2019 IEEE Symposium Series on Computational Intelligence (SSCI)
December 6-9 2019, Xiamen, China

Improving the Accuracy of License Plate Detection and Recognition in General Unconstrained Scenarios

Zhongze Zhang
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
zhangzhz17@lzu.edu.cn

Yi Wan
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
wanyi@lzu.edu.cn

***Abstract*—Nowadays, Automatic License Plate Recognition (ALPR) is widely used in commercial applications. Some of the existing methods focus on the detection of license plate (LP) images with approximate frontage, while others deal with the unconstrained LP of images, but the result of LP recognition under some extreme conditions is unsatisfactory. This work mainly focuses on improving the existing unconstrained LP capturing method [1] to extract the LP and recognize it in several extreme cases as many LPs may be seriously distorted, even not parallelogram. Our main contributions are the following. First is to modify the loss function of the existing Convolutional Neural Networks (CNNs) so that the output parameters can form a parallelogram directly, then, LPs could be detected more accurate in a single input image first. Secondly, two parameters need to be estimated are added to the output of the networks' fully connected layer structure to make the corresponding output parameters form an arbitrary quadrilateral, which is closer to the LP shape of the actual imaging. In this way, it can detect and correct LPs with various shapes (LP may be deformed into irregular quadrilateral due to various reasons) in a single image. Thirdly, a feedback mechanism would be added to identify and process the LP images which are not detected correctly after the LP characters are recognized by Optical Character Recognition (OCR) network, then, return the processed image to the detection network for re-detection. In general, LP can be detected correctly. Compared to the other benchmark methods, the experimental results have demonstrated that the proposed method achieved the best performance, especially in the extreme conditions.**

Keywords*—license plate; Deep Learning; Convolutional Neural Networks

I. Introduction

Because of the wide application of license plate (LP) detection and recognition technology in commercial field, people have been paying close attention to the development of this technology. From checking traffic violations, charging violations to accident monitoring, automatic LP detection and recognition is one of the key tools used by law enforcement agencies and parking lots around the world. With the rapid development of science and technology, parallel processing and Deep Learning (DL) help improve many computer vision tasks [2]–[6], such as object detection/recognition and Optical Character Recognition (OCR). This is also useful for Automatic License Plate Recognition (ALPR) system. In fact, the application of Convolution Neural Network (CNN) to vehicle and LP detection has become a leading machine learning technology. However, due to the limitation of camera location and LP type, the accurate detection and recognition of LP in various situations is still a challenging problem. Some methods of recognizing and detecting the front images of LP (such as fixing the camera position, selecting a specific perspective with a specific resolution, using templets of LP, inspection in common scenes like toll monitoring and parking lot) have achieved good results at present. However, there are various of random image acquisition scenarios in our life (for example, law enforcement officers who walk with mobile cameras or smartphones). The LP images captured in these special scenarios may be inclined views, and the LPs extracted from them may be highly distorted. Nowadays, though the existing methods can get some information by special processing of these images, the recognition result is still unsatisfactory.

This work intends to improve the related technique [1], mainly in three aspects. Firstly, the related technique gets the four vertices of LP by affine transformation of six parameters of CNNs' fully connected layer output. Without changing the structure of CNN, we modify the loss function to make the six coordinates of the LP's three vertices equal to the six parameters of the output of the networks' fully connected layer directly, and then calculate the last vertex coordinate of the LP through the three vertices' coordinates. So that it can detect the LP more accurately which can get more accurate LP characters. Secondly, the related technique [1] cannot be handled well in some worse conditions, because the LP detected by this method is a parallelogram, but in practice, the LP photographed may be irregular quadrilateral due to distortion. So, two parameters need to be estimated are added to the output of the networks' fully connected layer structure base on the first step. Correspondingly, modify the loss function to make the eight coordinates of the LP's all four vertices equal to the eight parameters of the output of the networks'

978-1-7281-2485-8/19/$31.00 ©2019 IEEE

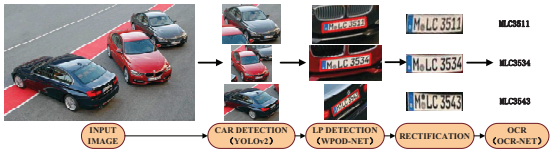


Fig. 1. Indication of the whole process [1].

fully connected layer directly. In this way, the detection of LP can be arbitrary quadrilateral because there is no constraint relationship between the eight parameters. Even if the LP taken at extreme angle is not a parallelogram, complex deformation LP can be identified by network. Thirdly, we add a feedback mechanism on the basis of the first two improvements. The feedback mechanism identifies and processes the LP images which are not detected correctly after the LP characters are recognized by OCR network. The way to deal with is to paint the misidentified area of the image black, essentially to set the three-channel value of the corresponding area to zero. Then, we return the processed image to the detection network for re-detection, and generally get the correct results.

The remainder of this work is organized as follows. In Section 2 we introduce how to realize LP detection and recognition by the related technique [1]. In Section 3 we describe in detail the improvements we have made on the basis of related technique. We modify the loss function of CNN and add a feedback mechanism. The overall evaluation and final results are presented in Section 4. Finally, in the Section 5, the conclusion of this paper was summarized and some opinions on future work are put forward.

II. Related Technique

The related technique [1] use You Only Look Once (YOLO) network [7], [8] to detect numbers and locations of vehicles in an image, and deal with the extracted vehicles separately. The method using sliding window approaches or candidate filtering coupled with CNNs can be find in [2], [3]. Because of the lack of shared computing, Faster RCNN [4] and Mask RCNN [9] are inefficient. In this paper, we use YOLOv2 network (his second version) to detect the vehicles in the image. Then each individual vehicle image is input into the detection network to detect and rectify its LPs. At last, all the corrected LP images are sent to OCR network to recognize characters above. We can understand the whole process intuitively through Fig. 1.

The specific algorithm is as follows:

Step 1: Considering some standards, the existing YOLOv2 vehicle detection system is still used to perform vehicle model shaping in images.

Step 2: Send the detected vehicle images to Warped Planar Object Detection Network(WPOD-NET) [1] for LP detection and correction.

Step 3: The corrected LP images are sent to OCR network for LP character recognition.

A. Vehicle Detection by YOLOv2

Because YOLOv2 network has the advantages of fast execution speed (about 70 FPS), good precision and recall

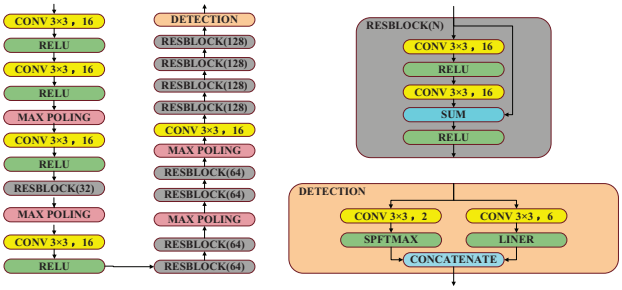


Fig. 2. Detailed WPOD-NET architecture.

compromise, the first step we input images into YOLOv2 network to detect vehicles. We haven't made any changes or improvements to the YOLOv2 network. We just think of it as a black box, integrating the output associated with vehicles (cars and buses) and ignoring other categories. The second step, the LP is recognized and corrected by WPOD-NET. Next, it explains in detail how WPOD-NET recognizes LPs from random images and corrects the direction, shape and size of LPs.

B. WPOD-NET

The structure of WPOD-NET is shown in Fig. 2. There are 21 convolutional layers, where 14 of are in residual blocks [10]. The size of all convolutional filters is fixed in 3×3. In addition to detection blocks, ReLU activations are used throughout the entire network. During the whole network, there are four max pooling layers with the size of 2×2, and stride 2 can reduce the dimension of the input network by 16 times. The detection block has two parallel convolutional layers: the first one is used to infer whether the area contains a LP object, activated by the softmax function. The second one is used to infer the six matrix parameters needed for affine transformation. It is activated by linear function, which is equivalent to output directly without activation function (which is equivalent to not using activation function).

Loss Function: Let $P_i = [x_i, y_i]^T, 1 = 1, ..., 4$, donate the fours corner points of an annotated LP, clockwise from the upper left corner. Let $q_1 = [-0.5, -0.5]^T$, $q_2 = [0.5, -0.5]^T$, $q_3 = [0.5, 0.5]^T$, $q_4 = [-0.5, 0.5]^T$, represents the corresponding vertex of a standard unit square centered on the origin.

For the input image with H height and W width, its the network stride size is given by $N_s = 2^4$(four maximum pooling layers), and the characteristic graph of the output of the network is composed of $M \times N \times 8$, in which $M = H/N_s$and $N = W/N_s$. For each point cell(m, n)in the feature graph, eight values are estimated: the first two values (v1 and v2) are the probability of objects or non-objects, and the last six values (v3 to v8) are used to construct the local affine transformation T_{mn} of the image. The computational formul of T_{mn} is (1), where v3 and v6 are used to maximize the function to ensure that the diagonal is positive (to avoid unwanted mirroring or excessive rotation).

$$T_{mn}(\boldsymbol{q}) = \begin{bmatrix} max(v_3, 0) & v_4 \\ v_5 & max(v_6, 0) \end{bmatrix} \boldsymbol{q} + \begin{bmatrix} v_7 \\ v_8 \end{bmatrix} \quad (1)$$

In order to match the output resolution of the network, $\boldsymbol{P_i}$ points are re-scaled by the reciprocal of the network stride and re-centered according to each point in the element diagram. This is accomplished by applying normalized function A_{mn}. The computational formul of A_{mn} is as follows:

$$A_{mn}(\boldsymbol{P}) = \frac{1}{\alpha} \left(\frac{1}{N_s} \boldsymbol{P} - \begin{bmatrix} n \\ m \end{bmatrix} \right) \quad (2)$$

Among them, α is the scaling ratio constant representing the square side length, which divides the maximum and the minimum of the LP dimension by the average point between the network steps. This paper order $\alpha = 7.75$ is used to enhance training data.

Assuming that there is an object (LP) in the point cell (m, n), the first part of the loss function f_{affine} is to consider the error between the deformed version of the specification square and the standardized annotation point of LP, which is given by the following equation:

$$f_{affine}(m, n) = \sum_{i=1}^{4} \left\| T_{mn}(\boldsymbol{q_i}) - A_{mn}(\boldsymbol{P_i}) \right\|_1 \quad (3)$$

The second part of the loss function f_{probs} deals with the probability of having or not non-objects in (m, n), which is given by the following equation. It is similar to SSD the loss of signal-to-noise, essentially the sum of two logarithmic loss functions.

$$\begin{aligned} f_{probs}(m, n) =& logloss\left(I_{obj}, v_1\right) \\ &+ logloss\left(1 - I_{obj}, v_2\right) \end{aligned} \quad (4)$$

The I_{obj} is the object indicator function. If there is an object within the point cell (m, n), it returns number 1, otherwise it returns number 0, in which $logloss(y, p) = -yloss(p)$. If the rectangular bounding box of an object presents IoU greater than the threshold Υ_{obj} (Υ_{obj} is set to 0.3 based on experience), the object is considered to be within the point cell (m, n). The boundary box compared with the above box has the same size, (m, n) is the center.

The final loss function $loss$ is given by the combination of terms defined in the above two equations:

$$loss = \sum_{m=1}^{M} \sum_{n=1}^{N} \left[I_{obj} f_{affine}(m, n) + f_{probs}(m, n) \right] \quad (5)$$

C. LP Character Recognition by OCR

Character segmentation and recognition on the whole LP is carried out through the modified YOLO network, with the same architecture presented in [6].

III. Proposed Method

In this paper, we have improved the existing LP monitoring and correction methods, which can make the input images recognize the LP faster than before under the same recognition accuracy. Besides, based on this method, an improved method with better LP detection accuracy is proposed.

A. Network Simplification

We do not construct affine transformations, but change the construction of loss function in (1). In the WPOD-NET, for each point cell (m, n) , a total of eight values are estimated for its corresponding four vertex coordinates. In this paper, the last six values of the network output (v3 to v8) are taken as the coordinates of the three vertices of clockwise from the upper left corner. So the new loss function T_{mn} equation is obtained as follows:

$$T_{mn}(i) = \begin{bmatrix} v_{i+2} \\ v_{i+5} \end{bmatrix} \qquad i = 1, ..., 3 \quad (6)$$

With the change of (1), (3) also changes as follows:

$$f_{l1}(m, n) = \sum_{i=1}^{3} \left\| T_{mn}(\boldsymbol{i}) - A_{mn}(\boldsymbol{p_i}) \right\|_1 \quad (7)$$

The final loss function $loss$ is given by a combination of the terms defined in (4) and (7):

$$loss = \sum_{m=1}^{M} \sum_{n=1}^{N} \left[I_{obj} f_{l1}(m, n) + f_{probs}(m, n) \right] \quad (8)$$

When we finish the training network model, we will get the coordinates of the three vertices counted clockwise from the upper left corner, and then predict model based on them. For the shape of the LP is mostly rectangular, the LP in the images are almost parallelogram when imaging. Parallelogram can infer the coordinates of the fourth point $point_4$ on the premise when three vertex coordinates are known. According to the characteristics of parallelogram, we can get:

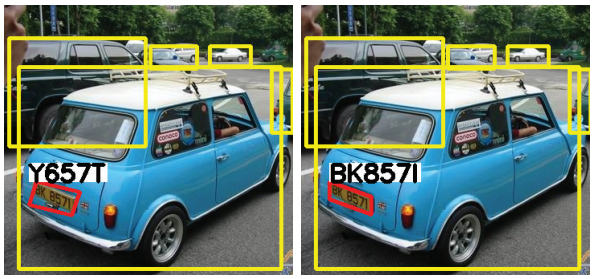
$$point_4 = point_1 + point_3 - point_2 \quad (9)$$

Thus, we can get four vertex coordinates of a LP, and the tilted LP can be corrected by perspective transformation of the LP image according to the four vertex coordinates.

B. Comparison of Recognition Effect

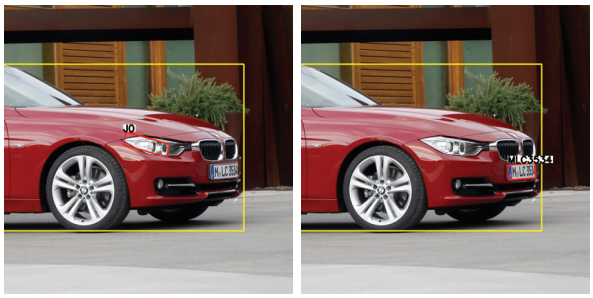
From Fig. 3 we can see that our Simplified Warped Planar Object Detection Network (SWPOD-NET) method is more accurate than the WPOD-NET of LP detection, and the four vertices of the LP are more accurate. The WPOD-NET is not very accurate in finding the second and fourth vertices. So there will be an error when transferring to OCR network for LP character recognition.

As can be seen from Fig. 4, WPOD-NET mistakenly detects the headlamp of a car as LP, so the recognized characters are naturally incorrect. SWPOD-NET can accurately detect LP, input LP into OCR for character recognition, and finally get the correct LP characters.



(a) WPOD-NET (b) SWPOD-NET

Fig. 3. WPOD-NET(L) and Ours Method(R).



(a) WPOD-NET (b) SWPOD-NET

Fig. 4. WPOD-NET(L) and Ours Method(R).

C. Network Improvement

Based on the above operations, we can find a drawback. The LPs detected base both on two method are parallelogram. But for some special angles or special situations, when the LP is irregular quadrilateral in imaging, the above two methods will produce errors. Therefore, we have made new improvements to SWPOD-NET's Detection Layer. We improve the network by changing six parameters into eight, and making the values of these eight parameters correspond to the coordinates of the four vertices of the LP one by one. The Detection structure of Improved Warped Planar Object Detection Network (IWPOD-NET) can be seen from Fig. 5 is as follows:

So we need to modify loss funtion Equation (6) as follows:

$$T_{mn}(i) = \begin{bmatrix} v_{i+2} \\ v_{i+6} \end{bmatrix} \qquad i = 1, ..., 4 \tag{10}$$

The Equation (7) has changed as follows:

$$f_{l1}(m,n) = \sum_{i=1}^{4} \left\| T_{mn}(\boldsymbol{i}) - A_{mn}(\boldsymbol{p_i}) \right\|_1 \tag{11}$$

The final loss function is still shown in Equation (8).

D. Network Improvement Effect

As can be seen from Fig. 6, the LP in the image is not an approximate parallelogram because of the distortion of the LP image. WPOD-NET by affine transformation and SWPODNET by parallelogram complementation cannot map to the four vertices of LP correctly. Our IWPOD-NET is not limited to finding parallelograms, so it can correctly find irregular LPs.

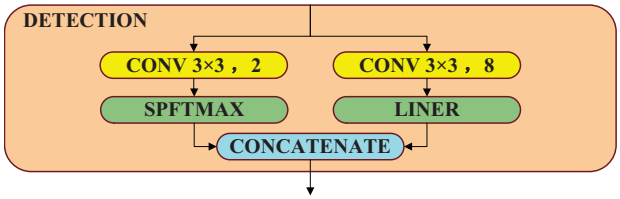
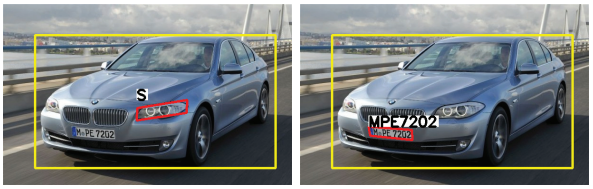


Fig. 5. IWPOD-NET Detection structure.



(a) WPOD-NET (b) IWPOD-NET

Fig. 6. WPOD-NET(L) and Ours Method(R).

As can be seen from Fig. 7, when the LP is partially covered or when the LP is small and irregular, the WPOD-NET can not find the LP position, and the wheel is mistakenly detected as the LP. But IWPOD-NET can find the LP position correctly and recognize it.

E. Feedback Mechanism

After improving the WPOD-NET, the LP images with some special shapes can be recognized more accurately than the WPOD-NET. But there are still clear pictures that mistake other parts of the car for LPs. Based on this, we make further improvements. In other words, a feedback regulation mechanism is added at the end of OCR recognition network. The flow chart is shown in Fig. 8.

The IWPOD-NET/SWPOD-NET detects the LP and sends it to OCR network for recognition. When the range of LP is detected incorrectly (for example, the lamp is detected as a LP), OCR generally detects no more than three characters. We set the three-channel pixel value of the detected area to zero when the recognized character is less than a certain number (if the parameter is 4, we can modify the parameter in the network), even if it turns black. Then the processed image is returned to the IWPOD-NET/SWPOD-NET for re-detection. Repeat the above process until the recognition character meets the set conditions or the maximum number of repetitions (we set the maximum number of repetitions to be 4).

We can clearly understand the process of LP detection and recognition by the Improved Warped Planar Object Detection Network with Feedback(IWPODF-NET) in Fig. 9.

IV. Experimential Result

A. Data Set

The data set we use for training and testing is a part of Cars Dataset [11]. The Cars dataset contains 16,185 images

(a) WPOD-NET

(b) IWPOD-NET

Fig. 7. WPOD-NET(T) and Ours Method(B).

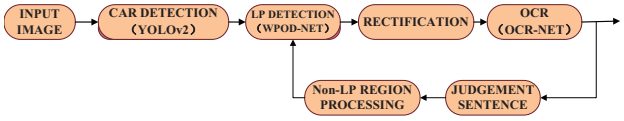


Fig. 8. The whole process with Feedback Regulation.

of 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images, where each class has been split roughly in a 50-50 split. Classes are typically at the level of Make, Model, Year, e.g. 2012 Tesla Model S or 2012 BMW M3 coupe.

Test pictures we selected 832 images with clearer LPs from 8144 training pictures as the raw data for our method comparison. And we set these 832 pictures as CD832. From CD832, we selected 106 pictures of vehicles under extreme angle as another contrast option. We designated these 106 pictures as EX106.

As shown in Fig. 10 is part EX106 of the picture, the LPs in the picture are in an extreme tilt state. It is very difficult for the computer to detect and recognize the LP. Our method can recognize more LPs than the traditional method.

B. Training Mode

In explaining the improvement of the method, we give a comparison of the LP detection and recognition results for some pictures. Next, we compare the three methods only for LP detection. Both vehicle detection and OCR adopt the same method. Our training set is 105 pictures in Cars Dataset with four vertex coordinates of the tagged LP. We trained the network with 300k iterations of mini-batches of size 64 using the ADAM optimizer [12].

C. The Result of Comparison

In addition to the above database, we selected 22 basically positive and clear LP images from 105 training pictures as the comparison of other parameters. From (3), (7), (11) we can see that our loss function is L1 norm, so we use three methods (WPOD-NET, SWPODF-NET and IWPODF-NET) to detect the LP of these 22 images, and compare the L1 norm by four vertices detected with the original training data. We compare

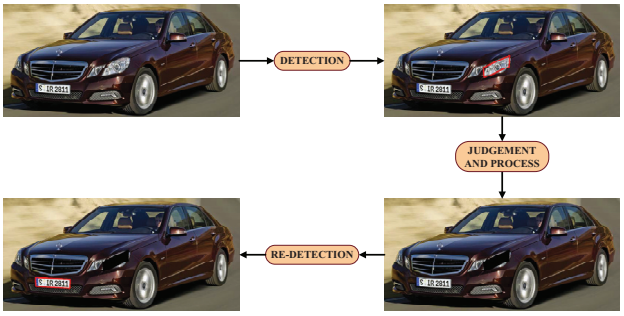


Fig. 9. The whole process with Feedback Regulation.

Fig. 10. A Part of EX106.

the average L1 norm of each vertex, the average L1 norm of the first three vertices and the average L1 norm of the four vertices of 22 pictures.

From the results of Table 1, we can see that SWPOD-NET can be greatly improved compared with the WPOD-NET when evaluating a large number of data sets (CD832). Although IWPOD-NET does not improve the result recognition too much, we find that when the number of training iterations is relatively small(for example 200K), it can achieve better results than 300K iterations. In our opinion, 105 training pictures are still too few for unconstrained training of four vertices, so the fitting phenomenon will occur when the number of iterations is too many. When the number of training pictures increases, we think that IWPOD-NET will achieve better results than the other two methods under the same training conditions.

In the detection of LP images under extreme conditions (EX106), our methods with feedback mechanism (SWPODF-NET & IWPODF-NET) is far better than the two methods without feedback mechanism (SWPOD-NET & IWPOD-NET).

At the same time, we compare the accuracy of three methods(WPOD-NET, SWPOD-NET and IWPOD-NET) to detect the pixel level of the effect by detecting the less difficult pictures. The reason why we only choose these three methods is that the difficulty of image detection is low, so the feedback mechanism will not be triggered, so whether there is

TABLE I
ACCURACY EVALUATION RESULTS

	WPOD -NET [1]	SWPOD -NET	IWPOD -NET	SWPOD F-NET	IWPOD F-NET
CD832	87.02%	88.82%	88.58%	91.11%	92.91%
EX106	78/106	84/106	79/106	86/106	90/106

TABLE II
LOSS EVALUATION RESULTS

	WPOD-NET [1]	SWPOD-NET	IWPOD-NET
$L_1$1st	1.47	1.74	1.6
$L_1$2nd	1.68	1.08	1.29
$L_1$3th	1.84	1.66	1.68
$L_1$4th	1.54	2.1	1.77
L_1first3	1.66	1.49	1.53
L_1all4	1.63	1.64	1.59

a feedback mechanism or not has the same result. From Table 2, we find that for the L1 norm of the whole four vertices, the SWPOD-NET is not much different from the WPOD-NET, and IWPOD-NET is more accurate than the two methods. The SWPOD-NET uses the coordinates of the first three points for training, so the detection accuracy of the point in the middle of the three points is more accurate than that of the other two methods. Relatively, the coordinate accuracy of the fourth point calculated from the other three points without training will be a little worse. The results of training at any three points are similar to this one, so we only give the results of the first three points.

V. CONCLUSION

This section describes the experimental analysis of the complete ALPR system, as well as the comparison of the most advanced methods and commercial systems. We focus on LP detection in extreme environments, not just in positive situations.

On the premise of the WPOD-NET, we first simplify the network, remove the affine transformation from six to four coordinates in the network, and focus on finding three points in the LP vertex. In this way, we can often find the selected three points more accurately. On the premise that the LP in the image is approximately a parallelogram, we can accurately calculate the coordinates of the fourth point. In the course of the experiment, the effect of training any three points is basically the same. Although the errors between the coordinates of the point in the middle of the three detection points and the actual coordinates are small, and the errors between the coordinates inferred from the diagonal point of the middle point and the actual coordinates are slightly larger (no more than the error range), which does not actually affect the results of the whole LP detection.

Secondly, the SWPOD-NET cannot achieve good results for the LP images distortion caused by the reasons of LP shooting. Because both WPOD-NET and SWPOD-NET are essentially looking for a LP approximating a parallelogram. At this time, we put the coordinates of the four points into the training on the basis of the SWPOD-NET. When the coordinates of the four points are not constrained, they can correctly detect the LP under more extreme conditions..

Finally, on the basis of these two methods, we add a feedback mechanism. The feedback mechanism identifies and processes the LP images which are not detected correctly after the LP characters are recognized by OCR network. We return the processed image to the detection network for re-detection, and generally get the correct results.

For future work, we hope to modify the OCR part and output fewer strings on the recognition error detection image, so that the feedback mechanism can be judged more accurately.

REFERENCES

[1] Montazzolli. Silva. S and Rosito Jung. C, "License Plate Detection and Recognition in Unconstrained Scenarios", In Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 580-596.

[2] Bulan, Orhan, et al. "Segmentation- and Annotation-Free License Plate Recognition With Deep Localization and Failure Identification." IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 9, pp.2351-2363, 2017.

[3] Kurpiel, Francisco Delmar, Rodrigo Minetto, and Bogdan Tomoyuki Nassu. "Convolutional neural networks for license plate detection in images." international conference on image processing, 2017, pp. 3395-3399.

[4] Ren, Shaoqing, et al. "Faster R-CNN: towards real-time object detection with region proposal networks." neural information processing systems, 2015, pp. 91-99.

[5] Li, Hui, et al. "Reading car license plates using deep neural networks." Image and Vision Computing, 2018, pp. 14-23.

[6] Silva, Sergio Montazzolli, and Claudio Rosito Jung. "Real-Time Brazilian License Plate Detection and Recognition Using Deep Convolutional Neural Networks." brazilian symposium on computer graphics and image processing, 2017, pp. 55-62.

[7] Redmon J, Divvala S K, Girshick R, et al. "You Only Look Once: Unified, Real-Time Object Detection." computer vision and pattern recognition, 2016, pp. 779-788.

[8] Redmon, Joseph, and Ali Farhadi. "YOLO9000: Better, Faster, Stronger." computer vision and pattern recognition, 2017, pp. 6517-6525.

[9] Kaiming H , Georgia G , Piotr D , et al. "Mask R-CNN." international conference on computer vision, 2017, pp. 2980-2988.

[10] He, K., Zhang, X., Ren, S., Sun, J. "Deep Residual Learning for Image Recognition." computer vision and pattern recognition, 2016, pp. 770-778.

[11] https://ai.stanford.edu/ jkrause/cars/car_dataset.html

[12] Kingma, Diederik P. , and J. Ba . "Adam: A Method for Stochastic Optimization." arXiv: Learning (2014).