A method of character recognition based on general characteristic and connected regions

Meng Qingyuan
College of Sciences, North University of China
North University of China, NUC
Taiyuan, China
mengqy1983@163.com

Hu Hongping
College of Sciences, North University of China
North University of China, NUC
Taiyuan, China
huhongping@nuc.edu.cn

Bai Yanping
Institute of Microelectronics, Peking University
Beijing, China
baiyp@nuc.edu.cn

Abstract—In this paper, by analyzing the number of concave domain and cycles and the characteristic of connected domain of the characters, we present a method of character recognition based on above mentioned features. At first, the characters are divided into two classes according to cycles: the character with cycles and the character without cycles. Then, we detect if the character has sunk parts in its four directions (up, down, left, right).According to the results, the character can be recognized or divided into six classes. At last, we recognize the character using the characteristic of connected domain, and obtain the final recognition results. The experiment shows that the algorithm has a good performance in English characters recognition.

Keywords: character recognition; cycle; concave domain; connected domain;

I. Introduction

With the development of the economy, the amount of automotive vehicles is increased sharply, which brings tremendous pressure to the traffic administrative department. In order to relieve this pressure and improve the work efficiency, the Intelligent Transport System (ITS, for short) rises in response to the proper time and conditions. ITS, which is applied to communication and transportation fields, is a total management system. It is represented by the information technology and fuses advanced information technology, transducer technology, control technology and computer processing technology, etc[1].License Plate Recognition (LPR) system is an important part of ITS. Meanwhile, LPR is also an important application fields on research of Optical Character Recognition (OCR). The LPR system is widely used in freeway management, electronic police and parking management [2].

For the decades, the researchers arrived at a variety of recognition ways, which can be carved up to statistic character oriented ways and structure character oriented ways [3]. The character image's statistic character usually contains point density, moment, feature region, connected domain, similarity, the maximum width of the character, the[1] number of strokes of the character, etc. The structure character includes cycle, endpoint, cross point, stroke, outline and sunken area, and so on [4]. Despite we can recognize character effectively using either of the two methods, the false and rejected identification also arise inevitably for various reasons. The reason of false identification is the effect of external factors, such as illumination, weather-factor and the shooting angle, and so on. Because of the influence of the factors, the characters may be reformative and the strokes become disconnected. The algorithm selected also has an influence on recognition rate. Theoretically, we can use amount of methods to recognize the same character, analyzing its characteristics by synthesis and arriving at the best result. But in application, we have to choose one or several features owing to the factor of efficiency. So we can select several features respectively in the statistic ways and the structure ways, and integrated use those features to recognize the characters. With this idea, we can choose fewer features and achieve better results.

In this paper, the author present a new English character recognition method by analyzing the cycles and concave domain contained in a character's binary image and the character image's connected domain. According to the method, the author divides the characters into characters with cycles and that without cycles by analyzing if the characters have cycles. Then, by judging if the characters have concave domains, some characters are recognized, and others are divided into six classes. At last, the characters are recognized by its connected.

II. The architectural feature of character image

By analyzing the characters image's outline, we find the characters have two distinguished features. The first one is a

Funding: This work was supported in part by the National Natural Science Fund of China (60876077), the Science Fund of Shanxi (2009011018-3) and the National Postdoctoral Science Foundations of China.

978-0-7695-4356-7/11 $26.00 © 2011 IEEE
DOI 10.1109/CMSP.2011.25

domain encircled by the character's strokes, which is called cycle. And the other one is its concave-convex regions, which is called concave domain. So if a domain is a cycle, all of the lines led from the domain's discretionary point's eight directions (up, down, left, right, top left, top right, lower left and lower down) must intersect the character's stroke area. But for the concave regions, there is at least one point in the region, whose eight lines above can arrive at the image's edge without intersecting the stroke region. The number of lines met the above conditions is greater than one and less than three.

2.1 The background field of the character image

The elements of a character's binary image can be sorted into two categories: background elements and foreground elements (stroke elements). Generally, the value of the background elements is zero, and the foreground ones' are 1(or 255, choose 1 here). We can extend eight line met the above conditions from any point. Supposing the number of lines intersecting the stroke regions is *n*, we called n as the point's background field. The rule of calculating the background field of the point A is as follows: if A is not an edge point, the value of its background field is the number of the lines intersecting the stroke area; if A is an edge one, we consider that the lines exceeding the image's edge don't intersect the stroke area. From the rule, we can know that the pixel's background field is an integer from 1 to 8. As shown in Fig.1, if the point A is not a edge element (as shown in Fig 1(a)), we can lead eight line from A, and the background field is 8 in this case; if the point A is on the edge (as shown in Fig 1(b)), we can obtain five lines from A, and the value of A is 3.

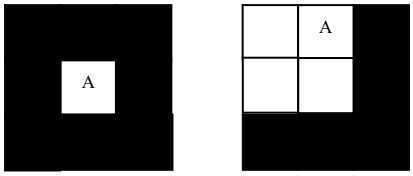

Fig.1 the background field of pixel

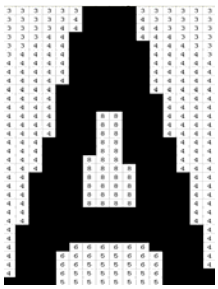

Fig.2 the background field array of character

Abstracting the connected domain which the background field is greater than five in the character's binary image, and analyzing the average value of every domain, we can reach the cycles and indentations of the image: if the region's average value is equal to eight, the region is a cycle; if the region's average value is less than five, the region is an indentation.

As is shown in Fig 2, this is a background field array of character A. In the array, the middle part, which is encompassed by the stroke area, is a cycle, for its average background field is eight. And the bottom region of the array has an average background field of 5, so this region is a hollow area.

2.2 Extracting the cycles of character

From the above analysis, we can know that a background region is a cycle if it is surrounded by the stroke area. Therefore, a discretionary pixel of a area has a 8-background-field if the area is a cycle. According to the character of cycles, we can confirm a way to extract cycles from character image:

i) extracting pixel whose background field is greater than 5 in order;
ii) extracting the connected regions comprised of the above pixel;
iii) calculating the average value of every region's background field, if the value is equal to eight, it is a cycle; otherwise, it is a hollow region.

2.3 Extracting the concave region of character

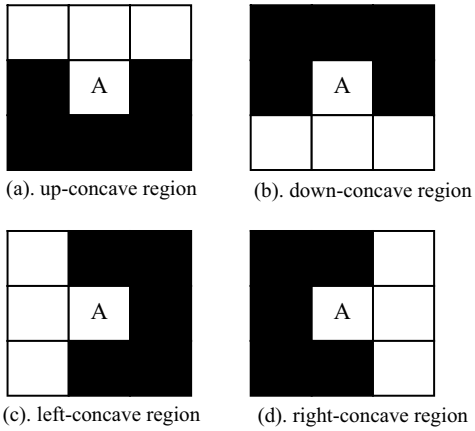

Fig.3 the concave region

The character has another important feature which is called concave region except cycle. Judging by the position of the region, we can divide the concave into four classes: the up-concave region, the down-concave region, the left-concave region and the right-concave region (as shown in Fig.3). In Fig 3, the white check stands for a background region, and the black check stands for a foreground one. The region A is a concave area as is shown in Fig 3. The lines are spread out in four directions (up, down, left and right) of a point of A. If the line extended upwards arrives at the edge without passing the stroke regions, we call A as a concave region [5] (as shown in Fig 3(a)). In a similar way, we can define the down-concave region (shown in Fig.3 (b)), the

left-concave region (shown in Fig.3(c)) and the right-concave region (shown in Fig.3 (d)).

Known from the above definition, in an English character image, a region can be not only an up-concave area (or a down-concave one), but also a left-concave region (or a right one) (as shown in Fig.4). But it mustn't be a up-concave region and a down-concave region (or a left one and a right one). The reason is that, if a region is an up-concave area and a down one, the region will divide the character strokes into two parts. It means that the character has two separate stroke connected regions (shown in Fig.5). But in a character image, there is one and only one stroke region. So the condition doesn't exist.

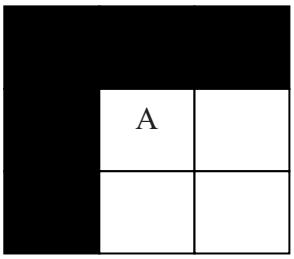

Fig.4 right-down concave region

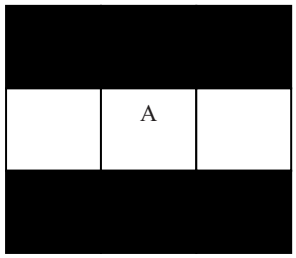

Fig.5 up-down concave region

For a binary image, we can get its array of background field easily. In the array, some parts stand for cycles and some other parts stand for concave region. By extracting the connected area comprised of the pixels whose average value of background field is bigger than 5 and less than 8, we obtain the concave regions preliminarily. But in these regions, some regions may be connected with each other. For example, the character E, which has two right-concave areas in structure, only has one concave region, for the two regions have linked together (shown in Fig.6). By analyzing this situation, we found that every pixel of the linked part has a 5-background field. And the pixels of the regions whose background field is greater than 5 are separate. So, in this case, we should deal with as follows: retain the pixels whose background field is greater than 5; for the pixel having a 5-background field, if one of its consecutive point's background fields is bigger than 5, reserve it, otherwise, clear it from the region. Through these steps, we can separate the linked regions (shown in Fig.7). If the area still can't be separated after treatment, if is considered as a single region.

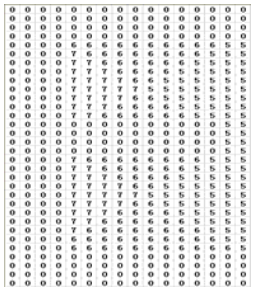

Fig.6 the concave region of 'E'

Through steps the above treated, we select a pixel in the region and analyze the class of the area according to the definition above.

III. The feature of character's connected domain

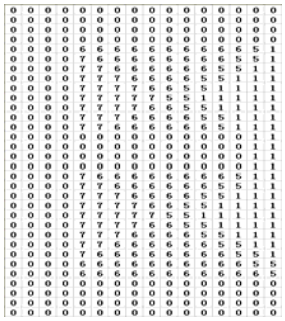

Fig.7 the concave region separated

The character matrix $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, which have the same dimension, have a number of connected regions as $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ and their coincident part has a number of $\boldsymbol{d}_3$. Supposing the absolute value of difference between $\boldsymbol{d}_1$ and $\boldsymbol{d}_3$ is $\boldsymbol{d}$, that is

$$\boldsymbol{d}=|\boldsymbol{d}_3-\boldsymbol{d}_1|.$$

We call $\boldsymbol{d}$ as the feature of connected region that $\boldsymbol{X}_1$ compared to $\boldsymbol{X}_2$. It is easy to know that, for most of English character, only when $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ stand for the same character, the d can take the minimum value. Consequently, if $\boldsymbol{A}$ is a template waiting for recognized and $\boldsymbol{B}$ is a standard template, it could be thought that the d can take the minimum value only when $\boldsymbol{B}$ is the correct identification result of $\boldsymbol{A}$. We can use the hypothesis to recognize the character [6].

IV. Classification

Based on the research above, we extract the cycles and concave regions from standard template. The results are shown in Table 1.

Table.1 the cycles and concave regions of character

	A	B	C	D	E	F	G	H	I	J	K	L	M
Cycle	1	2	0	1	0	0	0	0	0	0	0	0	0
Up-concave	0	0	0	0	0	0	0	1	0	1	1	1	1
Down-concave	1	0	0	0	0	1	0	1	0	0	1	0	1
Left-concave	0	0	0	0	0	0	0	0	2	1	0	0	0
Right-concave	0	1	1	0	2	2	1	0	2	0	1	1	0
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Cycle	0	1	1	1	1	0	0	0	0	0	0	0	0
Up-concave	1	0	0	0	0	0	0	1	1	2	1	1	0
Down-concave	1	0	1	1	1	0	2	0	0	1	1	0	0
Left-concave	0	0	0	0	0	1	1	0	0	0	1	0	1
Right-concave	0	0	1	1	1	1	1	0	0	0	1	0	1

By analyzing the data in the table, we found, the characters can be classed in the character with cycle and that without cycle. Then the characters are classified further by its distribution of the concave regions. The procedure of recognition is shown in Fig 8 and Fig 9.

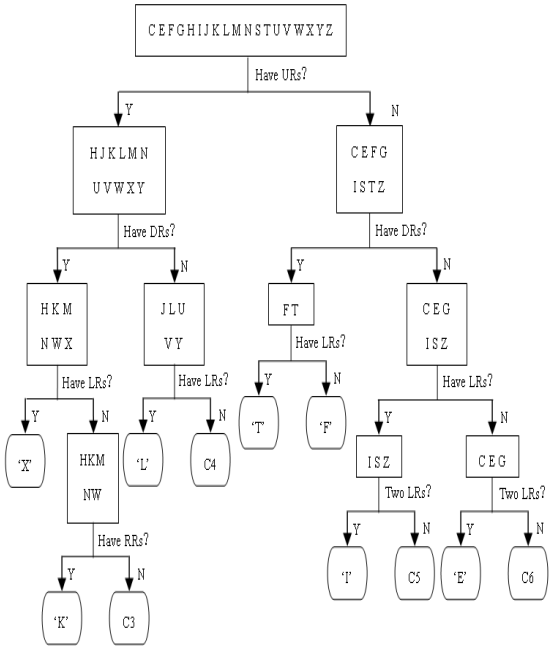

Fig.8 the recognition of character without cycles

UR: up-concave region DR: down-concave region

LR: left-concave region RR: right-concave region

By analyzing a large number characters extracted from the license plates, we found that the character Q and W have their exceptions. In the standard character set, the character has a down-concave region and a right-concave region. And its down-concave region is very small. The character W has two up-concave regions and a down-concave region. Because of the external factors' influence on the image during the sampling process and the losing of the image details in image processing, the character Q often lost its down-concave region, and the two up-concave regions of character W are linked together.

According to the facts, we carried out the corresponding treatments during the recognition process. For example, we class the characters on the basis of their right-concave regions when the set contains the character Q, we don't class the characters further in accordance with the character's two up-concave regions when the set includes the character W.

Through the recognition process shown in Fig. 8 and Fig. 9, a character binary array will be recognized or divided into one of the six character sets. Then we recognize the character secondly using the method of connected region, and get the final result.

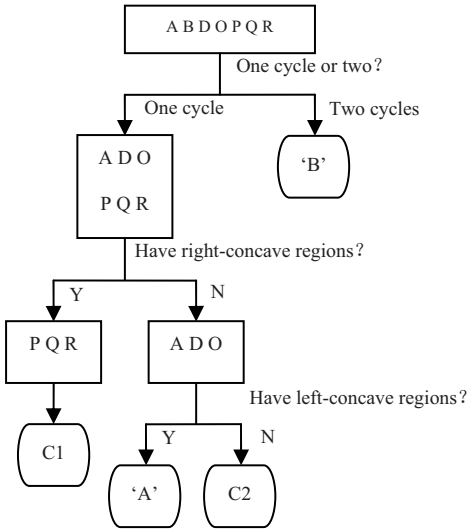

Fig. 9 the recognition of character with cycles

V. Experiment

We extract 2624 English characters from 1189 photos (all of the characters don't contain I and O). And the paper programs the above algorithms using MATLAB and does the simulation experiments. The result is shown as the Table.2 and Table.3.

Table.2 recognition results

character	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***
quantity	269	121	107	134	117	100	106	92
correct recognition	264	121	91	124	112	98	92	86
recognition rate(%)	98.1	100	85.0	92.5	95.7	98.0	86.8	93.5
character	***J***	***K***	***L***	***M***	***N***	***P***	***Q***	***R***
quantity	108	98	94	112	101	103	95	109
correct recognition	101	98	85	110	101	103	91	109
recognition rate(%)	93.5	100	90.5	98.2	100	100	95.8	100
character	***S***	***T***	***U***	***V***	***W***	***X***	***Y***	***Z***
quantity	97	133	71	84	103	87	95	88
correct recognition	94	133	67	84	103	87	92	84
recognition rate(%)	96.9	100	94.4	100	100	100	96.8	95.5

Table.3 total recognition rate

total quantity	correct recognition	recognition rate
2624	2530	96.4%

According to the analysis of the experimental result, we found the mainspring that affecting the recognition rate is the false identification of similar characters. In especial, The similar characters 'C' and 'G', have the greatest influence on the rate of identification. From the above table, we can find that the recognition rates of 'C' and 'G' are significantly lower than other characters. Meanwhile, as a result of the image binarization, the indistinct strokes and the lost of concave regions are also important factors affecting the recognition rate.

In the 2624 characters, there are some noisy ones. The program also has a good performance. Some noisy characters and their recognition results are shown in Fig. 10.

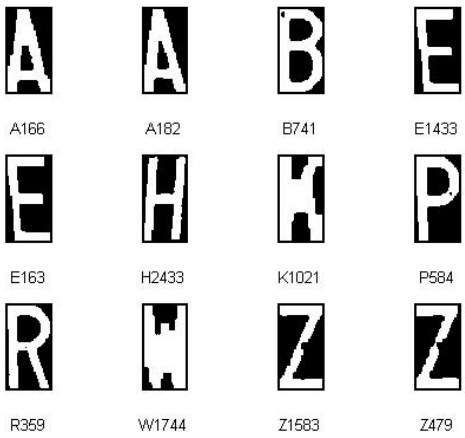

Fig.10 the recognition results of noisy characters

VI. Conclusion

In this paper we propose a method for license plate character recognition that relies on the distribution of character's cycles and concave regions. At first, we class the character as that with cycles and without cycles by checking if the cycles exist in character image. Then, on the basis of the distribution of the character's concave region, the character is classified secondly. The character is recognized or categorized to a classification in the process. At last, we use the character's feature of connected area to recognize it, and get the final result.

Because the way is based on the character's general features (the character's concavo-convex feature), we needn't do complex transactions of the character's local details, and spare a mass of computational time. Meanwhile, the character has the same general characteristics whether it is oblique or its stroke becomes fuzzy in image processing. Therefore, using the algorithm to recognize characters, we certainly can get a satisfactory result. The experimental result also shows that the way in this paper has a good performance in English character recognition.

References

[1] Shuang Wei, "The Research on Techniques of Car License Plate Recognition Based on Neural Network", master's thesis, Shanxi, North University of China, April 2009.

[2] Li Yanping, "Application study of digital image processing in ITS", master's thesis, Liaoning, Dalian University of Technology, Dec. 2005.

[3] Bian Zhaoqi, Zhang Xuegong, Pattern Recognition, Beijing,China:Tsinghua University, 2000,.

[4] Lou Zhen, Hu Zhongshan, Yang Jingyu,"Handwriting Numerals Recognition Based on Segmented Contour Feature", Chinese J.Computer, vol. 22, Oct. 1999, pp. 1065-1073.

[5] Gong Caichun, Liu Rongxing,"Fast Hand-Written Digital Character Recognition Based on Global Feature", Computer Engineering and Applications, Jan. 2004,vol. 9, pp. 81-83.

[6] Meng Qingyuan, "The Study of Vehicle Template Character Recognition Technology Based on Connected Domain Characteristic", Journal of Test And Measurement Technology, in press.