Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 4073 – 4078

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Character Segmentation System Based on C# Design and Implementation

Zekai Zheng, Jingying Zhao*, Hai Guo, Luping Yang, Xueai Yu, Wei Fang

Dalian Nationality University,Dalian,116000,P.R China.

Abstract:

At present, most of the OCR recognizing through individual character, thus the quality of character segmentation is the key point to affect the quality of OCR recognition system. This paper introduces the formula of projective method in analysis of preliminary segmentation for images. Moreover it applied analysis for connected spatial domain, the correct results shows that writing image well matched. After two analyses and segmentation, characters can be segmented correctly. In order to provide useful solutions to these two problems that characters keying must be performed rapidly and documents digitizing can be conserved for a long time. Therefore, we must place emphasis on the research and development of the character segmentation.

Keywords: pattern recognition; character segmentation; C#; connected spatial domain; pretreatment

1. Introduction

One of the most important aspects of pattern recognition is text recognition. It is a process that scanning on existing traditional text first , then analysis and disposal to results on scanned images. Of course , text recognition offers solutions to the problems that documents keying must be performed rapidly and text messages can be conserved for a long time as well. It will be divided into steps as follows: graphic input, image pretreatment, image recognition, recovery of layout.

Preprocessing is an important part of the character recognition system. The performance properties of Preprocessing directly affects performance of the whole recognition system. Therefore,it is necessary to study the technology and method which are applicable to the preprocessing of text recognition according to the characteristics of grapheme and writing style for various kinds of characters. Preprocessing carries

* * Corresponding author.
E-mail address: zjy@dlnu.edu.cn.

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.01.622

out noise reduction, graph double value , slant correction, layout analysis, character segmentation, normalization and edge detection, etc.

Character segmentation refers to the character information it obtains from scanned images. That is , extract character image from images. At the present stage , individual character recognition technology is more mature than others and under normal circumstances its accuracy rate of recognition is still quite high. If segmentation part of preprocessing were in low quality,recognition rate of the whole character recognition has been very low as a result. So, the quality of character segmentation directly affects operations of the whole recognition system.

2. C# and .Net Framework

C # is the object -oriented language, simple but nice and type safe. Developers can take advantage of it to build all kinds of applications operating in .Net Framework that is safe and reliable. C# could create simple client applications of Windows, XML Web services, distributed component, C/S applications, database applications, etc.

Through CLR(Common Language Runtime)， the program compiled by C# will run steadily on computers with .Net Framework. Application developers normally need not be concerned with using processors or Language. Tools described herein will run on it so long as with .Net Framework. Guaranteed compatibility and operation efficiency.

.Net Framework has provided many high effective tools for dealing with bitmaps. Such as Bitmap, BitmapData, Image, etc. Graphic image processing tools supplied with .Net bring great convenience for graphic processing. The class provided by .Net Framework taking account of both ease and speed. Choosing the tools it needs, according to the needs or standards of the user. But, if you have requirements for the drawing processing efficiency, perhaps a byte array built by BitmapData can meet your needs.

3. Common Sense of Segmentation

The OCR of the traditional image makes up of three phases: image pretreatment-> segmentation->recognition

The preprocessing algorithm includes image gradation, binary conversion, slant correction, and noise reduction , etc.

So the focal point of this article is to tell how to segment. The segmentation methods of this paper are main two kinds : one is based on projection, the other is based on connected spatial domain. Both approaches are so simple and extensively used.

3.1. Syncopation Based on Projection

Syncopation based on projection is projection method of being based on pixels of characters in a particular direction. Projection can be divided into two directions: horizontal and vertical projection, respectively. However, projection two of directions work on the same principle.

Usually syncopation based on projection is a method for segmentation which operations are conducted after the binary image processing and slant correction. Images after binarization are divided into two pixels: black and white. First, calculating the black pixel quantity of images in a horizontal direction. Then, margin between lines can be identified through finding troughs which reflected in statistical datas between their lines. But for reasons that region of character is a condensation which represented by closely spaced black pixel dots, while regions between lines dominated by white pixel.

When every line of boundary results gotten, make comprehensive analysis of vertical projection measured in unit one per line. Vertical projection works in very much the same way as horizontal projection. First, scanning the black pixel quantity of a line in a vertical direction to find troughs to get boundary results of column. Then, the initial results of projection segmented boundary can emerge after scanned twice. Syncopation based on projection so far has mostly done.

Syncopation based on projection is more effective for Individual character cutting, nonetheless, under condition of aliasing character or touching symbols or multiple character separation, would still be easily regulated to the structure of character writing. That is, for example , two characters could be syncopated to one word or one character might be broken down into multiple words.

3.2. Syncopation Based on Connected Region

First , this paper presents a analysis for searching connected area in binary image. And its operating principle is simple,as below:

Find the first black pixel dot of a image. If found, continue; or else, skip to 6)

a complete counterclockwise rotation

if met other black pixel dots, skip to 4), or else, skip to 5)

The current pixels would be set to white pixel dots, then turn into new pixels. skip to 3)

a single scan finished,skip to 1)

exit

Once connected component labeling strategy Implemented, we will get rapidly various kinds of information you want, It also includes amount of connected components, the starting and stopping point of each connected component object, the boundary of each connected component object, etc. Creating combination of above multiple sets of information, you can begin ana lysing the results to know the following: the area and relative location of each connected component, the state of aliasing, and so on.

Whereas, a direct rule–analyzing approach for connected component analysis is becoming more complex. Complicated relationships between connected regions makes worse. Therefore, connected component analysis is a incredibly effective way for rule–analyzing method of connecting area in binary images that quantity of connected components is not many.

4. Functions

In this segmentation system applications , the focus is Chinese characters(Fig1)(Fig 2),

Nakhi pictographs(Fig3), writing of Dai nationality (Fig4). Certainly, by focusing on customizing the rules primarily depend upon the segmentation of the Dai Wen.It is not acceptable to rely solely on syncopation based on projection for segmentation of Dai wen. Consequently, more analysis to be required. A sampling to segmentation as follows:

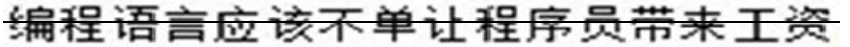

Fig 1 The original photo of Chinese segmentation

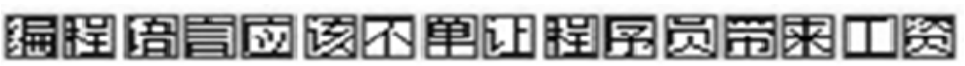

Fig 2 The designsketch of Chinese segmentation

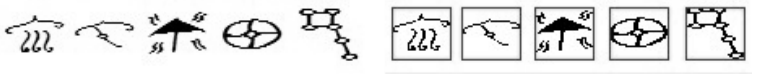

Fig 3 The designsketch of Nakhi segmentation

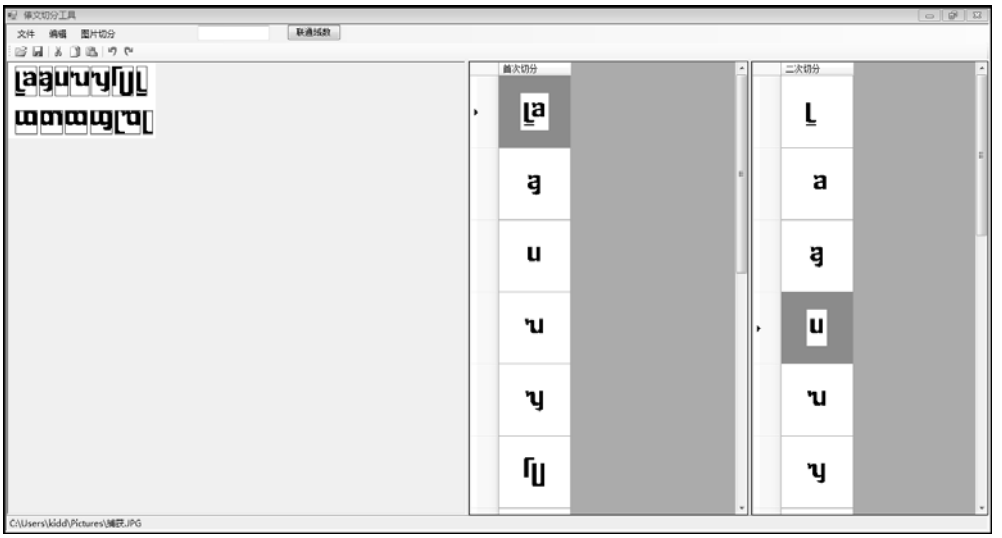

Fig 4 The designsketch of Dai Wen segmentation

5. Realization

The major function of experiments lies in researching and developing character segmentation system. Since certain characters cause aliasing when initial attached to final, the segmentation of Dai Wen is not based solely on projection or connect region.In Char Segment part , after several representative algorithms in existence is studied , a tactic is proposed , which is combined with connect region and projection to cut out images.

Bitmap and BitmapData class which is a part of C# are mainly used to process the image information in image processing. Although bitmap class can accomplish all the work on its own, in terms of processing efficiency, a byte variant array created by BitmapData manipulates the data more efficient.

Firstly, images even have a rudimentary segmentation, namely the syncopation based on projection. Secondly, after a continuous type byte array created by BitmapData class, next ,this array are then entered into the unidirectional analysis class that is AnalyzeCount to analyze their data .You could select horizontally analysis first , then vertically. OK also on the contrary. Both horizontally and vertically analysis that inherit from AnalyzeCount class. As shown in Fig 5.

As can be seen above , Rather great errors occur when processing of character images is designed only by using projection analysis . The association of Initials and finals,this phenomenon but projection analysis can't solve it. Therefore, after the first segmentation of rudimentary images, an Internal analysis of connection domain implemented according to the over-segmented images in a specified rectangular area. Usually the aliasing effect of Initials and finals in Daiwen pertains to left -right mechanism. So,make a rule in advance here,The left -right mechanism of two connection domains is not one word, while the upper and lower structure of two connection domains is one word. Neither left -right mechanism nor upper-lower structure could guarantee Images excursion to changes the results.So we

have need to set a threshold to identify whether the structural relationship between two graphs is left - right or upper-lower.

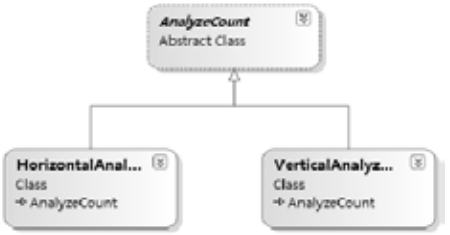

Fig 5 relevant subclass provided by AnalyzeCount and AnalyzeCount class

Combined horizontally analysis with vertically, we would get preliminary segmentation results based on projection. The analysis results can be displayed on GUI. To date,primary segmentation has been basically completed.when dealing with DaiWen, because of limitations of projection segmentation, the clustered space was divided to several parts, and each part corresponded with analysis of connected.Connected methods has been established as two sorts, the first is based on analysis of 8-connectivity, the other is 4-connectivity neighbor. In this application, progressive scanning is selected by default first. And then making a verbal scanning for each row. That way initial segmentation results obtained,but incomplete. Generally divided into two types, simply connected region and 4-connectivity neighbor in Dai Wen.Also the main structure as upper-lower. Projection segmentation alone will changes two or more words into one word caused by the writing of DaiWen. Analysis of connected are needed for primary segmentation regions. Yet the quantity of connected regions determines whether there is a single character. In other words, if the quantity was just one, it would identified as a single character soon;But if not, structural analysis would taken to judge whether it was upper-lower or left-right mechanism.Connected component analysis mainly adopts the marked analysis techniques of 4-connectivity neighbor with 8-connectivity one. Class graphs of connected component analysis shown as follows:(Fig 6)

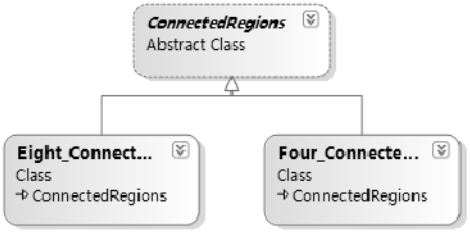

Fig 6 class and inheritance relationship in a connected component analysis

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China No. 60803096, and the Fundamental Research Funds for the Central Universities. So thankful for the support from the project of "TaiYangNiao " and the innovative experimental project of Dalian Nationalities University.

Reference

[1]Hai Guo, Jing-ying Zhao, Ming-jun Da, "A Preprocessing method for NaXi Pictograph Character Recognition", Journal of Convergence Information Technology, Vol. 5, No. 2, pp. 59 ~ 66, 2010.

[2]Hai Guo, Zhao jingying, Xiao-niu Li, "Preprocessing Method for NaXi Pictographs Character Recognition Using Wavelet Transform", International Journal of Digital Content Technology and its Applications. Vol. 4, No. 3, pp. 117 ~ 131, 2010.

[3]Hai Guo, Jing-ying Zhao, Ming-jun Da, Xiao-niu Li , "NaXi Pictographs Edge Detection Using Lifting Wavelet Transform ", JCIT: Journal of Convergence Information Technology, Vol. 5, No. 5, pp. 203 ~ 210, 2010

[4]Hai Guo， Jing-ying Zhao, "Segmentation Method for NaXi Pictograph Character Recognition ", JCIT: Journal of Convergence Information Technology, Vol. 5, No. 6, pp. 87 ~ 98, 2010 （Registered by EI）

[5]YI LU;M.Shridhar，Characters Segmentation in Handwritten Words-An Overview [Foreign Periodical] 1996(01)

[6]Richard G.Case;Eric Lecolinet，A Survey of Methods and Strategies in Character Segmentation

[7]O D Trier;A K Jain;T Taxt， Feature Extraction Methords for Character Recognition -A Survey [Foreign Periodical] 1996(04)