. Ass. ComPuter [M]. Mach. 18, 1971,255~264

[4] Keghbati H K, An Overview of Data ComPression Technique. ComPuter, 1982,14(4):71~76

[5]Mori K, LMasuda. Advances in Recognition of Chinese characters, Proe. Of 5thInten. eonf. On Pattern Recognition, 1980:692~720.

[6]]Whilchello A P, Yan H. Linking broken character borders with variable sized marks to improve recognition. Pattern Recognition, 1996, 29(8): 1429 ~ 1433.

[7] Yu T, Tang Y Y. The feature extraction of Chinese character based on Contour Information. Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999:637~640

[8] Aftabul l, Md R H, Razwanur R et al. Designing ANN using sensitivity and hypothesis correlation testing. 10th International Conference on Computer and Information Teehnology,2008:l~6.

[9] Huang L, Huang X, Multiresolutiong Recognition of Offiine Handwritten Chinese Characters with Wavelet Transform,ICDAR,2001,63

[10] Wang X F, Ding X Q, Liu C S, Optinimized Gabor Filter Based Feature Extraction for Character Recognition, ICPR,2002,223