

Captcha Automatic Segmentation and Recognition Based on Improved Vertical Projection

Lili Zhang¹, Yuxiang Xie¹, Xidao Luan², Jingmeng He¹

¹National University of Defense Technology, Changsha, 410073, China

²Changsha University, Changsha, 410073, China

e-mail: zhanglili12@nudt.edu.cn

Abstract—Captcha recognition plays an important role in Internet security, and conglutination characters segmentation is still the bottleneck of captcha recognition research. In this paper, a method of captcha segmentation based on improved vertical projection can efficiently solve the segmentation problem of different types of conglutination characters in captcha combining both numbers and letters, so as to improve the accuracy of captcha recognition. Firstly, take preprocessing to captcha image including removing interfering background, denoising and binarization, getting binary captcha image with less noise. Secondly, apply improved vertical projection method to captcha characters segmentation, propose targeted segmentation method to different conglutinate type characters, getting split single character image. Thirdly, take the overlap rate of sample and template character pixels as matching rate, using template matching algorithm for recognition. Finally, through program experiment to achieve the algorithms of segmentation and recognition proposed above. Experimental result suggests that the method proposed in this paper has efficient segmentation effect on conglutinate characters, consequently improving the accuracy of captcha recognition.

Keywords—captcha; conglutinate characters; segmentation; recognition

I. INTRODUCTION

As a widely used means to authenticate users, captcha plays an important role in Internet security, but there are still some loopholes. Study of captcha recognition can found and improve loopholes in time, test security of various captcha and is helpful to design much more safe and reliable captcha. So far, captcha mainly has the following several kinds: conglutination, distortion character, character of fuzzy, background interference, italic characters, and combination of these types and so on. The research idea of captcha recognition generally is the method of first segmentation then recognition [1]. The segmentation effect has direct influence on recognition accuracy. In the study of captcha recognition, segmentation is much more difficult than recognition. Once the characters are divided, recognition problem can be easily solved by machine learning or other algorithms. Since each website's captcha has different noise in background and different character types, it's unable to design a recognizer can identify all captchas automatically yet. Essentially it has not yet developed a generic captcha character segmentation method.

From the point of the research status at home and abroad, the more commonly used captcha characters segmentation algorithms are: drop algorithm, vertical projection [2], the segmentation method based on clustering analysis [3-4]. In 1995, Congedo G firstly proposed drop algorithm [5], using it to solve the segmentation problem of handwritten conglutination numeric characters by simulating rolling path of water drop from high to low. Water dripping down from the top string along the character profile, can avoid the lost of character information caused by linear cutting. Traditional vertical projection is mainly applied on segmentation of not conglutinate characters, by looking for pixel projection valued zero to determine the location of a gap between the adjacent characters then determine the segmentation point. In 1996, Biensaid A. M applying clustering analysis algorithm in image segmentation field [6], in captcha characters segmentation, clustering analysis algorithm is mainly through agglomerating the captcha image pixels of characters into classes, divided them into different clusters, so as to achieve the effect of the segmentation character. But these algorithms can only be applied for a certain conglutination type of characters, unable to solve a variety of conglutination types hybrid captcha character recognition problem, the generality is not strong. Therefore, this paper puts forward an improved vertical projection method, which can automatically identify and segment different conglutination types captcha characters, and to obtain the good segmentation effect.

II. RECOGNITION PROCESS

In this paper, captcha recognition process is shown in Fig. 1: firstly read captcha images in the dataset, and gray them. Captcha image preprocessing part mainly includes the circle detection based on Hough transform algorithm, through color reversal inside the circle to remove interference, denoise on the residue interference based on median filtering algorithm and binarization of the image. Character segmentation part is mainly on the basis of vertical projection, distinguish the conglutinate type of characters before segmentation, then according to the conglutinate type, divide the segmentation into segmentations of not conglutination, part conglutination and high conglutination characters. Apply template matching method to captcha automatic recognition. First extract reference template, then recognize characters, finally get the results.

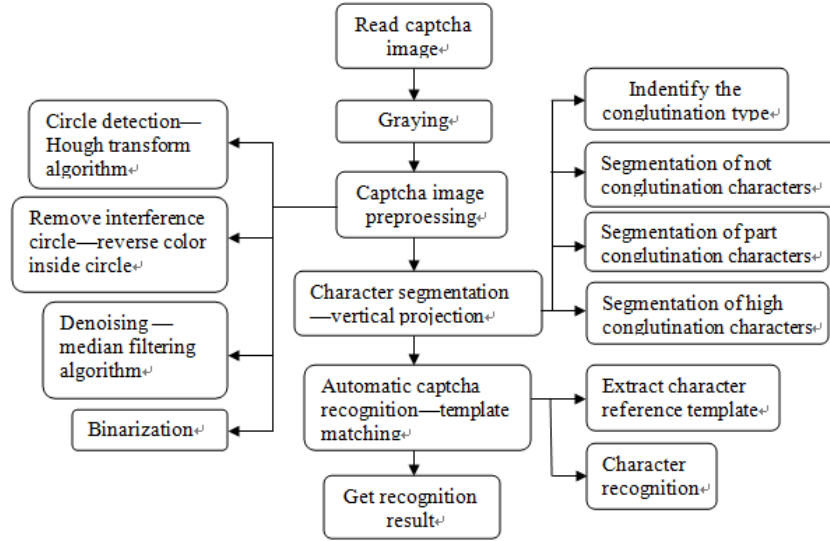


Figure 1. Recognition process

III. IMPLEMENT METHOD

A. Preprocess

The captcha dataset used in this paper came from ‘the Second National Big Data Technology Competition—Captcha Recognition Competition’ undertaken by DataCastle. This paper selected the dataset containing 20000 captcha images and randomly selected 100 pieces of image from the dataset for experiment. Sample pictures are shown in Fig. 2.



Figure 2. Captcha images

We can see from Fig. 2 in this paper, the captcha we study has the following features:

- 1) Background interference are the black circles, the overlap part of the circle and the characters inside circle is white, the characters outside of the circle is black;
- 2) The characters are the mixture of capital English letters and Numbers.

- 3) Characters are for regular italic font, there is no distortion and irregular etc;
- 4) The width of the English letters and Numbers are different;
- 5) There are some smaller, partial conglutination between adjacent characters.

According to the characteristics of the captcha image above, take pretreatment operation including removing interference, denoising and binarization to the captcha image.

1) *Remove circular interference based on hough algorithm*

To remove the circular interference in background, first by Hough transform algorithm [7] detect the black interference circle in captcha images. Then remove the interference circle and leave the characters inside circle at the same time, reverse the color of interference circle, make the picture only complete black character with white background, so as to achieve the aim of removing interference circle.

Hough circle detection results are shown in Fig. 3:



Figure 3. Hough circle detection result



Figure 4. Removing circular interference effect



Figure 5. Denoising effect

The effect of removing circular interference are shown in Fig. 4.

From Fig. 4 we can see that after removal of interference circle, there are part of residual outer of interference circle remaining in captcha images, causing disturbance to the captcha recognition. Therefore the captcha images need denoising processing.

2) Denoise by median filtering algorithm

Median filtering algorithm [8] is using the gray value of adjacent pixels to replace the gray value of one pixel, is a typical sort of filter. Through the median filtering algorithm for denoising of residual interference round contour in the captcha image, the effect is shown in Fig. 5.

From Fig. 5 we can see that after denoising by median filtering algorithm, most part of residual interference circle outline has been removed, leaving the characters much more clean and clear and interference with captcha character recognition greatly reduced.

3) Binarization

Digital image data can be represent by matrix, generally using matrix theory and matrix algorithm for digital image analysis and processing. To transform captcha image into 0-1 matrix, first of all do the image binarization processing, set the gray value of character pixels to 255 and the rest of the pixels' gray value should be set to 0. Then set the matrix element corresponding to the gray value of 255 (white) pixels to 0, the matrix element corresponding to the gray value of 0 (black) pixels to 1, then get the 0-1 matrix of captcha image.

At this point captcha image preprocessing is done.

B. Automatic Segmentation of Different Types of Conglutination Characters

Captcha characters segmentation effect directly affects the accuracy of character recognition, the information of single character segmented can't lack or be interfered, otherwise it will lead to the subsequent character recognition errors [9]. The automatic segmentation of different conglutination types of characters is the bottleneck of captcha study. Therefore this paper combines the location, size features of captcha characters and the vertical projection histogram together, put forward the Improved vertical projection method, discuss the segmentation methods of not conglutination, part conglutination and high conglutination characters, part conglutination is divided into edge tangent conglutination and edge horn conglutination of two types.

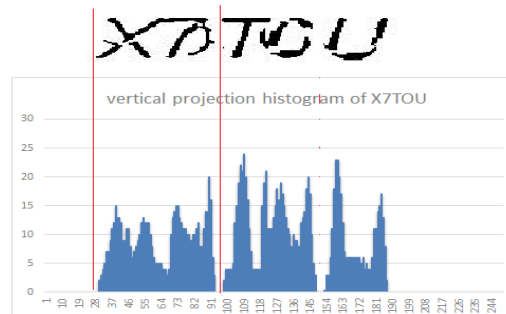
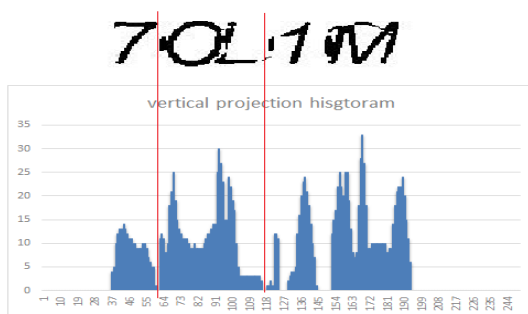


Figure 7. Vertical projection histogram of part conglutination

1) Identify the conglutination type

Longitudinal scanning image matrix of captcha images, get the vertical projection histogram of captcha image. If there is point of projection valuing 0 exiting between adjacent characters and the point locates reasonably (the point can divide the projection area into two areas of a character's width), it means that there is column blank space between adjacent characters, namely the adjacent characters are not conglutination characters. As shown in Fig. 6, the position that the arrow points to projection values 0, and is a gap between the characters.

If there is no point of projection valuing 0 between adjacent characters, it means there is some degree of conglutination between the adjacent characters. As shown in Fig. 7, there is no position of 0 projection value between adjacent characters O, L and X, 7, namely that they are both conglutination characters. If in the conglutination area of vertical projection histogram exits the highest peaks, and deviation between the position of the peak and the middle position of two characters is less than 5 pixels, then the conglutination characters are edge tangent conglutination characters. If in the conglutination area of vertical projection histogram exits the lowest trough, and deviation between the position of the trough and the middle position of two characters is less than 5 pixels, then the conglutination characters are edge horn conglutination characters. From Fig 7 we can see that adjacent characters O, L are edge tangent conglutination characters, adjacent characters X, 7 are edge horn conglutination characters. Both edge tangent conglutination and edge horn conglutination are part conglutination characters.

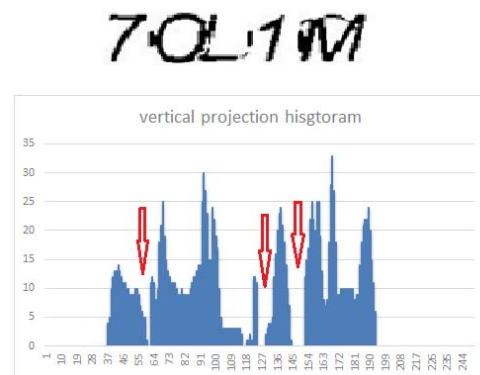


Figure 6. Vertical projection histogram of captcha image

If there is no obvious characteristics in the conglutination area of vertical projection histogram, the adjacent characters will be judged to be high conglutination, namely high degree of conglutination and irregular. As is shown in figure 8, in the vertical projection histogram, there is no obvious highest peak and lowest trough appearing in the projection area of conglutination characters W, F, therefore they are judged to be high conglutination characters.

2) *Segmentation of conglutination characters*
After identifying the conglutination type of character, find out the segmentation point according to the distribution features of captcha characters in vertical projection histogram: the segmentation point of not conglutination characters is the position of projection valuing 0, the segmentation point edge tangent conglutination characters is the position of the peak in adjacent characters' projection area, the segmentation point of edge horn conglutination characters is the position of trough in adjacent characters' projection area, the segmentation point of high conglutination is the middle position of adjacent characters' projection area. The specific segmentation steps are as follows:

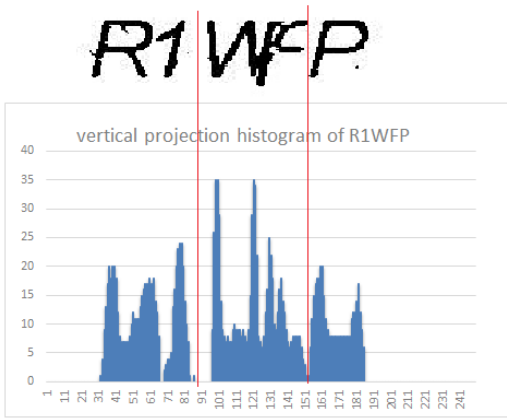


Figure 8. Vertical projection histogram of high conglutination characters

1) Add up the element of the same column in captcha image 0-1 matrix, get a one- dimensional array, named as col[], each element value of the array corresponding to the total number of black pixels in each column.

2) Iterate through array col[], the sequence number of elements value 0 represent for blank column's corresponding abscissa, record it to a one-dimensional array, named as p[].

3) Iterate through p[], orderly do p[j]-p[j-1], getting the distance of the most close two points of projection valuing 0, and compare with the average width d of characters. If p[j]-p[j-1]<d, it suggests that these two points are in the same blank area. If 2d> p[j]-p[j-1]>d, it suggests that the points in position p[j] and p[j-1] belong to two different blank areas, then (p[j-1], p[j]) determines one character area and is the segmentation point of not conglutination characters.

4) If p[j]-p[j-1]>2d, it suggests that there is point of projection not valuing 0 exiting the two characters width area, there is no blank gap between adjacent characters, namely these adjacent characters are conglutination characters.

Iterate through the value of col[i], (i ∈ (p[j-1], p[j])), find out the maximum value max, compare the value of i corresponding to max minus p[j-1] with character width d, if they are nearly equal, then i will be the segmentation point of edge tangent conglutination characters.

Iterate through the value of col[i], (i ∈ (p[j-1], p[j])), find out the minimum value min, compare the value of i corresponding to min minus p[j-1] with character width d, if they are nearly equal, then i will be the segmentation point of edge horn conglutination characters.

If the above two cases are not set up, it will be high conglutination character, then adopt the method of average cutting, reserve the characteristics of single character information to the most degree.

C. Character Recognition

For the captcha studied in this paper is made up of 10 numbers and 26 letters mixed, and with unified, standardized font and no distortion. In order to reach the goal of being concise and efficient, take template matching method for character recognition after segmentation [10].

TABLE I. CAPTCHA EXPERIMENTAL TREATMENT EFFECT

No.	Original image	Filter interference	Denoising	Segmentation effect	Recogniti-on result
1					7CL1M
2					X7T0U
3					R1WFP
.....
99					2P5N3
100					97XDJ

1) Extract character reference template

Through preprocessing and segmentation for captcha images in dataset, get complete character template with no fragmentary, no noise and no interference, as shown in Fig. 9:

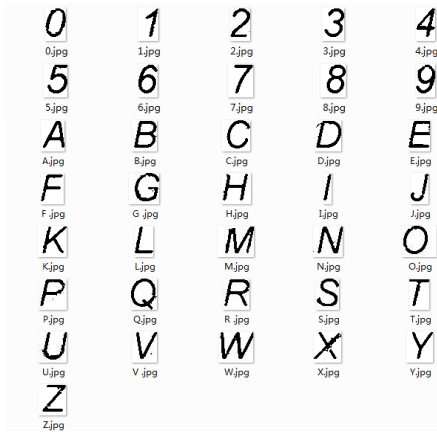


Figure 9. Character template

2) Translation and alignment for the character area of sample and template

After segmentation we get single character images, the positions of each character in the image are different, that is to say the distances between character and image edge are different. So before template matching, the sample must be translated and directed to template, guaranteeing the maximum overlap of the character in sample and template. Respectively store the pixels coordinate information of single character image and template in arrays `pieces[]` and `mods[]` in the order of from left to right, top to bottom. The difference of the two abscissas and ordinates in the first element of two arrays are the horizontal and vertical distance x and y of sample and template translate. The abscissas and ordinate of each element from the `mods[]` respectively plus x , y , so as to achieve the alignment effect of character template and sample.

3) Calculate matching degree of samples

After the alignment of sample and template, calculate the overlap degree of characters, namely matching degree of sample and template. Iterate through the elements in characters pixels coordinate arrays `pieces[]` and `mods[]` of sample and templates. The same element in these two arrays means that the pixels of sample and template are overlapped at this coordinate. Iterate through the elements of these two arrays, find out all overlapped pixels, name the total number as n and name the total number of character pixels of template as N , then the matching degree of sample and

template is. Match each single character sample with all templates, the template character with the highest matching degree is the recognition result of the sample.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Realize the automatic captcha recognition algorithm proposed by this paper in Python environment. Experiment with the captcha images dataset, the experimental results are shown in Table I.

As shown in Table I, in the recognition result of the first captcha image, character 'O' was wrong recognized as 'C', this is because character 'O' and 'L' are edge tangent characters, character 'O' lost some information in segmentation position when segmenting, leading to the broken character and being wrong recognized as 'C' when template matching. It suggests that the segmentation and recognition algorithm of this paper need to be further optimized and improved.

Take preprocessing measures including graying, denoising and binarization to the captcha images of other websites, then apply the improved vertical projection method for character segmentation proposed in this paper to the captcha recognition problem in other websites. The results are shown in Table II:

From Table II we can see that, the improved vertical projection method for segmentation proposed in this paper has good effect on other websites' captchas, and has better applicability, which laid the foundation for the subsequent recognition work. The statistics results are shown in Table III:

The captcha image dataset used in this paper is shown in the No. 1 picture, with more interference noise and more complicated conglutination, the accuracy of recognition is 92.3%. In the No. 2 and No. 3 captcha image dataset, the interference noise is line more and conglutination is relatively simple, the accuracy of recognition reached 97% and 96.4%.

Compare the improved vertical projection method proposed in this paper with traditional one for segmentation and experiment with the captcha image dataset of this paper, getting the recognition accuracy of two methods shown in Table IV.

From Table IV we can see that comparing with traditional vertical projection method, the improved vertical projection method proposed in this paper has much better segmentation effect, more intactly reserve the information of single character after segmentation, which laid the foundation for the subsequent recognition work, thus to improve the captcha recognition accuracy.

TABLE II. EXPERIMENTAL RESULTS OF DIFFERENT CAPTHAS

No.	Original image	Graying	Denoising	Segmentation effect	Recognition result
1					93608
2					NyezX

TABLE III. EXPERIMENTAL TESTING RESULT STATISTICS


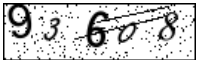

No.	Captcha image	Num of samples	Total chracters	Accuracy
1		100	500	92.3%
2		100	500	97%
3		100	500	96.4%

TABLE IV. CHARACTER SEGMENTATION ALGORITHM EXPERIMENTAL RESULTS

Segmentation algorithm	Traditional vertical projection	Improved Vertical projection in this paper
Accuracy	76.5%	92.3%

V. CONCLUSION

With the proposed improved vertical projection method can effectively identify and segment not conglutination, part conglutination, and high conglutination characters, help to improve the accuracy of the captcha recognition. On this basis, but also put forward the oblique direction of the projection method, solve the segmentation problem of italic characters. The shortcoming is that the robustness of this method is not strong enough. It is easily affected by previous captcha image preprocessing effect. If preprocessing is not good enough and too much interference noise residue remains on captcha image, it will greatly reduce the segmentation results. In this paper, the research on segmentation problem of captcha recognition has certain reference significance on related areas of research and application. To improve reliability of captcha and maintain the network security problem has certain application prospect.

ACKNOWLEDGEMENT

This work is supported by National Science Foundation of China (No. 61571453) and Major Project of Education Department of Hunan Province, China (No. 15A020). The authors are grateful for the anonymous reviewers who made constructive comments.

REFERENCES

- [1] Bursztein E, Martin M, Mitchell J. Text-based CAPTCHA strengths and weaknesses.[C]// ACM Conference on Computer and Communications Security, CCS 2011, Chicago, Illinois, Usa, October. 2011:125-138.
- [2] Ran L F. An Algorithm of Characters Segmentation based on Vertical Projection for License Plate[J]. Communications Technology, 2012(45):89-91,98. [in Chinese]
- [3] Li C. Cluster-Based Method of Characters' Segmentation of License Plate[J]. Computer Engineering & Applications, 2002, 221-222,256. [in Chinese]
- [4] Grira N, Crucianu M, Boujemaa N. Fuzzy Clustering with Pairwise Constraints for Knowledge-Driven Image Categorization.[C]// Knowledge-Based Media Analysis for Self-Adaptive and Agile Multi-Media, Proceedings of the European Workshop for the Integration of Knowledge, Semantics and Digital Media Technology, EWIMT 2004, November 25-26, 2004, London, UK. 2004:299--304.
- [5] Congedo G, Dimauro G, Impedovo S, et al. Segmentation of numeric strings[C]// International Conference on Document Analysis and Recognition. IEEE Computer Society, 1995(2):1038-1041.
- [6] Bensaid A M, Hall L O, Bezdek J C, et al. Partially supervised clustering for image segmentation[J]. Pattern Recognition, 1996, 29(5):859-871.
- [7] Feng Y E, Chen C J, Lai Y Z, et al. Fast circle detection algorithm using sequenced Hough transform[J]. Optics & Precision Engineering, 2014, 22(4):1104-1111. [in Chinese]
- [8] Liu G H. Application of improved arithmetic of median filtering denoising[J]. Computer Engineering & Applications, 2010,46(10):187-189. [in Chinese]
- [9] Lian Xiaoyan, Dengfang, CAPTCHA recognition based on image recognition and neural networks[J]. Journal of Central South University(Science and Technology), 2011(42):48-52. [in Chinese]
- [10] Wei W, Zhang Q S, Wang M J, et al. A method of number-plate recognition using templates matching[J]. China Journal of Highway & Transport, 2001, 14(1):104-106.[in Chinese]