License Plate Character Segmentation Algorithm Based on Improved Regression Model

Yuping Yang

1 Smart health Institute, Chongqing College of Electronic Engineering, Chongqing, 401331, China

Abstract. Many character segmentation algorithms in license plate recognition system are not very good for segmentation of complex characters, and there is a phenomenon of character sticking, which results in slow recognition speed and not very good recognition effect of the license plate with complex characters. For such problems, a segmentation algorithm based on the improved regression model on statistical characteristics is presented. Compared with other algorithms, experimental results show that less time is consumed and segmentation effect is better. The recognition rate of license plate can be improved effectively.

1. Introduction

License plate character segmentation is the key content of ITS [1][2]. At present, the commonly used algorithms of license plate character segmentation are projection algorithm, template matching algorithm, clustering analysis algorithm and connection domain algorithm, etc. The algorithm proposed in document [3] can remove certain noise and vertical border interference on both sides of the license plate [3], but it has poor segmentation effect on Chinese characters under uneven illumination and harsh environment; Literature [4] uses horizontal cutting and vertical projection to solve the problem of character recognition [4], but this method is very sensitive to license plate tilt. Literature [5] can better solve the complex situation of Chinese character disconnection [5], character cohesion and border cohesion, but it is sensitive to noise. Generally, the projection algorithm is relatively simple to implement, but it is very sensitive to tilt. At the same time, the problem of disconnection and cohesion of Chinese characters has not been solved very well. Therefore, literature [6] uses the edge information of the license plate frame to correct the tilt of the license plate, which combines vertical projection with template matching [6]. However, the algorithm improves the accuracy of character segmentation, but improving the complexity of the algorithm. Literature [7] solves the problem of character segmentation of license plate under complex background better [7]. But it needs a certain prior knowledge and cannot segment the glued characters correctly. The linear regression model can make full use of the rich data information provided by each component of the colour image RGB to construct a multiple linear regression model for image segmentation. A large number of experiments have proved that the algorithm has strong robustness to noise and other harsh environments, and can better solve the problem of character disconnection and cohesion, with a wide range of applications.

2. Establishment of Regression Model for License Plate Recognition

2.1. Problem of Regression Model
Definition 1: Multivariate linear regression model is a linear regression model with multiple explanatory variables, which is used to explain the linear relationship between the explanatory variables and other explanatory variables. The mathematical model is following formula (1).

\[ y = \varepsilon + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \]  

(1)

\( \varepsilon \) is a random error, called residuals. \( \beta_0, \beta_1, \cdots, \beta_n \) are partial regression coefficients. Model Formula (1) indicates that negative variables are interpreted jointly by explanatory variables. The sampling distribution of the estimated partial regression coefficients follows the following formula (2):

\[ \beta \sim N \left( \beta, \frac{\sigma^2}{\sum_{i,j=1}^n (x_{ji} - \bar{x}_i)^2} \right) \]  

(2)

Under the original hypothesis, t-test statistics can be constructed:

\[ t_i = \frac{\beta}{\sigma / \sum_{i,j=1}^n (x_{ji} - \bar{x}_i)^2} \]  

(3)

2.2. Establishment of Regression Model for License Plate Recognition

2.2.1. Procedure of regression model establishment

- Determine the explanatory variables and interpreted variables in the regression model;
- Sample the population;
- Determine \( \beta_0, \beta_1, \cdots, \beta_n \) in formula (1) by the commonly used least squares estimation method combined with formula (2) formula, and then get a regression equation;
- Test the significance of the regression equation with (3) formula.

Formula (1) can be expressed in matrix form \( Y = \varepsilon + X \beta \), and \( X = [x_1, x_2, \cdots, x_n] \), \( \beta = [\beta_0, \beta_1, \beta_2, \cdots, \beta_n]^T \). According to the least square method, \( \beta \) can be calculated by \( \frac{\partial (\varepsilon^T \varepsilon)}{\partial \beta} = -2X^T Y + 2X^T X \beta = 0 \), and that is \( \beta = \left( X^T X \right)^{-1} (X^T Y) \).

2.2.2. RGB Component Regression Model of License Plate Character Background

Firstly, the R component in the RGB component with significant correlation is taken as the interpreted variable; the G component and the B component are taken as the explanatory variables. And since the image can be represented by two-dimensional coordinates, it can be assumed that the background image of the license plate character is \( BS(i, j) \), then the corresponding component is expressed as \( BS \_ R(i, j), BS \_ G(i, j) \) and \( BS \_ B(i, j) \). Combining with formula (1), the regression model can be further expressed as formula (4):

\[ BS \_ R(i, j) = \beta_0 + \beta_1 \cdot BS \_ G(i, j) + \beta_2 \cdot BS \_ B(i, j) \]  

(4)

Secondly, 50 background pixels of license plate characters are extracted as samples, and the RGB component values of the sample points are extracted to form the variable matrix.

Thirdly, according to the least square method, \( \beta = \left( X^T X \right)^{-1} (X^T Y) \) can be obtained. Sampling distribution of the obtained values is verified by formula (2) and the regression equation \( R = \beta_0 + \beta_1 \cdot G + \beta_2 \cdot B \) is obtained.

Finally, the significance of the regression equation was tested by formula (3).
3. Improvement of regression model

Through experiments, we found that the noise in the background colour cannot be removed when the background colour is removed by regression model alone. And the removal effect is not ideal at the edge of the boundary between the background colour and the character colour. The main reason is that the background colour we sampled is relatively single, so we will further improve the sampling method. Firstly, we can find the boundary between the license plate characters and the background colour by calculating the difference image and mark it. Then we can calculate the average values of the left, right, up and down pixels at the marks. Because some parts of the average values are calculated from the image at the different character boundary. When the average values are directly used to estimate the value of $\beta$, the fitness of the regression equation is poor. Therefore, further optimization of the norm is needed to achieve better fitness of the regression equation. The specific process chart is as follows:

![Character Segmentation Process](image)

**Figure 1. Character Segmentation Process of Improved Regression Model**

3.1. Differential Image Construction of License Plate Image

The license plate images $\text{RGB}(i,j)$ with $\text{R}$, $\text{G}$ and $\text{B}$, which have been positioned and processed well, are decomposed into three components: $\text{R} \_ \text{S}_{\text{temp}}(i,j)$, $\text{G} \_ \text{S}_{\text{temp}}(i,j)$ and $\text{B} \_ \text{S}(i,j)$. Assume that $\text{R} \_ \text{S}_{\text{temp}}1(i,j)$, $\text{G} \_ \text{S}_{\text{temp}}1(i,j)$, and $\text{B} \_ \text{S}_{\text{temp}}1(i,j)$ are vertical differential images of these three components respectively, and $\text{R} \_ \text{S}_{\text{temp}}2(i,j)$, $\text{G} \_ \text{S}_{\text{temp}}2(i,j)$ and $\text{B} \_ \text{S}_{\text{temp}}2(i,j)$ are transverse differential images of these three components respectively. Each component is substituted into formula (5) for vertical and horizontal differential operations.

$$
\begin{align*}
\text{S}_{\text{temp}}1(i,j) &= \text{S}(i,j+1) - \text{S}(i,j) \\
\text{S}_{\text{temp}}2(i,j) &= \text{S}(i+1,j) - \text{S}(i,j)
\end{align*}
$$

(5)

Let $\text{R} \_ \text{mark}(i,j)$, $\text{G} \_ \text{mark}(i,j)$ and $\text{B} \_ \text{mark}(i,j)$ separately be labelled images of three components of the same size as the original image, traverse $\text{R} \_ \text{S}_{\text{temp}}2(i,j)$, $\text{G} \_ \text{S}_{\text{temp}}2(i,j)$ and $\text{B} \_ \text{S}_{\text{temp}}2(i,j)$, and find out the maximum and the second largest values of them by traversing, and assign values of 1 at $\text{R} \_ \text{mark}(i,j)$, $\text{G} \_ \text{mark}(i,j)$ and $\text{B} \_ \text{mark}(i,j)$ corresponding positions respectively. Set $\text{RGB} \_ \text{mark}(i,j)$ as the labelled image of the colour image of the license plate of the same size as the original image, the initial value is all 0. The following formula (6) is used for calculation.

$$\text{RGB} \_ \text{mark}(i,j) = 1, \quad \text{R} \_ \text{mark}(i,j) \& \& \text{G} \_ \text{mark}(i,j) \& \& \text{B} \_ \text{mark}(i,j) = 1$$

(6)

3.2. Find the Mean Value of Marked Points

The marked point $\text{RGB} \_ \text{mark}(i,j)$ is the position where the values of each component of RGB vary greatly. However, the position $\text{RGB} \_ \text{mark}(i,j)$ of the pixels in which the RGB component values may be the background colour of the pixel values, or the character colour of the pixel values, or may be between the character and the background pixel values. Since we generate the differential
images vertically and horizontally, in order to make the extracted pixel values more universal, we will calculate their average pixel values in the way of Figure 2.

![Figure 2. Computation Way of Average Pixel Value](image)

In Figure 2, the pixel block has eight pixel values, "○" denotes the pixel value that does not participate in the calculation, "●" denotes the pixel value that participates in the calculation of the average value, "※" denotes the location of the marked point, and also is the location stored after the calculation of the average value of the pixel. Let \( R_{AV}(i,j) \), \( G_{AV}(i,j) \) and \( B_{AV}(i,j) \) store the calculated average value of the RGB component pixels, \( R_S(i,j) \), \( G_S(i,j) \) and \( B_S(i,j) \) are the three component pixel images respectively.

### 3.3. Optimization with Norms

In order to better remove the background colour of characters of license plate, not only the background pixel values far from the edge of the characters, but also the background pixel values of the character edge need to be sampled. Such sampling also results in a larger range of pixel values and a larger fitting degree of the constructed regression model, which reduces the effect of background removal. Therefore, it is necessary to optimize the norm of the sampled pixel values. Taking the centre line of the license plate as the standard, 50 pixels at the bottom and top of the license plate are taken as background pixel sampling points farther from the edge of the character, totalling 200 pixels, and the RGB component pixel values of these pixels are expressed \( R_{BAK}(i,j) \), \( G_{BAK}(i,j) \) and \( B_{BAK}(i,j) \) respectively. \( R_{AV}(i,j) \), \( G_{AV}(i,j) \) and \( B_{AV}(i,j) \) denote the RGB component pixel value of the edge of the character of background, which is calculated in the way of Figure 2. The following norm optimization formula (7) is used for optimization.

\[
\begin{align*}
\tilde{R}(i,j) &= \frac{R(i,j)}{\sqrt{R(i,j)^2 + G(i,j)^2 + B(i,j)^2}} \\
\tilde{G}(i,j) &= \frac{G(i,j)}{\sqrt{R(i,j)^2 + G(i,j)^2 + B(i,j)^2}} \\
\tilde{B}(i,j) &= \frac{B(i,j)}{\sqrt{R(i,j)^2 + G(i,j)^2 + B(i,j)^2}}
\end{align*}
\]

(7)

After norm optimization, the RGB component values obtained from the two samples will be estimated and determined by combining formula(4), (2), (3), and the regression equation will be constructed. Finally, the built regression equation is used to remove the background of license plate characters, and the traditional vertical projection and template matching are used to recognize the characters.

### 4. Contrast and Analysis of Experimental Data

The experimental work of this algorithm is based on OpenCL, which is an open source visual library, and 200 license plate pictures are randomly selected for experiment. In order to reflect the advantages and disadvantages of each algorithm, and have a more comparative significance, when compared with literature [5], literature [6], and literature [7], this paper adopts the same way in the stages of license plate pre-processing and character recognition, such as license plate tilt correction with Hough transform; character recognition with template matching.
Figure 3 shows the effect of different algorithms applied to the same license plate. Figure 3 (a) is the original image, which has been preliminarily processed by positioning and slope correction. Figure 3 (b) is the effect of character segmentation, edge extraction and binarization by using the algorithm of literature [5]. Figure 3 (c) is the result of segmentation and binarization of literature [6]. Figure 3 (d) is the result of segmentation and binarization of literature [7]. Figure 3 (e) is the algorithm of this paper, and it can be seen that the improved sampling method is used. For the sampling of the pixels at the junction of characters and background, the constructed regression model can process the edges of characters better, especially in the dense strokes of Chinese characters, which improves the subsequent character recognition rate.

![Figure 3](image)

Table 1 is a comparison table of running time and recognition rate of each algorithm. The time-consuming of literature [6] algorithm is the shortest, the time-consuming of literature [7] algorithm is the longest, and the time-consuming of this algorithm is in the middle and closer to the shortest. At the same time, the recognition rate of literature [6] algorithm is the lowest, and the recognition rate of this algorithm is also the highest, almost 96%. Generally speaking, the proposed algorithm is superior to the other three algorithms.

| algorithms          | Number of license plates | Recognition time (unit: s) | Recognition accuracy |
|---------------------|--------------------------|----------------------------|----------------------|
| literature[5]       | 200                      | 1.1                        | 93%                  |
| literature[6]       | 200                      | 0.8                        | 90%                  |
| literature[7]       | 200                      | 1.2                        | 92%                  |
| proposed algorithm  | 200                      | 0.9                        | 96%                  |

5. Conclusion
The effect of license plate character segmentation is very important to improve the recognition rate of license plate. This paper proposes an improved license plate character segmentation algorithm based on statistical regression model. Pixels are sampled from character background and the intersection of character and background. By calculating representative mean and norm optimization, a more fitting regression equation is constructed. Removing the background of license plate characters has achieved good results.

Acknowledgments
I wish to acknowledge assistance and encouragement from colleagues. The work described in this paper was fully supported by a grant from the General Scientific Research Project of Chongqing College of Electronic Engineering of China in 2019 "Research on Intelligent Robot of Home Care
Based on AI (NO.XJZK201905)" and Intelligent Robot Technology Research Centre of Chongqing College of Electronic Engineering (XJPT201705).

References
[1] MUHAMMAD SALMAN. Vehicle License Plate Detection and Characters Recognition [D]. South China University of Technology, 2018.
[2] Lyimo Martha Donald. Segmentation and Extraction of Vehicle License Plate [D]. South China University of Technology, 2017.
[3] License Plate Detection and Character Segmentation Using Adaptive Binarization Based on Superpixels under Illumination Change [J]. IIEICE Transactions on Information and Systems, 2017, E100.D (6):1384-1387
[4] El-said, Shaimaa Ahmed. Shadow aware license plate recognition system [J]. Soft computing: A fusion of foundations, methodologies and applications, 2015, 19(1): 225-235.
[5] Wang, Jailing, Huang, He, Dai, Yoking, etal.Sequence recognition of Chinese license plates [J]. Neurocomputing, 2018, 317 (Nov.23): 149-158.
[6] R. Kh. Khamdamov, H. E. Rakhmanov. A Character Segmentation Algorithm for Vehicle License Plates[J]. Cybernetics and Systems Analysis, 2019, 55(4).
[7] Chen, Bei, Cao, Wenlun, He, Yuyao. An Automated Vision System for Vehicle License Plate Localization, Segmentation in Real Life Scene[J]. International journal of pattern recognition and artificial intelligence, 2018, 32