An Efficient Character Segmentation Algorithm for Printed Chinese Documents

Yuan Mei[1,2], Xinhui Wang[1,2], Jin Wang[1,2]

[1] Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science &Technology, Nanjing, Jiangsu, 210044,China

[2] School of Computer & Software, Nanjing University of Information Science &Technology, Nanjing, Jiangsu, 210044, China

Abstract. The character segmentation technology for printed documents is applied in many fields. This paper proposes an efficient character segmentation algorithm for Chinese printed documents, which is suitable for paper watermarking system. This algorithm is composed of three main steps: connected regions recognition, connected regions merging, and fine-gained segmentation, through what the algorithm succeeds in achieving Chinese character segmentation with high accuracy and high consistent segmentation between the digital version and print-scanned version of images from the same documents. Experiments show the effectiveness of the proposed algorithm.

Keywords: Printed document images; Chinese character segmentation; Connected region segmentation; Connected region merging

1 Introduction

Character segmentation of printed documents plays an important role in many fields, such as optical character recognition (OCR), identification for ticket information, recognition for zip code, automatic license plate recognition, and identification for printed circuit boards, as well as character labels on varieties of industrial components. Up to now, various methods have been proposed for complicated printed documents character segmentation, and can be classified as: *Projection-based segmentation methods* [1-3], in which vertical projection or histogram are used to locate reasonable split points between characters; *Recognition-based segmentation methods* [4-8] , which adapt prior knowledge to screen all possible segmentation schemes; *Feature extraction-based segmentation methods* [9-13], which segment and recognize the characters through different exacted features; *Skeleton analysis-based segmentation methods* [14-15], in which the skeletons of the characters are extracted for segmentation.

In this paper, we focus on the research of the segmentation algorithm for Chinese characters, which is applied in the paper watermarking system (this algorithm can also be applied to the printed Chinese character recognition or other related fields). The system requires an accurate and consistent character segmentation between the digital

version (i.e. digital images formed directly from the documents) and the print-scanned version (i.e. images scanned from printed paper documents) of the same documents, so as to embed the watermark information. In order to satisfy the requirements of the application, our algorithm needs to achieve the following purposes:

(1) Chinese character segmentation in the setting of mixed fonts.

(2) Maximize the consistency of segmentations between digital and print-scanned versions of images from the same documents.

(3) Independence on image resolution.

As for a complete character segmentation system, despite of the specific operation of character's segmentation, there are a series of pre-processing operations, including denoising, binarization, skew adjusting, and page segmentation. And in this paper, the character segmentation algorithm is proposed under the premise that all pre-processing operations have been completed.

The remainder of this paper is organized as follows: In Section Ⅱ, we introduce the principle and process of our algorithm. Section Ⅲ describes the specific process steps in detail, followed by Section Ⅳ, which presents simulation results. And we conclude the paper in Section Ⅴ.

2 Principle of Algorithm

In this section, we give an introduction to the principle and process of character segmentation algorithm.

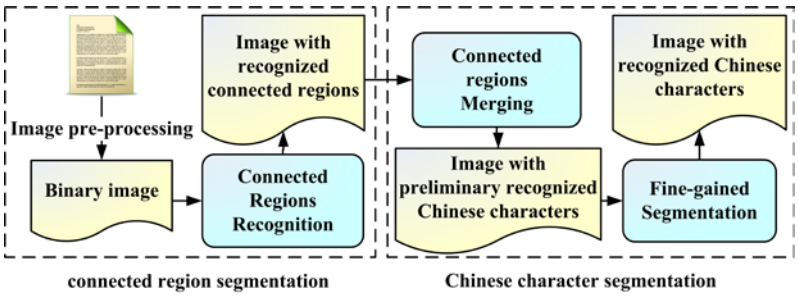


Fig. 1. Process of the Chinese character segmentation algorithm

Fig. 1 shows the basic process of our algorithm, which can be divided into two parts: the connected region segmentation stage, and the Chinese character segmentation stage.

A. Connected region segmentation

Connected region in this paper refers to the character part whose vertical integral is nonzero and continuous (on the assumption that background pixel is 0 and the text pixel is 1), as shown in Fig. 2. Different colors represents different connected region. The main tasks of this stage are position location and data acquisition (such as boundary coordinates, the value of dimensions, and the number of pixels) for

connected regions in the binary document images. In this stage, vertical projection is adopted as the segmentation algorithm, which will be elaborated in Section Ⅲ.

B. Chinese character segmentation

Fig. 2 shows that some of the connected regions we get from the first stage are not complete Chinese characters, but radicals or components of them. Therefore, the main task of this stage is to merge several connected regions in accordance with the rules we've set, and finally make them a complete Chinese character. The main work has two parts: 1)The merging of connected regions and the processing of special characters; 2) The precise segmentation of Chinese characters. The specific descriptions about these processes are in Section Ⅲ.

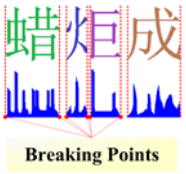


Fig. 2. Connected regions recognition through vertical projection

3 Proposed Scheme

This section will describe the process and steps of our algorithm in detail.

3.1 Connected region segmentation based on vertical projection

The main idea of this part is to locate the borders of the connected regions through the vertical projection. As shown in Fig. 2, since there are intervals between characters, the vertical integral of the entire text line is not continuous. These "breaking points" turn out to locate the positions of the connected regions' borders. The continuous part between dashed lines is the connected region. The steps of connected region segmentation for one text line are summarized as follows:

Step1: Implement the vertical projection for the text line and get the breaking points. According to whether the point represents a change from none-integral to continuous-integral or not, the left or right border of a connected region can be located and their positions are stored in arrays $col\_left[]$ and $col\_right[]$ respectively;

Step2: Locate and count all black pixels in the range of $[col\_left[i], col\_right[i]]$ (the left and right borders of the connected region), and store the position and amount information. Then, by looking for the minimum and maximum ordinate values from the position information, the positions of the connected regions' upper and lower borders are obtained. And the whole information we need is completely collected;

Step3: Repeat *Step2* until the segmentation of the entire text line is finished, and the connected regions are stored in array $char[]$.

After the implementation of the above steps for each text line, the connected regions' information of the entire document could be obtained.

3.2 Connected Regions Merging

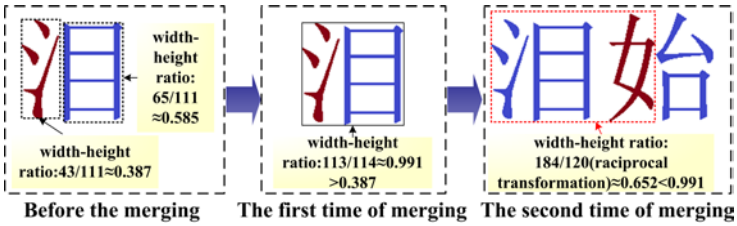


Fig. 3. Process of connected regions merging

3.2.1 Special Characters Processing

Before the merging of connected regions, some interference factors need to be taken into considerations, such as punctuations (，。、“”·) frequently appear in Chinese documents, as they are always different from normal Chinese characters in the location of the text line and the amount of pixels. Through these features, the special characters are easy to be recognized and labeled as “unavailable” afterwards, and will not be involved in the merging operation follows.

3.2.2 Connected Regions Merging

The main idea of the connected region merging is described as: Based on the fact that the width-height ratios of most Chinese characters are close to 1, for the combination of the chosen connected region and its next connected region, if its width-height ratio is more close to 1 than that of the chosen connected region, then it is recognized as the part of one Chinese character, as shown in Fig. 3. Specific rules and procedure are as follows:

Step1: Select the array $char[]$ for every text line. For each connected region $char[i]$ in this array, it will be merged if it matches the three rules as follows:

(a) The width-height ratio of the combination of $char[i]$ and $char[i+1]$ is more close to 1 than that of $char[i]$ (if the width-height ratio is bigger than 1, take reciprocal transformation).

(b) The distance between $char[i]$ and $char[i+1]$ is less than threshold s (the value of s is determined by experimental results).

(c) The connected region $char[i]$ is labeled as “available”.

Step2: After the first time of merging, the new connected region is obtained as $char[i]'$, and if the new one matches the three rules above, the second time of merging is taken, otherwise, it will be put into the array $m\_char[]$ and all the connected regions constituting the whole new $char[i]'$ are labeled as “unavailable”.

Step3: Repeat *Step1* and *Step2* until the merging of whole $char[]$ is finished, and the new array $m\_char[]$ is the preliminary segmentation result of Chinese characters.

3.3 Fine-gained Segmentation

Most of the Chinese characters are accurately separated after the preliminary segmentation, but there are also some special cases:

(1) The mistaken segmentation of high width-height ratio connected regions caused by adhesions. We solve this problem through three steps as follows:

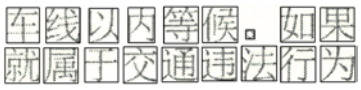
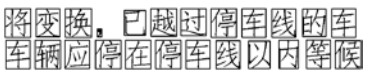
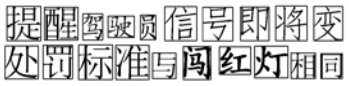
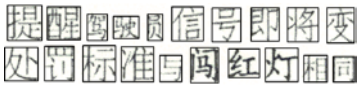
Step1: Calculate the width-height ratio of every connected region, if it's larger than the threshold we set, turn to *Step2*.

Step2: Judge the type of high width-height ratio through the height data, if it isn't caused by special characters like "一", then turn to *Step3*.

Step3: Find the weakest connections in the connected region and separate them through the vertical projection (only try its best for segmentation). Turn to *Step1*.

(2) For some Chinese characters with special structures, the whole character may be separated into different ones because the width-height ratio of one part is more close to 1 than that of itself. For instance, the "亻" of the Chinese character "悄" has a larger width-height ratio than "忄". In this case, we analysis the features of the connected region, if it matches the features like "亻", we force it not to be an independent Chinese character, but a part of its adjacent connected region.

Tab. 1. Segmentation results for document images in digital and print-scanned versions

Sample types	Digital document images	Print-scanned document images
Song ; 16-point		
Regular Script ; 14-point		
Fangsong ; 10.5-point		
Font: Mixed Size: Mixed		

4 Experimental Results and Performance Analysis

We randomly select some different documents as the origin of the digital and print-scanned images. Moreover, in each version of images, there are samples of different font types and sizes. The time of segmentation is about 1.5 seconds. Tab.1 shows the segmentation results of images of different font types and sizes in two versions, and both the resolutions of the digital and print-scanned images are 600dpi.

A. Accuracy and Consistency

Tab 2 shows the statistic data for segmentation accuracy and consistency of all experimental samples. As shown in Tab. 2, the proposed algorithm achieves the accuracy of about 99% for the character segmentation, and the consistency between digital and print-scan versions of document images reaches about 99.5%, which satisfy the watermarking system's requirements.

B. Resolution-independence

We also randomly have batch of samples scanned in 300dpi, and compare them with those in 600dpi, as shown in Tab. 3. Through the comparison, it is obvious that the segmentation of document images scanned in 300dpi are of lower accuracy, as they are more vulnerable to external factors. However, the accuracy is close to 99%. So, we can have the conclusion that our segmentation algorithm has a good performance on resolution-independence.

Tab. 2. Statistic data for accuracy and consistency

Sample types	Accuracy of digital versions	Accuracy of print-scanned versions	Consistency of two versions
Song ; 16-point	99.80%	99.73%	99.87%
Song ; 14-point	99.43%	99.16%	99.62%
Song ; 10.5-point	99.65%	99.20%	99.54%
Regular Script ; 16-point	99.71%	99.64%	99.90%
Regular Script ; 14-point	99.37%	99.18%	99.75%
Regular Script ; 10.5-point	99.65%	99.43%	99.78%
Fangsong ; 16-point	99.86%	99.79%	99.93%
Fangsong ; 14-point	99.49%	99.43%	99.81%
Fangsong ; 10.5-point	99.65%	99.57%	99.92%

Tab. 3. Accuracy of segmentation for document images scanned in different resolutions

Sample types	Scanned in 600dpi	Scanned in 300dpi
Song ; 16-point	99.73%	99.46%
Song ; 14-point	99.16%	98.87%
Song ; 10.5-point	99.20%	99.08%
Regular Script ; 16-point	99.64%	99.29%
Regular Script ; 14-point	99.18%	98.93%
Regular Script ; 10.5-point	99.43%	99.38%
Fangsong ; 16-point	99.79%	99.58%
Fangsong ; 14-point	99.43%	99.12%
Fangsong ; 10.5-point	99.57%	98.44%

5 Conclusion

In this paper, we propose an efficient characters segmentation algorithm for Chinese printed documents for the application in paper watermarking system. The experimental results show that the proposed algorithm achieved three goals: Chinese character segmentation in the setting of mixed fonts, maximizing the consistency of segmentations between digital and print-scanned versions of images from the same

documents and the independence on image resolution, which suit the requirements of paper watermarking system. In the subsequent research, we will concentrate on eliminating the influence on segmentation caused by numbers and punctuations, and improving the quality of damaged scanned images caused by external factors.

References

1. Wu, C.-D., FAN, Y.-q., ZHANG, Y.-z., LIU, M.: License Plate Character Segmentation Based on Differencing Projection and Preferably Segmented Character. Journal of Northeastern University(Natural Science), Vol.29(2008) 920-923.
2. Wenzhe W: A Method of Characters Segmentation and Its Application in Digital Textual Material Repairment. New Technology of Library and Information Service, Vol.3 (2010) 82-85.
3. Shuying C, Yongjie Z, Shijie D: Research on method of train's code image segmentation based on feature of space and the vertical projection. Journal of Hebei University of Technology, Vol.40(2011) 59-61.
4. Yungang Z, Changshui Z: Segmenting Characters of License Plate by Hough Transformation and the Prior Knowledge. Chinese Journal of Computers , Vol.27(2004) 130-135.
5. Chengyong Z, Hong L: Vehicle license plate characters segmentation method using character's entirely and blob analysis. Journal of Huazhong University of Science and Technology(Natural Science Edition) , Vol.38(2010) 88-91.
6. Jingming G, Yunfu L: License Plate Localization and Character Segmentation With Feedback Self-Learning and Hybrid Binarization Techniques. IEEE Transactions on Vehicular Technology, Vol.57(2008) 1417-1424.
7. Anagnostopoulos C-N, Anagnostopoulos IE, Psoroulas ID, Loumos V, Kayafas E: License Plate Recognition from Still Images and Video Sequences: A Survey. IEEE Transactions on Intelligent Transportation Systems, Vol.9(2008) 377-391.
8. Senapati, D., Rout, S., Nayak, M: A Novel Approach to Text Line and Word Segmentation on Odia Printed Documents. IEEE Third International Conference on Computing Communication & Networking Technologies (ICCCNT). (2012) 1-6.
9. Yongzhong L, Yulei W, Zhenzhen L: Study on Printed Tibetan Character Recognition Technology. Journal of Nanjing University(Natural Sciences), Vol.48(2012) 55-62.
10. Khurshid, K., Faure, C., Vincent, N.: Word spotting in historical printed documents using shape and sequence comparisons. Pattern Recognition, Vol.45(2012) 2598-2609.
11. Liying Zheng, Abbas H. Hassin, Xianglong Tang: A new algorithm for machine printed Arabic character segmentation. Pattern Recognition Letters, Vol.25(2004)1723-1729.
12. Ota,T, Wada,T: Classification based character segmentation guided by Fast-Hessian-Affine regions. IEEE First Asian Conference on Pattern Recognition (ACPR). (2011)249-253.
13. Chen, Z.-X., Liu, C.-Y., Chang, F.-L., Wang, G.-Y.: Automatic License-Plate Location and Recognition Based on Feature Salience. IEEE Transactions on Vehicular Technology, Vol.58(2009) 3781-3785.
14. Jianxiong Guo, Lihua Yang: Approach to Segment Multi-Size Machine Printed Characters by Removing Serifs. Pattern Recognition and Artificial Intelligence, Vol.19(2006) 702-707.
15. Nikolaou, N., Makridis, M., Gatos, B., Stamatopoulos, N., Papamarkos, N.:Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths. Image and Vision Computing, Vol.28(2010) 590–604.