High-speed Railway Real-time Localization Auxiliary Method based on Deep Neural Network

Dongjie Chen[1, 2, b)], Wensheng Zhang[1, 2, a)] and Yang Yang[1, c)]

[1]*Institute of Automation, Chinese Academy of Sciences*, [2]*University of Chinese Academy of Sciences*

[a)] *Corresponding author: zhangwenshengia@hotmail.com*
[b)]*chendongjie2013@ia.ac.cn*
[c)]*yang.yang@ia.ac.cn*

Abstract. High-speed railway intelligent monitoring and management system is composed of schedule integration, geographic information, location services, and data mining technology for integration of time and space data. Assistant localization is a significant submodule of the intelligent monitoring system. In practical application, the general access is to capture the image sequences of the components by using a high-definition camera, digital image processing technique and target detection, tracking and even behavior analysis method. In this paper, we present an end-to-end character recognition method based on a deep CNN network called YOLO-toc for high-speed railway pillar plate number. Different from other deep CNNs, YOLO-toc is an end-to-end multi-target detection framework, furthermore, it exhibits a state-of-art performance on real-time detection with a nearly 50fps achieved on GPU (GTX960). Finally, we realize a real-time but high-accuracy pillar plate number recognition system and integrate natural scene OCR into a dedicated classification YOLO-toc model.

I. INTRODUCTION

During the past twenty years, China High-speed Railway monitoring system has already accumulated massive amounts of monitoring data including whole data of catenary components. In catenary components, the underground base and pillar are fixed together in order to guarantee catenary stability. As shown in Fig.1, high-definition camera is installed to the train roof and it captures the pillar images from far to near. Each pillar refers to a unique point which can help promoting the train localization accuracy. Developed methods are not well up to standard requirement because of the limitation of special targets representation, calculation speed and even other actual application problems. Taking various factors into actual consideration, we choose deep CNN to achieve real-time multiple targets detection ultimately. The serial numbers of pillar plate along each railway line conform to an encoding rule that each number corresponds to a unique geographical location. The position of the railway pillar can be identified automatically through identifying serial number by text OCR methods [1-4], and then the algorithm confirms real-time location of the train along the railway, which will greatly help high-speed railway localization.

The contributions of this paper are as follows: 1) our main contribution is a practical, real-time, high-precision, robustness, and an end-to-end natural scene recognition system adapting to the complex and changeable environment for high-speed railway intelligence management system. 2) Our second contribution is proposing a real-time high-speed railway positioning auxiliary methods, a deep convolution neural network for high-precision localization of pillar plate, and real-time pillar plate number recognition by using natural scene OCR method.

II. REAL-TIME LOCALIZATION AUXILIARY METHOD

The first stage is the high-speed railway pillar plate text spotting. Inspired by YOLO [10], we propose YOLO-toc network predicting bounding boxes and associated class probabilities directly from global image. This model avoids the computational complexity of generating candidate boxes with exhaustive overmuch box filtration. The pillar plate labels are defined as $T=\{t_1,t_2,\cdots,t_K\}$, where K is the number of plate classes. As show in Figure 1, the image is divided into $S_1 \times S_2$ grids as the candidate bounding boxes. YOLO-toc is a much smaller and faster network designed for text OCR containing multiple convolution layers, pooling layers, full-connected layers. Two typical activation function is used in the CNNs. The linear activation function is applied for the final connected layer. The leaky rectified linear activation is defined as follow:

$$\phi(x)=\begin{cases} x & \text{if } x>0 \\ 0.1x & \text{otherwise} \end{cases}$$

In our model, each grid cell predicts $B=2$ bounding boxes and corresponding object confidence scores C and label probabilities P. Each bounding box consists of 5 predictions: x, y, w, h and confidence. The (x,y) coordinates represent the center of the box relative to the bounds of the grid cell. The width and height (w,h) are predicted relative to the whole image. The confidence scores reflect the degree of bounding box contains a target and label probabilities show the accuracy that the box is belong to which target. Formally the object confidence P is defined as the intersection over union (IOU) between the predicted box and the ground truth. The label probabilities P represented by a vector is similar with the classification problem in neural network. The final layer of our network is a long conventional vector which can be transformed into a $S_1 \times S_2 \times (B \times 5 + K)$ tensor of predictions.

The sum-squared error is chosen to be the loss function for convenience. In order to remedy the equal weights for different grids, the loss from bounding box coordinate predictions is increased and the loss form confidence prediction for boxes that do not contain objects is accomplished this. Two different parameters $\lambda_{coord}=5$ and $\lambda_{noobj}=0.5$ are set to accomplish this. Generally, one bounding box prediction is assigned to be responsible for predicting an object based on which prediction has the highest current IOU with the ground truth.

During training we optimize the following, multi-part loss function:

$$L=\lambda_{coord}\sum_{i=1}^{S}\sum_{j=1}^{B}[\![_{ij}^{obj}\left[\left(x_i-\hat{x}_i\right)^2+\left(y_i-\hat{y}_i\right)^2\right]+\lambda_{coord}\sum_{i=1}^{S}\sum_{j=1}^{B}[\![_{ij}^{obj}\left[\left(\sqrt{w_i}-\sqrt{\hat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h}_i}\right)^2\right]$$
$$+\sum_{i=1}^{S}\sum_{j=1}^{B}[\![_{ij}^{obj}\left(C_i-\hat{C}_i\right)^2+\lambda_{noobj}\sum_{i=1}^{S}\sum_{j=1}^{B}[\![_{ij}^{obj}\left(C_i-\hat{C}_i\right)^2+\sum_{i=1}^{S}[\![_{ij}^{obj}\sum_{k=1}^{2}\left(P_{ik}-\hat{P}_{ik}\right)^2$$

Where $[\![_{i}^{obj}$ denotes if object appears in cell i and $[\![_{ij}^{obj}$ denotes that the jth bounding box predictor in cell i is responsible for that prediction. The loss function only penalizes classification error if a target is present in that grid cell and bounding box coordinate error if that predictor is responsible for the ground truth box.

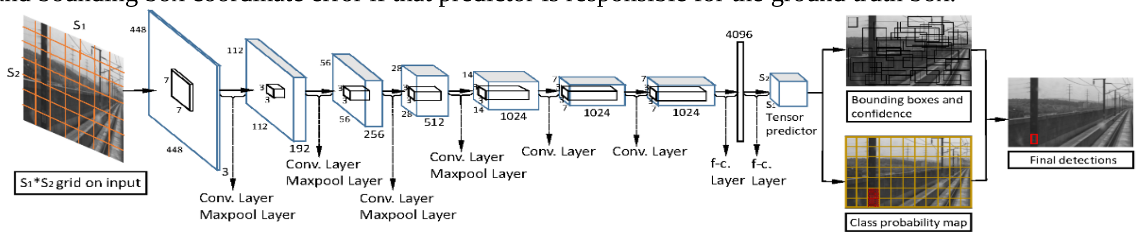


FIGURE 1. Architecture of YOLO-toc network

The second stage of our work is the pillar plate text recognition. The details of high-speed railway pillar plate text OCR are demonstrated in Figure 2. The procedure of OCR can be divided into three parts: preprocessing, character segmentation, and character recognition. Firstly, we have completed the grayscale image binarization in OCR part. Secondly, we segmented the whole word into several single characters. At last, the single character was recognized accurately through the lexicon look-up.

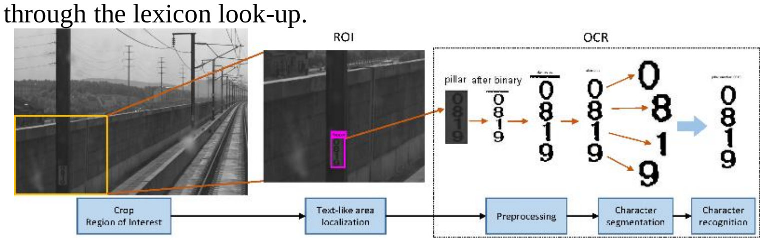


FIGURE 2. Natural scene OCR procedure for high-speed railway pillar plate

III. EXPERIMENT AND ANALYSIS

A. *Experiment setup*

Comparative methods: In order to verify the performance of the YOLO-toc network, we demonstrate an evaluation of different methods for text spotting. We choose Template Match, DPM [5], and Faster R-CNN [6-9] methods as test benchmarks.

Evaluation metric: To evaluate the performance of the proposed automatic pillar plate number recognition, we use the non-contact inspection system to capture the original images. In order to be consistent with the real world condition and to evaluate the performances of different approaches, the two situations (i.e. normal & darkness) are both taken into consideration during our experimental study.

We use Precision and Recall as the metric for performance evaluation, which are defined as follows:

$$\text{Precision} = \frac{\textit{the number} \text{ of correctly detected pillar plate}}{\textit{total} \text{ number of detected pillar plate}}, \quad \text{Recall} = \frac{\textit{the number} \text{ of correctly detected pillar plate}}{\textit{total} \text{ number of pillar plate}}$$

B. *Result and analysis*

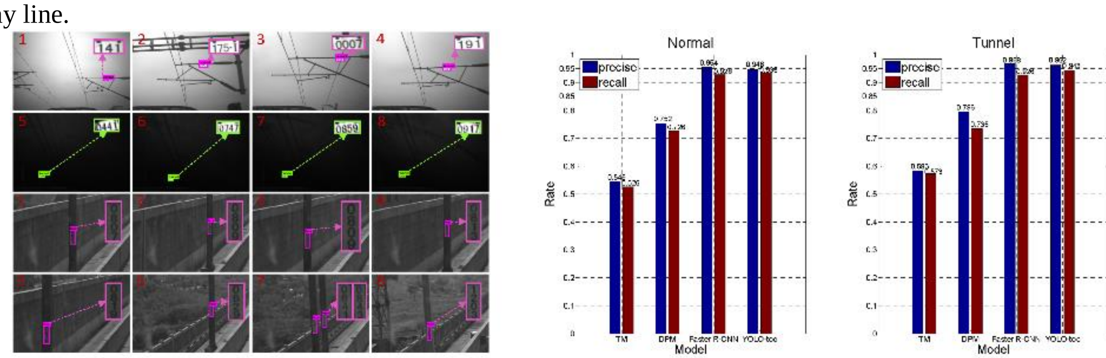
The experimental result is reported in Table.1 and Figure.3(a). YOLO-toc has good performance on the experimental dataset, it exhibits excellent accuracy and efficiency both on detection precision and calculation speed. It is quite obvious that the average accuracy of text detection based on YOLO-toc is much higher than other models, nearly reaching a precision of 95.5%. YOLO-toc model performs significantly better than DPM method on several classes. The text detection based on Faster R-CNN was eliminated because of its low framerate. According to the experiment result, we can draw a conclusion that the YOLO-toc network has a good performance on multiple experimental datasets with a nearly 45fps using a GPU (GTX960).

TABLE 1. Text detection performance of different methods

Method	**Precision (%)**	**Recall (%)**	**Framerate(fps)**
TM	54.6	56.8	0.25
DPM	76.3	78.6	6.2
Faster R-CNN	95.6	96.5	16.5
YOLO-toc	95.5	96.2	45.5

Pillar plate text consists of consecutive numbers according with a regular rule. We can get a real-time location of the running train once the pillar plate is recognized accurately along the railway line. Figure.3(b) displays comparison of the pillar plate OCR precision in different environments for different methods. The normal environment denotes the situation which original image sequences captured outside at day. The precision of pillar plate recognition in tunnel environment is nearly the same as in normal environment.

After the first step done, we use natural scene OCR technique to recognize the pillar plate number. At the second step, binarization is performed with different methods for scene text. At the last step, a category called "double-detection" is developed to promote text recognition precision. We try the category for the purpose of optimizing text recognition precision. Pillar plate text consists of consecutive numbers according with a regular rule. We can get real-time location of a running train once the pillar plate number is recognized accurately along the railway line.



(a) (b)

FIGURE 3. (a) Pillar plate text detection result; (b) Comparison of text recognition in different situation

The performance of our proposed framework for operational use is shown in Figure 4. Results on number recognition from the whole dataset show that our scheme can achieve excellent performance combined with OCR method. Besides, natural scene OCR is error-prone in a larger percentage than conventional OCR and a post-processing correction solution is necessary. We plan to start the precision refinement based on prior knowledge in the future work.

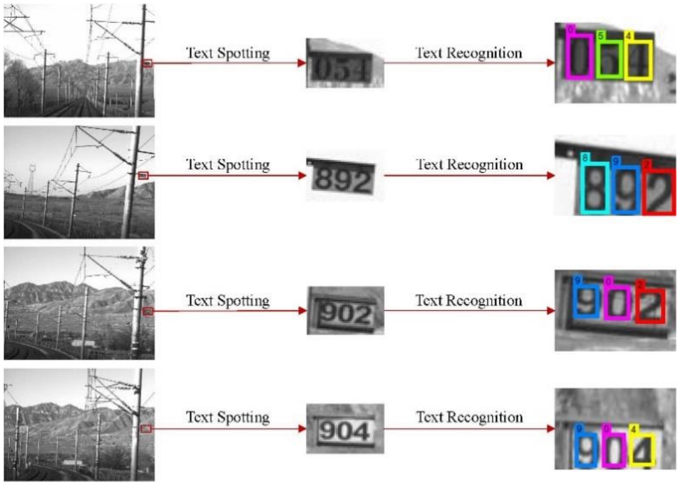


FIGURE 4. Pillar plate number recognition result

IV. CONCLUSION

We propose a novel framework and solution for high-speed railway localization auxiliary for the CHR intelligence monitoring system, furthermore, it promotes the high-speed railway intelligent management greatly. Monitoring data analysis is greatly helpful to intelligent high-speed operation management. Pillar plate location and number recognition methods based on YOLO-toc model show excellent properties such as real-time, high-accuracy and robustness. In the future work, we still need more research on high-speed railway monitoring system, such as the connections between the frames before and after the current frame which called sequence learning.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant No.U1636220, Grant No.61602484, Grant No.61502495, and Grant No.61432008. We gratefully acknowledge the support of Chengdu Sino Rail Electronics Corporation with the association of data and devices used for this research.

REFERENCES

1. Jaderberg M, Simonyan K, Vedaldi A, et al. Reading Text in the Wild with Convolutional Neural Networks[J]. International Journal of Computer Vision, 2016, 116(1):1-20.
2. Porat S, Carmeli B, Domany T, et al. Dynamic masking of application displays using OCR technologies[J]. Ibm Journal of Research & Development, 2009, 53(6):10:1-10:14.
3. Huang W, Qiao Y, Tang X. Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees [M]// Computer Vision – ECCV 2014. Springer International Publishing, 2014.
4. Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition[J]. Eprint Arxiv, 2014.
5. Felzenszwalb P, Mcallester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. DBLP, 2008:1-8.
6. Uijlings J R R, van de Sande K E A, Gevers T, et al. Selective search for object recognition [J]. International journal of computer vision, 2013, 104(2): 154-171.
7. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2015, 37(9): 1904-1916.
8. Girshick R. Fast r-can[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.
9. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016:1-1.
10. Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. Computer Science, 2015:779-788.