Research on Segmentation and Recognition of Printed Chinese Characters

Xi Wang, Hong Du, Xiping Wen

College of Communication Engineering, Chengdu University of Information Technology, Chengdu610225, China
College of Communication Engineering, Chengdu University of Information Technology, Chengdu610225, China
College of Communication Engineering, Chengdu University of Information Technology, Chengdu610225, China
wxcuit2018@163.com  duhong@cuit.edu.cn  starrwen@163.com

Abstract. The printed character recognition (PCR) is an efficient and automatic information entry method, which will become the important function interface of the computer in the future. A good character segmentation method can greatly improve the speed and accuracy of character recognition. This paper proposes a Chinese character segmentation algorithm based on OpenCV open source library and parallel processing, which can segment Chinese characters efficiently and accurately. And then, using the Tesseract-OCR character recognition engine to identify Chinese characters, the experimental results show that the average recognition time of a single Chinese character is 0.07s, and the recognition rate of Chinese characters is 96.2%, which proves the feasibility of the method.

1. Introduction

With the development of the information age, it has become the mainstream to use computers to organize, edit and save massive information on media such as disk. However, the contradiction between the low-speed input of massive data and the high-speed information processing has become the bottleneck of information interaction. Relying on manual input, labor intensity is high, error-prone and inefficient. So it is a good choice to recognize printed characters automatically and efficiently by computer, and the accurate and efficient character segmentation plays an important role in the process of recognition, which can greatly improve the efficiency of recognition.

In recent years, many methods have been proposed in the field of character segmentation and recognition. Yu Ming, Zhang Yanyun and others proposed a template segmentation algorithm for identifying handwritten Chinese characters of the same person[1]. Fu Lujing, Qian Junhao have proposed a Chinese character segmentation algorithm based on connected components of Chinese characters[2]. However, the accuracy of recognition is still lower than that of English characters and numeric characters. Therefore, how to improve the integrity of Chinese characters is further studied. Chen Nian, Huang Xiao and others proposed a Chinese character segmentation algorithm based on projection and structure information[3]. Liu Juning proposed a maximum backtracking method. Using the height of Chinese characters as a reference to determine the width of Chinese characters[4]. The above methods have some shortcomings in the efficiency or accuracy of Chinese character...
segmentation. This paper proposes a Chinese character segmentation algorithm; and then, the character recognition engine Tesseract-OCR is used to recognize Chinese characters.

2. The Segmentation of Chinese Characters

This paper presents a Chinese character segmentation algorithm based on OpenCV open source library and parallel processing, which greatly improves the segmentation speed and accuracy of printed Chinese characters. The sample image shown in Fig. 1. Including its enterprises registration number, enterprise name, enterprise scope and so on. The enterprise name in the sample image is divided and identified in this paper, that is, the selected part of the red box in the sample image.

In the experimental process, the sample image processing mainly includes the following steps: Firstly, inputting the sample images, then according to the computer configuration and the number of images to create multiple work threads automatically to maximize efficiency. Secondly, the sample image is preprocessed, including a series of operations such as de-watermarking, gray conversion, binarization, denoising and so on. Thirdly, using algorithms to achieve accurate segmentation of text lines. Fourthly, character interpolation is used to segment the identified content to achieve fast and
efficient segmentation of Chinese characters. Finally, Chinese characters are recognized by Tesseract-OCR character recognition engine. The flow chart of the whole experiment is shown in Fig.2.

2.1. Preprocessing images
Preprocessing usually includes noise removal, binarization, tilt correction, row, and column segmentation, smoothing and normalization etc.[5]. The binary processing of the business license image in the recognition process is to highlight the characteristics of the characters. In the binary process, according to the brightness distribution histogram of the sample image, using the two-peak method[6] to remove the watermarking and transform the sample image into black background and white characters. The result of binarization is shown in Fig.3.

2.2. Segmenting of the text lines
This paper takes the case of enterprise name segmentation and recognition. The text line segmentation is also a key step in character segmentation and recognition processes[7]. The method of resampling based on region pixel relation is used to obtain the line feature of the text. The gap between lines after
projection is used to segment the text, and then using HAAR features[8] to detect colon and accurately segment the contents to be identified. The process of text segmentation is shown in Fig.4.

Figure 4 the enterprise name, “Lingzhi Fashion Sales (Tianjin) Co. Ltd.”

2.3. Character interpolation and segmentation
In order to improve the efficiency of recognition, a vertical projection method[9] combined with multi-threshold segmentation method is proposed to obtain the brightness distribution characteristics of each character. Then the character is interpolated and segmented according to the principle of maximum backtracking method[10]. Which achieve the recognition rate of Chinese characters from 84% to 96.2%. Firstly, the text content to be processed is projected vertically; secondly, the position data of independent Chinese characters are recorded with the height of Chinese characters as a reference, and then several segmentation thresholds are selected according to the brightness integral value of projection for brightness value segmentation. Finally, each Chinese character is segmented one by one. The principle of character segmentation is shown in Fig.5.

Figure 5 The process of characters interpolation and segmentation of enterprise name

2.4. Character Recognition Based on Tesseract-OCR
This paper chooses the Tesseract-OCR engine to recognize the characters. The mainly steps is divided into font library training and character recognition.

Font Library Training: Considering the main work of this paper is the research of Chinese character segmentation algorithm. This paper chooses Tesseract-OCR official Chinese character library as experimental font library[11].

Character Recognition: Putting the Tesseract-OCR font library into the engineering catalog and through the program to achieve character recognition on the visual studio 2012 platform.

In this paper, the recognized characters are imported into the table. A part of the data results of recognition is shown in Fig.6. There are only a few recognition errors, such as the red box in the picture.
3. Experimental Results and Analysis

This paper collects 500 numbered linked sample images from WWW.TMALL.COM, which included 6423 Chinese characters in all enterprise name. The recognition result of all sample images is shown in Table 1, and the recognition result of Chinese characters in all enterprise name is shown in Table 2.

| Table.1 The recognition results of the picture |
|-----------------------------------------------|
| The number of sample images | The number of images correctly identified | The number of images mistakenly identified | The rate of recognition of picture |
| With process | 500 | 453 | 47 | 90.6% |
| Without process | | 253 | 247 | 49.4% |

| Table.2 The recognition results of Chinese character |
|--------------------------------------------------|
| The number of Chinese characters in all enterprise name | The total numbers of character | The number of characters correctly identified | The rate of recognition of Chinese characters |
| With process | 6423 | 6179 | 244 | 96.2% |
| Without process | | 5395 | 1028 | 84% |

This article uses Visual Studio 2012 as the development platform. Uses C++ as the development language, combining OpenCV open source library and Tesseract-OCR open source library to recognize printed Chinese characters.

From the Table.1 and Table.2, here are five hundred sample images with 6423 Chinese characters of all enterprise name, which 453 pictures and 6179 Chinese characters correctly recognized, the characters recognition rate in the enterprise name can reach 96.2%.

4. Conclusion

In the paper, a vertical projection method combined with multi-threshold segmentation method is proposed to obtain the brightness distribution characteristics of each character. Then the character is interpolated and segmented according to the principle of maximum backtracking method. Finally, the Tesseract-OCR open source character recognition engine is used to recognize Chinese characters. The experimental results show that the processing time of each Chinese character is about 0.07s, the
recognition rate of enterprise name can reach 96.2%, the recognition rate has reached the requirements of the practical application.

References

[1] MingYu, Yanyun Zhang, Cuihong Xue, Linjuan Sun. The Character segmentation of single Chinese characters in handwriting image [J]. Computer engineering and application, 2010, 46(09):180–182+245

[2] Lujing Fu, Junhao Qian, Yunfei Zhong. The method of printing image layout segmentation based on the connected component of Chinese characters [J]. Computer engineering and application, 2015, 51(05):178-182.

[3] Nian Chen, Xiao Huang. Chinese handwriting segmentation based on projection and structure information [J]. Journal of chizhou university, 2008(05):22-23+48.

[4] Jining Liu. Research and implementation of printed Chinese character recognition system [D]. Dalian University of technology, 2011.

[5] Fang Wang, Muning Kang, Xianguo Li. Printing Chinese character recognition technology [J]. Journal of information, 2004(02):32-33.

[6] Zengkang Zhang, Weihong Ma. Research on threshold segmentation of yarn image based on bimodal method [J]. Chemical fiber and textile technology, 2008, 47(03):34-38.

[7] Xiaojuan Yang Kai Song. Document image segmentation algorithm based on the projection method [J]. Journal of Chengdu University (natural science edition), 2009, 28 (02):139-141.

[8] Ping Shao, Luming Yang, yaorong Zeng. Improvement of integral image algorithm for calculating rotating Harr features [J]. Computer technology and development, 2006(11):146-147+181.

[9] Xiaojuan Yang, Kai Song. Document image segmentation algorithm based on the projection method [J]. Journal of Chengdu University (natural science edition), 2009, 28 (02):139-141.

[10] Bing Tian. Preprocessing of multi-font character recognition -- segmentation [J]. Journal of Beijing University of posts and telecommunications, 1991(04):82-86.

[11] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2007, pp. 629-633.