A Chinese Character Segmentation Algorithm for Complicated Printed Documents

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

[1]*Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science &Technology, Nanjing, 210044, China*
[2]*School of Computer & Software, Nanjing University of Information Science &Technology, Nanjing , 210044, China*

Abstract

The character segmentation technology for printed documents plays an important role in optical character recognition, ticket information identification, postal code identification, automatic license plate recognition and so on. In this paper, a Chinese characters segmentation algorithm for complicated printed documents is proposed for the application in paper watermarking system. In this application, the algorithm aims to achieve high accuracy Chinese character segmentation and high consistent segmentation between the digital version images and print-scanned version images for the same documents. In this method, three main steps are included: connected regions recognition, connected regions merging, and fine-gained segmentation. 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* [1-5], in which vertical projection or histogram are used to locate reasonable split points between characters; *Recognition-based segmentation methods* [6-12], which adapt prior knowledge to screen all possible segmentation schemes; *Feature extraction-based segmentation methods* [13-18], which segment and recognize the characters through different exacted features; *Skeleton analysis-based segmentation methods* [19-21], 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. In a real document file, characters are always presented in different font types, different sizes, and sometimes bold or italic. Through software or print-scan processing, these document files will be converted into images with various features and objects. Compared with the existed researches on objects with simplified features, the algorithm in this paper concentrates on objects which are more complex and closer to the actual conditions. And in order to satisfy the requirements of paper watermarking system, the algorithm needs to achieve adaptive segmentation for multiple objects at the same time.

(2) Maximize the consistency of segmentations between digital and print-scanned versions of images from the same documents. When embedding watermarks, the carriers are commonly digital images of low noise or always without noise; but while extracting, the objects are images converted from one or several times of print-scan processing, which have normal geometric deformation and serious noise pollution caused by printers, scanners, and man-made factors (such as folding, dirt, scratches, etc.). Hence, the segmentation algorithm needs to solve different problems in different situations. Although fault-tolerant coding could be a solution for the problem of consistency in paper watermarking system, the co-operation with character segmentation of high consistency is indispensable.

(3) Independence on image resolution. The resolution of scanned images depends on uncontrollable factors like hardware features of scanner, users' choice etc. Therefore, resolution-independence is another requirement in our algorithm design.

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.

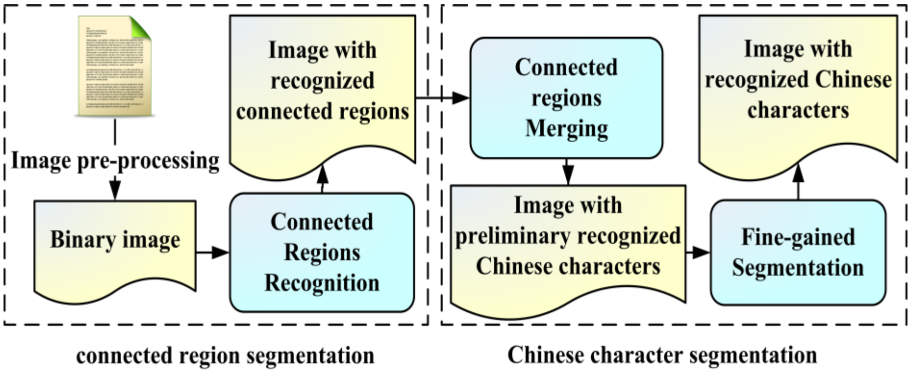
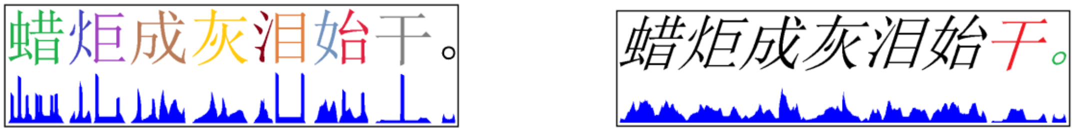


Image with recognized connected regions
Connected regions Merging
Image with recognized Chinese characters
Image pre-processing
Binary image
Connected Regions Recognition
Image with preliminary recognized Chinese characters
Fine-gained Segmentation
connected region segmentation
Chinese character segmentation

Figure 1. Process of the Chinese Character Segmentation Algorithm



(a) Connected regions and their vertical projections

(b) Vertical projection of italic characters

Figure 2. Connected Regions Recognition through Vertical Projection

2. Principle of Algorithm

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

Figure 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 Stage

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 Figure 2(a). 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 Stages

Figure 2 (a) 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.

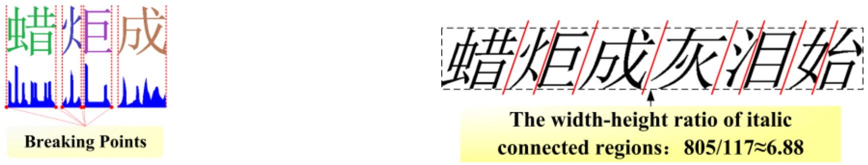
The main work has two parts:

1) The merging of connected regions and the processing of special characters.

The algorithm and rules for connected regions' merging will be introduced in Section. Special characters here refer to characters much more different from the Chinese box-shaped characters, such as punctuation (，。、“”·), strokes of Chinese characters and those shaped like punctuations, which will affect the subsequent operation of character segmentation. Hence, after preliminary processing of connected regions, these special characters should be identified and treated specially in the following merging operation. In addition, the vertical projection may cause the italic characters being divided into a connected region, as shown in Figure 2(b). So, the identification and re-segmentation of the italic region are needed to correct the error appears here.

2) The precise segmentation of Chinese characters.

The algorithm in this paper is not only suitable for ordinary digital images, but also suited to the scanned images. As scanned images are susceptible to external factors like noise and scratches, which finally causes the adhesions or broken strokes in the images, then after preliminary segmentation of the Chinese characters, incorrect segmentations caused above should be handled. The specific algorithm process is also shown in Section Ⅲ.


蜡炬成
Breaking Points
蜡炬成灰泪始
The width-height ratio of italic connected regions：805/117≈6.88

(a) Borders located by vertical projection (b) Italic connected regions segmentation through oblique projection

Figure 3. Connected Regions Segmentation

3. Proposed Scheme

This section will describe the process and steps of our algorithm in detail, according to the principle presented in Section.

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 Figure 3(a), 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, which are depicted in red dashed lines in Figure 3(a). 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 array $col\_left[\,]$ and array $col\_right[\,]$ respectively;

Step2: Locate and count the numbers of 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 in corresponding arrays. 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. At this moment, the connected regions’ information has been 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

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 these characters are always different from normal Chinese characters in the location of the text line and the amount of pixels, they are then called special characters in this paper. Thanks to these features, the special characters are easy to be recognized and labeled as “unavailable” afterwards. All the connected regions labeled as “unavailable” are not involved in the merging operation follows.

It can be seen from Figure 2(b) that the vertical projection may cause a mistaken segmentation for multiple italic characters. In this paper we implement a re-segmentation for these mistaken connected regions with oblique projection. As these connected regions usually have high width-height ratio and large amount of pixels, as shown in Figure 3(b), they can be recognized while the segmentation of connected regions, and be processed by an oblique projection (the red lines in Figure 3(b)). The slope of the oblique projection is obtained through the statistic data.

Nevertheless, it is need to be noticed that not all connected regions with high width-height ratio are created by italic characters, some special Chinese characters like “一” may be the reason and the adhesions of multi-character caused by noise or scratches may also lead to this phenomenon.

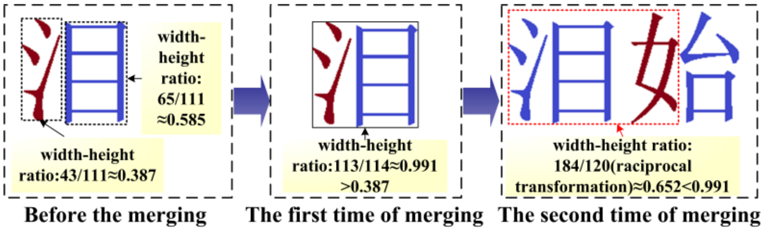


Figure 4. Process of Connected Regions Merging

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, while the chosen connected region and its next connected region are not treated as independent connected regions any more, as shown in Figure 4. 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 (the step only aims to try 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, which means a part of the character is identified as a real Chinese character. For instance, the “丶” of the Chinese character “悄” has a larger width-height ratio than “忄” when its font is “Fangsong”. In this case, we analysis the features of the connected region, if it matches the features like “丶”, we force it not to be an independent Chinese character, but a part of its adjacent connected region.

(3) Affected by the shade of printing and man-made scratches on papers, the whole Chinese character may also be separated into several ones, when we take the segmentation for scanned images. However, compared to the distances among characters, the distances among these separated parts are much shorter, and the width or height values are much smaller too. So, to take full advantage of these features, we can set another distance threshold here, once the distance between two connected regions is smaller than the threshold and their sizes are smaller than normal Chinese characters, the two connected regions are judged to be of the same one, regardless of the width-height ratio.

4. Experimental Results and Performance Analysis

In this section, we demonstrate a series of experimental results to evaluate the proposed technology. We randomly select some different documents as the origin of the digital and print-scanned images, in other words, every document has two versions of image. Moreover, in each version of images, there are samples of different font types (Song, Regular Script and Fangsong) and sizes (16-point, 14-point and 10.5-point). The whole experiment is implemented by HP LaserJet M1530 MFP Series PCL 6, and 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

Table 2 shows the statistic data for segmentation accuracy and consistency of all experimental samples. As shown in Table 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. Through the analysis of samples, we find that most of the mistaken segmentations are caused by numbers and punctuations which are hardly to be distinguished from normal Chinese characters(like （ and 《). Otherwise, external factors like shade of printing and paper pollution also lead to the mistaken segmentations.

B. Resolution-independence

We also randomly have batch of samples scanned in 300dpi, and compare them with those in 600dpi, as shown in Table 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.

Table 1. Segmentation Results for Document Images in Digital and Print-scanned Versions

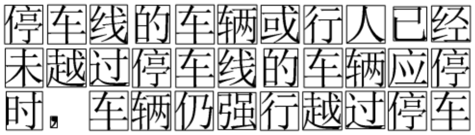
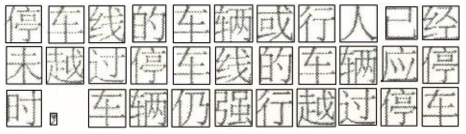
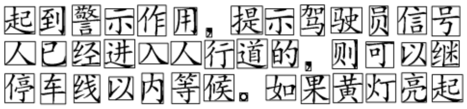
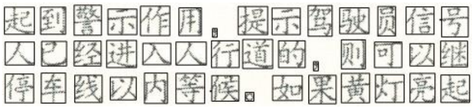
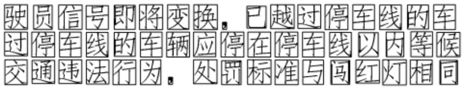
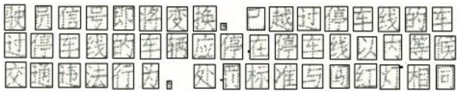
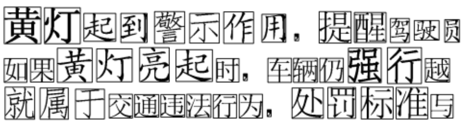
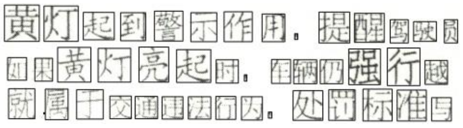
Sample types	Digital document images	Print-scanned document images
Font: Song ; Size: 16-point	 停车线的车辆或行人已经 未越过停车线的车辆应停 时，车辆仍强行越过停车	 停车线的车辆或行人已经 未越过停车线的车辆应停 时，车辆仍强行越过停车
Font: Regular Script ; Size: 14-point	 起到警示作用，提示驾驶员信号 人已经进入人行道的，则可以继 停车线以内等候。如果黄灯亮起	 起到警示作用，提示驾驶员信号 人已经进入人行道的，则可以继 停车线以内等候。如果黄灯亮起
Font: Fangsong; Size: 10.5-point	 驶员信号即将变换，已越过停车线的车 过停车线的车辆应停在停车线以内等候 交通违法行为，处罚标准与闯红灯相同	 驶员信号即将变换，已越过停车线的车 过停车线的车辆应停在停车线以内等候 交通违法行为，处罚标准与闯红灯相同
Font: Mixed; Size: Mixed	 黄灯起到警示作用，提醒驾驶员 如果黄灯亮起时，车辆仍强行越 就属于交通违法行为，处罚标准与	 黄灯起到警示作用，提醒驾驶员 如果黄灯亮起时，车辆仍强行越 就属于交通违法行为，处罚标准与

Table 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%

Table 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 a Chinese characters segmentation algorithm for complicated printed documents for the application in paper watermarking system. The algorithm achieves high accuracy Chinese character segmentation and high consistent segmentation between the digital version images and print-scanned version images for the same documents with three main steps, including: connected regions recognition, connected regions merging, and fine-gained segmentation. 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 how to eliminate the influence on segmentation caused by numbers and punctuations, and how to improve the quality of damaged scanned images caused by external factors like shade of printing.

Acknowledgements

This work is supported by National Science Foundation of China (No:61202138), China Postdoctoral Science Foundation (No:20110491413), Natural science fund for colleges and universities in Jiangsu Province (No:11KJB520012), JiangSu province Postdoctoral Science Foundation (No: 1102051C). This work was also supported by the Industrial Strategic Technology Development Program (10041740) funded by Ministry of Knowledge Economy (MKE）Korea, and by Natural Science Foundation of Jiangsu Province (No. BK2012461).

References

[1] W. Chengdong, F. Yuquan, Z. Yunzhou and L. Meng, “License Plate Character Segmentation Based on Differencing Projection and Preferably Segmented Character”, Journal of Northeastern University (Natural Science), vol. 29, no. 7, **(2008)**.

[2] W. Wenzhe, “A Method of Characters Segmentation and its Application in Digital Textual Material Repairment”, New Technology of Library and Information Service, vol. 3, **(2010)**.

[3] C. Shuying, Z. Yongjie and D. Shijie, “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, no. 2, **(2011)**.

[4] L. Zhen and L. Jianfeng, “Segmentation and Recognition of Printed Character in Check Image”, Computer Engineering, vol. 29, no. 9, **(2003)**.

[5] Z. Zhen, H. Shan, L. Daizhang and Y. Guoli, “A Method Applied for Precise Segmentation of the Characters in the ID Card”, Computer Engineering and Applications, vol. 13, **(2003)**.

[6] Z. Yungang and Z. Changshui, “Segmenting Characters of License Plate by Hough Transformation and the Prior Knowledge”, Chinese Journal of Computers, vol. 27, no. 1, **(2004)**.

[7] Z. Chengyong and L. Hong, “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, no. 3, **(2010)**.

[8] M.-C. Jung, Y.-C. Shin and S. N. Srihari, “Machine Printed Character Segmentation Method using Side Profiles”, IEEE International Conference on Systems, Man, and Cybernetics, vol. 6, **(1999)**.

[9] G. Jingming and L. Yunfu, “License Plate Localization and Character Segmentation with Feedback Self-Learning and Hybrid Binarization Techniques”, IEEE Transactions on Vehicular Technology, vol. 57, no. 3, **(2008)**.

[10] C.-N. E. Anagnostopoulos, I. E. Anagnostopoulos, I. D. Psoroulas, V. Loumos, E. Kayafas, “License Plate Recognition from Still Images and Video Sequences: A Survey”, IEEE Transactions on Intelligent Transportation Systems, vol. 9, no. 3, **(2008)**.

[11] Y. Sicong, W. Qing and Z. Rongchun, “An Optimal Character Segmentation Algorithm Based on Connected Component Recognition”, Journal of Image and Graphics, vol. 11, no. 1, **(2006)**.

[12] D. Senapati, S. Rout and M. Nayak, “A Novel Approach to Text Line and Word Segmentation on Odia Printed Documents”, IEEE Third International Conference on Computing Communication & Networking Technologies (ICCCNT),**(2012)**.

[13] L. Yongzhong, W. Yulei and L. Zhenzhen, “Study on Printed Tibetan Character Recognition Technology”, Journal of Nanjing University (Natural Sciences), vol. 48, no. 1, **(2012)**.

[14] K. Khurshid, C. Faure and N. Vincent, “Word spotting in historical printed documents using shape and sequence comparisons”, Pattern Recognition, vol. 45, no. 7, **(2012)**.

[15] R. Azmi and E. Kabir, “A new segmentation technique for omnifont Farsi text”, Pattern Recognition Letters, vol. 22, **(2001)**.

[16] L. Zheng, A. H. Hassin and X. Tang, “A new algorithm for machine printed Arabic character segmentation”, Pattern Recognition Letters, vol. 25, **(2004)**.

[17] T. Ota and T. Wada, “Classification based character segmentation guided by Fast-Hessian-Affine regions”, IEEE First Asian Conference on Pattern Recognition (ACPR), **(2011)**.

[18] Z. Chen, C. Liu, F. Chang and G. Wang, “Automatic License-Plate Location and Recognition Based on Feature Salience”, EEE Transactions on Vehicular Technology, vol. 58, no 7, **(2009)**.

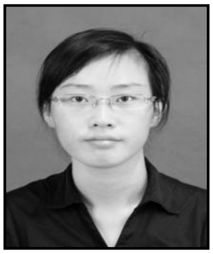
[19] L. Da, P. Wei and X. Mingpei, “Topographic Approach Recognizer for Merged Character Images Based on Skeletonization”, Computer Engineering and Applications, vol. 2, **(2002)**.

[20] G. Jianxiong and Y. Lihua, “Approach to Segment Multi-Size Machine Printed Characters by Removing Serifs”, Pattern Recognition and Artificial Intelligence, vol. 19, no. 6, **(2006)**.

[21] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos and N. Papamarkos, “Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths”, Image and Vision Computing, vol. 28, **(2010)**.

Authors

Yuan Mei received the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science & Technology in 2009, China. He is currently working at the school of computer and software, NanJing University of Information Science & Technology. His research interests include fingerprint recognition, image processing and machine learning.



Xinghui Wang received the Bachelor degree in software engineering from Nanjing University of Information Science & Technology in 2012, China. She is now carrying on her master education in Nanjing University of Information Science & Technology. Her research interest is the area of information security.

Jin Wang received the B.S. and M.S. degree from Nanjing University of Posts and Telecommunications, China in 2002 and 2005, and Ph.D. degree from Kyung Hee University Korea in 2010. During 2010-2011, he was a post-doctor in Kyung Hee University Korea. Now, he is a professor in School of Computer and Software, Nanjing University of Information Science and Technology. His research interests mainly include routing algorithm/protocol design, performance evaluation and optimization for wireless ad hoc and sensor networks. He is a member of the IEEE and ACM.