Real-time Traffic Sign Text Detection Based on Deep Learning

Ximing Peng¹, Xianqiao Chen¹,²,* and Chang Liu¹
¹College of Computer Science and Technology, Wuhan University of Technology, Wuhan, China
²Hubei Key Laboratory of Transportation Internet of Things, Wuhan University of Technology, Wuhan, China

*Corresponding author e-mail: 294407816@qq.com

Abstract. Aiming at the task of traffic sign text detection in natural scenes, a two-stage cascade detection model based on deep learning is proposed. The proposed model first locates the regions of interest (RoI) of text-based traffic sign by applying improved SSD (Single Shot MultiBox Detector) network. Then a rotation-based text detection network is used to detect text strings in the located RoI. Moreover, the lightweight convolutional neural network MobileNetV2 is combined with the deep learning component in two stages, which reduces network parameters and improves detection speed of the model. On the one hand, the proposed approach takes full advantage of the information of traffic signs. By this way the search area of text detection can be reduced, which makes it possible to simplify the text detector. On the other hand, it can obtain more complete text detection results by using rotation-based text detector. The experimental results show that the proposed method can perform well on different data sets, it can not only keep a high accuracy of detection, but also meets the real-time requirements.

1. Introduction

As part of modern transportation facilities, traffic signs play an important role in ensuring road safety and guiding driving. The common traffic signs can be divided into two types: symbol-based traffic signs and text-based traffic signs. However, at present, most researches of traffic sign detection [1-6] mainly focus on symbol-based traffic signs, such as vehicle indicating, no parking and limit speed sign, while the detection of text in traffic signs has received less attention [7-9]. Since text-based traffic sign are abundant and can provide plenty of valuable road information in the form of text, it is still worthy of studying. Besides, with the development of intelligent vehicles, the perception requirements of road environmental information are also increasing, so automatic detection of traffic sign text will be an indispensable part of intelligent vehicles.

Some researchers have proposed several methods for traffic sign text detection by making use of the characteristics of traffic signs, their basic idea is to extract the traffic sign area, and then detect the text lines in the extracted area. In the past years, Greenhalgh et al. [7] proposed a method that can localize and recognize traffic sign texts from video. Firstly, a combination of MSERs and HSV color thresholding are applied to localize candidate regions of traffic signs, and then MSERs is used to detect the text strings. After that, the detected text lines are recognized by OCR. Rong et al. [8] proposed a top-down model for automatic detection and recognition of traffic panels, which contains
two deep learning network. After that a deep recurrent model is trained to recognize the text lines. Zhu et al. [9] present a cascaded segmentation-detection network to detect traffic sign texts of different languages. Traffic sign area is segmented by a FCN-based network, and a fast text detection network is trained on multilingual data set to detect text lines.

Compared with the traditional methods, the deep learning method can perform better on traffic sign text detection, but there are still some shortcomings. For one thing, deep learning-based networks for object detection usually have a huge amount of parameters, which lead to insufficient memory and slow model detection speed. For another thing, the horizontal-specific text detection methods used in [8] and [9] cannot obtain complete text lines when the text regions are not horizontal. In this paper, we mainly focus on text-based traffic sign, and propose a novel a two-stage cascade detection model, which is shown in Fig. 1. Firstly, combined with MobileNetV2 [10], an improved SSD [12] network is proposed for quickly locating traffic sign areas. Secondly, we apply a rotation-based network inspired by EAST [13] to detect the text lines. Finally, extensive experiments are carried out to verify the effectiveness of the proposed method.

![Figure 1. The pipeline of the proposed method for traffic sign text detection](image)

2. MobileNetV2

In this paper, we apply MobileNetV2 to optimize our model. MobileNetV2 uses the depthwise separable convolutions structure, as shown in Fig. 2, it has two convolution layers: depthwise convolution layer and pointwise convolution. Firstly, depthwise convolution is used to filter to each input channel by applying a single convolution. Secondly, pointwise convolution with convolution kernel size of 1×1 is responsible for achieves the sparse representation by computing a linear combination of input channels.

![Figure 2. Depthwise separable convolution](image)

Based on residual net, an “residual” structure is designed to solve the gradient disappearance and gradient explosion, it makes skip connections in different separable convolutional layers. However, as residual structure with compression, convolution feature extraction and expansion mode is used
together with the depthwise separation convolution, there will be a problem that the amount of features extracted during the second step of the convolution feature extraction of the bottleneck is too small. Thus, the original residual structure is improved in MobileNetV2, and it becomes the process of expansion, convolution feature extraction, and compression.

3. Traffic sign detection

In this section, a real-time traffic sign detection network based on SSD is proposed. The advantages of SSD include three parts:

The design of multi-scale feature maps. In SSD, a variety of feature maps are designed to predict the locations and confidences in multiple convolutional layer, which can detect different sizes of targets. Especially, the feature maps from the shallower layer are applied for small target, and feature maps from deeper layer for big target.

The design of default boxes. SSD uses plenty of default boxes in different feature maps to fit the target position. Briefly, SSD generates a large number of bounding boxes for each position in the feature map. These bounding boxes have different sizes and aspect ratios according to the position and scale of the layer in which they are located.

Detection and recognition. In SSD, target category and location can be predicted together, which makes it faster than two-stage methods. A set of convolution kernels are added behind several different scales feature maps to get a fixed set of test results.

3.1. The improved SSD for traffic sign detection

The parameters of VGG-16 used in SSD are 138M and the trained model can only perform well on GPU. Therefore, combined with MobileNetV2, we proposed an improved SSD for traffic sign detection. In traffic sign detection network, the basic feature network is replaced with MobileNetV2, fully connected layer and pooling layer of which is removed. After that, four convolutional layers (Extras IRblock_1~4) composed of the inverted residual blocks are added as extra feature layers. After MobileNetV2 and four additional feature layers are added, two 3*3 convolution kernels are used for convolution operation on 6 feature maps (Inverted Residual 32*32, Inverted Residual 16*16, Extras IRblock_1 8*8, Extras IRblock_2 4*4, Extras IRblock_3 2*2, Extras IRblock_4 1*1). The output one is used for classification, while the other is used for bounding box regression. Table 1 lists the parameters of traffic sign detection network.

| Layer          | Channel | Repeat | Stride | Size    |
|----------------|---------|--------|--------|---------|
| Conv2D         | 32      | 1      | 2      | 256*256 |
| Inverted Residual | 16     | 1      | 1      | 256*256 |
| Inverted Residual | 24     | 2      | 2      | 128*128 |
| Inverted Residual | 32     | 3      | 2      | 64*64   |
| Inverted Residual | 64     | 4      | 2      | 32*32   |
| Inverted Residual | 96     | 3      | 1      | 32*32   |
| Inverted Residual | 160    | 3      | 2      | 16*16   |
| Inverted Residual | 320    | 1      | 1      | 16*16   |
| Extras IRblock_1      | 512    | 1      | 2      | 8*8     |
| Extras IRblock_2      | 256    | 1      | 2      | 4*4     |
| Extras IRblock_3      | 256    | 1      | 2      | 2*2     |
| Extras IRblock_4      | 128    | 1      | 2      | 1*1     |

3.2. Training of traffic sign detection

When training the proposed traffic detection network, we adopt the loss function of SSD. In SSD algorithm, plenty of default boxes in different feature maps are designed to predict the target bounding boxes. The default boxes are divided into positive and negative samples when training. Positive sample means the default box matched with the real target box, while negative sample means not matched. Let $x$ be the matching result between default boxes and real target position (for the $i$-th
default box and the \( j \)-th real target position, \( x_y = 1 \) means positive sample, \( x_y = 0 \) means negative sample.), \( c \) be the confidence of classification, \( l \) be the predicted offsets, and \( g \) be the real target position. The loss function can be defined as follow:

\[
L(x,c,l,g) = \frac{1}{N} \sum \left( L_{conf}(x,c) + \alpha L_{loc}(x,l,g) \right)
\]

where \( N \) represents the quantity of positive samples, \( L_{conf}(x,c) \) is 2-class soft-max loss, and \( L_{loc}(x,l,g) \) stands for the smooth L1 loss. \( \alpha \) is a constraint to adjust \( L_{conf} \) and \( L_{loc} \), which is set to 1 in experiment.

4. Text detection in traffic sign

By applying the results of Chapter 3, we can mainly detect text lines in a small region. Most of traffic sign text lines are horizontal, but they can be non-horizontal (Fig. 3 (a)) because of the various angle of view. Due to the large oriented angle of the texts, we may get a bad segment result (Fig. 3 (b)) by using horizontal bounding boxes. To solve this difficulty, we use a rotation-based text detector inherited from EAST to obtain entire text line (Fig. 3 (c)).

![Figure 3. Two types of residual structure](image)

4.1. The improved EAST for text detection

Fig. 4 shows the framework of the improved EAST for text lines detection, which is combined with three parts: feature extraction stem, feature merging branch and prediction module.

![Figure 4. The framework of the proposed improved EAST](image)

Firstly, we use MobileNetV2 network for the feature extraction stem to get different feature maps of the original image. In original EAST network, four convolution layers of the MobileNetV2 yield four scales of feature maps defined as \( f_i = 1, 2, 3, 4 \), whose sizes are \( 16*16*384 \), \( 32*32*256 \), \( 64*64*128 \), \( 128*128*64 \) for an input image of \( 512*512*3 \). This means that the feature map \( f_i \) is used to predict the small target whose width is around \( 1/32 \) of the width of the original image. However, there are no such small texts in the traffic sign, because the size ratio between texts and their corresponding traffic signs are large enough to be easily identifiable. Therefore, we changed the four scales of feature maps to \( 1/24, 1/12, 1/6, 1/3 \) of the input image, which is more suitable for detecting text on traffic signs.
Secondly, lower feature map is combined with higher feature map. By this way, different information of multi-scale of feature map can be fused together. The specific calculation process uses the following formula:

\[
g_i = \begin{cases} 
  \text{unpool}(h_i) & \text{if } i \leq 3 \\
  \text{conv}_{vs}(h_i) & \text{if } i = 4 
\end{cases}
\]  

(2)

Where \( g_i \) represents the feature map before merged, \( h_i \) stands for the feature map after merged.

Finally, the prediction layer outputs score map and geometry maps. It contains a 1*1 convolution kernel for the score map \( F_s \), five 1*1 convolution kernels for the geometry maps \( F_g \). Among the convolution kernels of geometry maps, four of them contain the corresponding score which is the distance from the pixel to the four borders of target box \( R \), and another one encodes the rotation angle \( \theta \) of target box. After that, we set a threshold on the score map, dropping candidates with lower scores. At last, we apply Non-Maximum Suppression (NMS) to remove the redundant target box, and get the text location results from the input image.

4.2. Training of text detection

The total loss \( L \) of the proposed text detection network consists of two parts: the loss \( L_s \) on the score map and the loss \( L_g \) on the geometry maps, which is formulated as:

\[
L = L_s + \lambda_g L_g
\]

(3)

\[
L_s = 1 - \frac{2\sum_i \hat{T}_i T_i'}{\sum_i \hat{T}_i + \sum_i T_i'}
\]

(4)

\[
L_g = -\log\text{IoU} (\hat{R}, R') + \lambda_\theta (1 - \cos (\hat{\theta} - \theta'))
\]

(5)

where \( \lambda_g \) is the weight to balance \( L_s \) with \( L_g \) and is set to 1.0 in our experiment. \( \hat{T} \) stands for the predicted confidences of classification, and \( T' \) is the real target on the score map. \( N \) is the total number of the scales. \( \text{IoU} (\hat{R}, R') \) is the ratio of intersection area to union area between the predicted location \( \hat{R} \) and the real target location \( R' \). \( \hat{\theta} \) and \( \theta' \) represent the predicted and the real rotation angle of text box respectively. Besides, \( \lambda_\theta \) is a constraint to adjust the rotation angle loss and the location loss, and is set to 15 when training.

5. Experiment

5.1. Experimental dataset

In this paper, our work focuses on text-based traffic sign, and the data sets used in our experiment are Traffic Guide Panel dataset collected from the highway, TT 100K and our self-collected dataset. There are 2315 traffic images containing various types of traffic panels in Traffic Guide Panel dataset, and the image size is 480*360. Among them, 1526 images are selected as training data for traffic sign detection network, and the rest are relabeled with text region bounding boxes to verify the proposed method. TT 100K is a benchmark dataset for traffic sign detection, but it also contains a lot of text-based traffic signs. Thus, we collect 910 images containing text-based traffic signs from TT 100K. Besides, we collect 1324 images from Internet for our experiment. The 2234 images we collected are also applied to train and test our model. We label all images with traffic sign bounding box and text line bounding box. Among them, 379 images are randomly selected for testing while others for training.

5.2. Implementation details

Our simulation hardware involves Gigabyte GTX1080Ti gaming OC 11G and INTEL I7 8700K processor. The software system is Ubuntu 18.04, CUDA10.0 and Pytorch.
The training process of the model can be divided into two parts. Our first part is to train the traffic sign detection network. Our proposed traffic sign detection network uses pre-trained MobileNetV2 as backbone network. During training, we use mini-batch SGD to optimize the training process and the batch size is set to 16. Images are resized to 512×512, learning rate initializes at 2 × 10^{-4}, and decreases progressively on epoch 60, 90 and 120 by the factor of 10. We use a weight decay of 0.0005 to avoid overtraining, and set the momentum to 0.9. The full training is composed of 150 epochs. In addition, we follow the data augmentation scheme used in original SSD when training, which includes photometric distortions and geometric distortions.

The second part is to train the text detection network. The pre-trained MobileNetV2 is also used in the text detection network. To make the network more robust, we first train it on ICDAR2017 [13] dataset for 100 epochs. After that, we fine-tune the network for 80 epochs using our collected training data with the learning rate set to 0.0001. Besides, weight decay is set to 0.0003 while momentum to 0.95.

5.3. Evaluation Protocol

We use general evaluation measures for object detection, which evaluates the model performance by Precision Rate (PR), Recall Rate (PR), and F measure. The F measure is the harmonic mean value of Precision and Recall. They are computed as:

$$PR = \frac{N_{tp}}{N_{total\_target}}$$  (6)

$$RR = \frac{N_{tp}}{N_{ground\_truth}}$$  (7)

$$F = \frac{2 \times PR \times RR}{PR + RR}$$  (8)

where $N_{tp}$ stands for quantity the number of correct target being detected. $N_{total\_target}$ and $N_{ground\_truth}$ denotes the quantity of targets detected the number of ground truths respectively.

5.4. Experimental Results and Analysis

Evaluation of traffic sign Detection. To verify the effectiveness of the improved SSD network, we compare it with SSD300 and SSD512 shown in Table 2. Precision, Recall and F measure of the improved SSD can achieve 94.2%, 94.7% and 94.4% respectively in traffic sign detection. The Precision of the improved SSD is 1.1% lower than original SSD512, but the Recall of it is 2.4% higher than SSD512. More importantly, the detection speed of the improved SSD is roughly the same as SSD300. As shown in Fig. 5, our improved SSD can detect the traffic sign with various colors (Fig. 5 (a)-(c)), different illumination (Fig. 5(d), (e)) and partially blocked object (Fig. 5 (f)).

| methods       | PR       | RR       | F-measure | Time(ms) |
|---------------|----------|----------|-----------|----------|
| SSD300        | 0.892    | 0.86     | 0.875     | 23       |
| SSD512        | 0.953    | 0.923    | 0.938     | 46       |
| Our improved SSD | 0.942    | 0.947    | 0.944     | 26       |
Figure 5. The experiment results of the proposed traffic sign detector

Evaluation of traffic sign text detection. We compare our proposed method with other methods, the results are shown in Table 3. It can be seen that our method has better real-time performance and detection accuracy. As we can see, by applying EAST to detect traffic sign texts directly, it can obtain a relatively good performance by using the input size of 1080*1080. But it takes 480ms for its large input size so that it will be hard to meet the real-time requirement. By applying the step of traffic sign detection, our proposed method is more efficient than EAST because of plenty of complex background being removed. Meanwhile, the proposed model is lightweight because MobileNetV2 is applied on two deep learning component, and both the traffic sign detection network and the text detection network can perform well with fast detection speed. Several detection samples on different data sets are shown in Fig. 6.

Table 3. Comparison with different methods for traffic sign text detection

| Methods            | PR   | RR   | F-measure | Time(ms) |
|--------------------|------|------|-----------|----------|
| Ours               | 0.93 | 0.91 | 0.92      | 120      |
| Zhu et al. [8]     | 0.90 | 0.87 | 0.88      | 154      |
| Rong et al. [9]    | 0.73 | 0.64 | 0.68      | 167      |
| EAST(1080*1080)    | 0.83 | 0.86 | 0.84      | 480      |
| EAST(512*512)      | 0.75 | 0.71 | 0.73      | 294      |
| EAST(300*300)      | 0.52 | 0.47 | 0.49      | 92       |

Figure 6. The experiment results of traffic sign text detection model
6. Conclusion

In this paper, we propose a novel traffic sign text detection method. The improved SSD network can efficiently locate the region of traffic sign area, which greatly reduces difficulty in text detection. Meanwhile, the rotation-based text detector used in our model can detect the non-horizontal text lines. More importantly, MobileNetV2 is combined with the traffic sign detector and text detector to improve the detection speed so that our proposed model can meet the real-time requirements. In the future work, we will apply our model to an embedded platform. Moreover, text recognition will be considered by using the result of the traffic sign text detection.

7. Acknowledgments

This work was financially supported by National Key R&D Program of China (2018YFB1600400).

References

[1] Greenhalgh J, Mirmehdi M. Real-time detection and recognition of road traffic signs [J]. IEEE Transactions on Intelligent Transportation Systems. 2012, 13(4): 1498–1506.

[2] González Álvaro, Garrido Miguel Ángel G, Llorca D F, et al. Automatic Traffic Signs and Panels Inspection System Using Computer Vision[J]. IEEE Transactions on Intelligent Transportation System, 2011, 12(2):485-499.

[3] Stallkamp J, Schlipsing M, Salmen J, et al. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition[J]. Neural networks: the official journal of the International Neural Network Society, 2012,32:323-332.

[4] Mogelmose A, Trivedi M M, Møeslund T B. Vision-Based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(4):1484-1497.

[5] Zhu, Z, Huang, X L, Liang, D. Traffic-Sign detection and classification in the wild[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016: 2110–2118.

[6] Yuan Y, Xiong Z, Wang Q. An incremental framework for video based traffic sign detection, tracking, and recognition[J]. IEEE Transactions on Intelligent Transportation Systems. 2017, 18(7):1918-1929.

[7] Greenhalgh J, Mirmehdi M. Recognizing text-based traffic signs. IEEE Trans. Intell. Transp. Syst., 2015, 16(3): 1360-1369.

[8] Rong X, Yi C, Tian Y. Recognizing text-based traffic guide panels with cascaded localization network [C]. Proceedings of European Conference on Computer Vision(ECCV). Springer, Cham, 2016:109–121.

[9] Zhu Y, Liao M, Yang M, et al. Cascaded Segmentation-Detection Networks for Text-Based Traffic Sign Detection[J]. IEEE Transactions on Intelligent Transportation Systems. 2018. 19(1):209-219.

[10] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]. 2018 CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 4510-4520.

[11] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]. Proceedings of the European Conference on Computer Vision, Amsterdam, Oct 11-14, 2016. Berlin, Heidelberg: Springer, Cham. 2016: 21-37.

[12] Zhou X, Yao C, Wen H, et al. EAST: An Efficient and Accurate Scene Text Detector[C]. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017:2642 - 2651.

[13] Shi B, Yao C, Liao M, et al. ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17) [C]. Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2018:1429–1434.