An End-to-End Practical System for Road Marking Detection

Chaonan Gu, Xiaoyu Wu(✉), He Ma, and Lei Yang

Communication University of China, Beijing, China
wuxiaoyu_hai@163.com, {gcn,mahe,yanglei}@cuc.edu.cn

Abstract. Road marking is a special kind of symbol on the road surface, used to regulate the behavior of traffic participants. According to our survey, it seems that no papers has yet proposed a mature, highly practical method to detect and classify these important fine-grained markings. Deep learning techniques, especially deep neural networks, have proven to be effective in coping with a variety of computer vision tasks. Using deep neural networks to construct road marking detection systems is a practical solution.

In this paper, we present an accurate and effective road marking detection system to handle seven common road markings. Our model is based on the R-FCN network framework, with the ResNet-18 model as backbone. SE blocks and data balancing strategies are also used to further improve the accuracy of the detection model. Our model has made a good trade-off between accuracy and speed, and achieved quite good results in our self-built road marking dataset.

Keywords: Road marking · Detection · R-FCN · SE block · Median frequency balancing

1 Introduction

Road markings refer to the special symbols that are drawn on the surface of the road, usually including arrows, characters, etc., which play an important role in guiding, restricting, and warning to the traffic participants. Although the problem of road marking recognition is solved in [1], there seems to be no mature, highly practical method to detect road markings currently, which not only requires to classify but also to locate the markings. The detection of road markings has great practical value, that could be used in many fields such as GPS positioning, the decision making and path planning of automatic driving systems. It is of great practical significance to construct an accurate and efficient detection system.

It is very difficult for computers to locate and identify the road markings from the images. As shown in Fig. 1, the road scene is changeable. Affected by weather, illumination, angle of view and other factors, outdoor road images will change to a lot of different forms. Furthermore, the worn-out road markings, the complex urban road environment, and the markings occlusion, also increase the difficulty of road marking detection. According to our survey, there seems to be no published paper presenting a

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11902, pp. 85–93, 2019.
https://doi.org/10.1007/978-3-030-34110-7_8

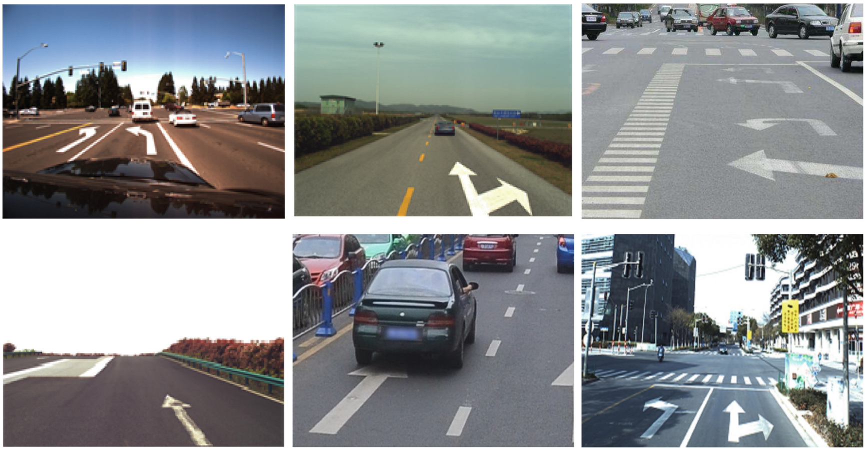

Fig. 1. Various types of road markings: under different lighting condition; worn out marking; marking occlusion.

practical classification and detection method, to deal with these important and variable fine-grained markings.

We plan to use deep learning algorithms to build our detection system, in which data plays a critical role. Currently, there are very few road marking datasets proposed. Existing datasets, such as Caltech, KITTI [2], Berkeley datasets, etc., contain only a small amount of image data, far from meeting the needs of deep learning. However, in a large dataset: Apollo [3], since the images are mainly extracted from the video, the changes between them are relatively small and cannot meet the requirements of diversity. In another large dataset, the Tsinghua-Tencent-100k dataset, some pictures are distorted, and only a few pictures contain road markings, which means it will take a lot of time to filter. All of these datasets have certain drawbacks. The lack of road marking data is a huge challenge for building our systems. In addition, there is another big problem in the data, that is, the numbers of various road markings are not balanced. The number of straight arrows is much larger than that of other classes, such as U-turns, which is very disadvantageous for deep learning.

At present, there are two main ideas for road marking detection algorithms using deep learning: based on object segmentation and based on object detection. The segmentation-based road marking detection algorithm (VPG [4], TT100k [5], Mask R-CNN [6]) spends a lot of computing resources on pixel-by-pixel segmentation, which results in long test times. However, the segmentation result is superfluous for road marking detection. Therefore, it is not suitable to construct a road marking detection system based on segmentation.

Currently, road marking detection algorithms based on object detection are more common. There are two main ideas in the field of object detection: one-stage networks, two-stage networks based on region proposal and classification [7–9]. Through our experiments, it is found that typical single-stage network structures, such as YOLO [10], SSD [11] and its improved method DSSD [12], although with high detection speed, are not accurate enough for detecting relatively small objects in images. It still has a certain gap from our requirements.

A two-stage detection network based on region proposal has higher accuracy than single-stage networks, but its detection speed is usually slow (e.g., Faster R-CNN [9]). Another two-stage network, the R-FCN model with two-stage network structure can achieve similar detection accuracy compared with Faster R-CNN while maintaining relatively fast speed. The R-FCN [13] model adopts fully convolutional network [14]. Through the design of the position-sensitive score map, all convolution calculations are shared by all candidate regions, thus getting a good trade-off between speed and accuracy. The R-FCN network is a relatively suitable basic model for road marking detection.

For the significant problems of detecting road marking: the insufficient and imbalance data, the lack of mature road marking detection model, we propose our own road marking detection system drawing on the SE (squeeze and excitation) blocks [15]: (i) we build our own dataset, which contains 40k images for seven common road markings; (ii) the basic model is R-FCN with ResNet-18 [16] as backbone; (iii) median frequency balancing is used to compensate for the adverse effects of imbalance data; (iv) SE blocks are adopted to further improve the accuracy of the model. Our model has a good balance between speed and accuracy, and achieves good results in self-built dataset.

2 Our Approach

The structure of our model is shown in Fig. 2. We use the R-FCN as basic model, with ResNet-18 as the backbone network. Two SE blocks are added to stage 4 and stage 5 of ResNet-18 to further enhance model performance with negligible effect on speed.

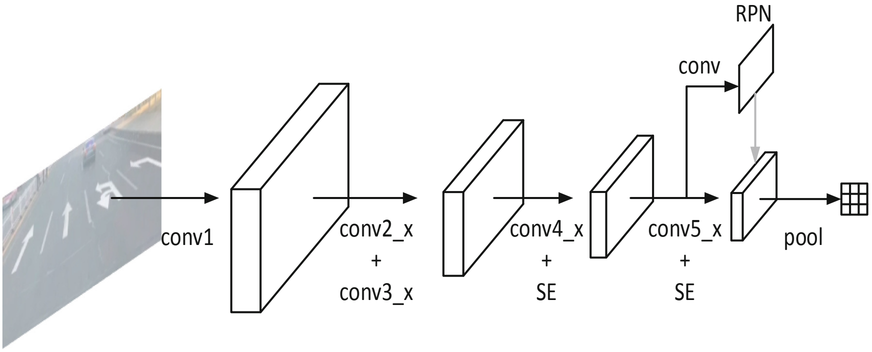

Fig. 2. Road marking detection model

2.1 Basic Model

We selected 3269 original images containing seven common road markings from the public dataset, and finally obtained our own dataset containing 40k images through data augmentation. The dataset is described in more detail in Sect. 3.1.

We build our road marking detection system based on the R-FCN network. Compared to other detection models, the R-FCN model is relatively fast, while still maintaining high accuracy. We use the residual network as the backbone network of the model. The detection model is fine-tuned using ResNet model pre-trained on ImageNet. Images were preprocessed before sent to the network. Since the road markings usually appear in the lower half of the image, the top pixels in the image do not contain the required information, so we cropped 1/3 of the image from the top. In order to achieve good trade-off between speed and accuracy, each cropped image is resized to 900 * 600.

We tested two residual network models, ResNet-50 and ResNet-18, and the experimental results are shown in Sect. 3.2. Compared to ResNet-50, the accuracy of ResNet-18 has dropped by 0.6%, but the time is shortened by half to 56 ms. The ResNet-18 network gets a better balance between speed and accuracy, making it more suitable for road marking detection.

2.2 Median Frequency Balancing

The dataset we get is not balanced. The number of straight arrows is much larger than the number of other classes of road markings, while, on the opposite, the number of U-turn arrows is much smaller than others. Since the goal of the neural network algorithm is to minimize the overall error rate, which may result in poor performance of classes with less instances. On the one hand, we try to ensure a balance of different road markings in the data augmentation process. On the other hand, we adopted median frequency balancing referring [17].

In RFCN [13], the loss of each RoI is defined as the summation of the cross-entropy loss and the box regression loss: $\mathrm{L}\big(\mathrm{s}, t_{x,y,w,h}\big) = L_{cls}(s_{c^*}) + \lambda\big[c^* > 0\big]L_{reg}(t, t^*)$. Here c^* means the ground-truth label of the RoI, s_{c^*} is the predicted probability of the RoI belonging to c^*. We keep the bounding box regression loss L_{reg} not changed, and change the cross-entropy loss of the classification to:

$$L_{cls}(s_{c^*}) = w_{c^*}[-log(s_{c^*})] \tag{1}$$

w_{c^*} is the weight the original loss of the RoI multiplied. It is defined as:

$$w_c = \frac{median(\{f_c | c \in C\})}{f_c} \tag{2}$$

Here, we have C classes of road markings to be detected, f_c is the frequency of the c-th category. We define f_c as:

$$f_c = \frac{Number\ of\ c\text{-}th\ instances\ in\ dataset}{Number\ of\ instances\ in\ dataset} \tag{3}$$

w_c varies from class to class to compensate for the adverse effects of the unbalanced data. The class with the most instances has the least weight, and the class with the fewest instances has the largest weight. By using the median frequency balancing strategy, the contribution of each class to overall loss will be more balanced. According to experimental, the accuracy of the model using the strategy is about 0.4% higher than the original model.

2.3 SE Blocks

Since the accuracy dropped with resnet18 as the backbone network, the SE module [15] is used to compensate. The literature [15] studies the relationship of channels and explicitly models the relationship between convolutional feature maps by using SE blocks. Through two operations, the model can use the global information to strengthen the beneficial channels and suppress the useless channels, thus achieving the characteristic response re-correction between the channels and improving the accuracy of our model.

The first operation of SE block is squeeze. Each feature map is compressed using global average pooling. The convolutional layer outputs a feature map of H*W*C, which will be compressed into a real sequence of 1*1*C through squeeze, so that the global information of the network can be utilized by all layers. The second operation of the SE block is called excitation. Using the gate mechanism, the output of the squeeze is transformed into a series of weights to measure the importance of each channel.

The SE block is easy to port and can be inserted into the non-linear operation after convolution of all standard models. Therefore, SE blocks can also be easily applied to ResNet-18. We place the SE blocks after the non-linear operation of the residual block, before the identity branch summation. In Sect. 3.4, we will examine in detail the impact of the position of SE blocks and the number of SE blocks.

Although each SE block introduces extra parameters, it does not significantly affect the detection speed while improving the accuracy of the model, which is very suitable for application to our model. The experimental results show that adding SE blocks to res4b and res5b respectively, the model achieves the best results. The accuracy was improved by 0.6% while maintaining almost the same detection speed.

3 Experiments

3.1 Dataset

Since existing datasets do not meet our research needs, due to their limitations, we have built our own road marking dataset shown in Table 1. We choose seven types of most common road markings as a study aims: right turn arrow, left turn arrow, straight arrow, straight or right turn arrow, straight or left turn arrow, U-turn arrow, and crosswalk.

Table 1. Weights for all classes

	Right turn arrow	Left turn arrow	Straight or right turn arrow	Straight or left turn arrow	Straight arrow	U-turn arrow	Crosswalk
Number of instances	10656	13171	10263	8550	23619	5994	15067
Weight	1.0	0.8204	1.0404	1.2511	0.4831	2.2189	0.7067

Then we picked out 3269 original images containing these road markings from public datasets such as KITTI, RoadMarking, Baidu Apollo and TT100k. Among these images, there are video frames extracted from the video sequence, and pictures taken in various different scenes to meet the requirements of different application scenarios. The pictures contain simple scenes such as highways, and also contains complex scenes such as urban areas. All original images are manually labeled.

We random selected 1k origin images as test set, others as training set. We adopted a series of data augmentation methods to improve the diversity and quantity of the training set. We adjust the brightness and contrast of the image to simulate the effects of different shooting times. We also rotated the original image by 5° to simulate the change in perspective. Other methods such as cropping and mirroring are also used. Images that do not match the actual situation have been deleted. Finally, we got a training of 40k images.

3.2 Basic Model

In general, deeper networks have abilities to learn more high-level abstract features. However, for road marking detection tasks, deeper network like the Resnet-50 network takes up a lot of computation time, reducing the practicability of the model. And because road markings are relatively simple, the features extracted by very deep network model are quite redundant. Network models with less layers like ResNet-18 may have better results.

Based on a single computer with a Nvidia 1080ti GPU, we studied the effects of different residual network models. The experimental results are shown in Table 2.

Table 2. Performance with different residual networks

	mAP (%)	Test time (ms)
ResNet-50	88.9	105
ResNet-18	88.3	56

According to the experiments results, the ResNet-18 model is sufficient enough for road marking feature extraction and does not lead to a sharp drop in accuracy. With the ResNet-18 network, the accuracy is dropped by about 0.6% to 88.3% compared to the model with ResNet50, and the time is shortened by half to 56 ms. This means that ResNet-18 is more suitable for road marking detection systems than ResNet-50.

3.3 Median Frequency Balancing

We explored the effect of median frequency balancing. The weights are calculated on the training set. The distribution of various class instances in the dataset as well as the calculation results of weights are shown in Table 1.

The effects of the median frequency balancing strategy are shown in Table 3. We can see that, whether with or without SE blocks, the accuracy of the model with median frequency balancing is improved compared to the model without it, which proves the validity and applicability of this data balancing strategy.

Table 3. The mAP of models with and without median frequency balancing

	Without (%)	With (%)
R-FCN + ResNet-18	88.3	88.8
R-FCN + ResNet-18 + SE	89.2	89.5

3.4 SE Blocks

We explore the impact of SE blocks in terms of both quantity and location. The reduction factor for all SE blocks is set to 16, which is consistent with [15]. The experimental results are shown in Tables 4 and 5.

Table 4. The impact of different SE block position

Position	mAP (%)
No SE block	88.3
Res2b	88.3
Res3b	88.5
Res4b	88.9
Res5b	88.6

Table 5. The impact of different number of SE blocks

Position	mAP (%)
No SE block	88.3
Res4b	88.8
Res4b+5b	89.2
Res3b+4b+5b	88.9

When using a single SE block, it works best on the second residual block of stage 4 called res4b which is the last residual block shared by the RPN subnetwork and the classification subnetwork. When two SE blocks placed in res4b and res5b respectively, the model achieves the highest accuracy, which is 0.6% higher than the model without any SE block. More than two SE blocks seem to have contributed nothing to improving accuracy.

Through calculation, two SE blocks in res4b and res5b add a total of 40k parameters. The model size increased by only 0.1M, and the test time on each image barely increases. Adding two SE blocks to the original model, combined with median frequency balancing in training process, we get the final model with an accuracy of 89,5% and 56 ms of test time for a single image, which means our model achieves a good balance between speed and accuracy.

In summary, ResNet-18, combined with median frequency balancing and SE module, ensures that our model has a relatively fast speed and a high accuracy. Results on different models are shown in Table 6, where MFB means median frequency balancing. The visualization of test result is shown in Fig. 3. Our model can deal with illumination changing, worn-out marking and marking occlusion relatively well.

Table 6. Results on different model

	mAP (%)
Basic model: R-FCN + ResNet-18	88.3
Basic model + MFB	88.8
Basic model + SE	88.9
Basic model + MFB + SE	89.5

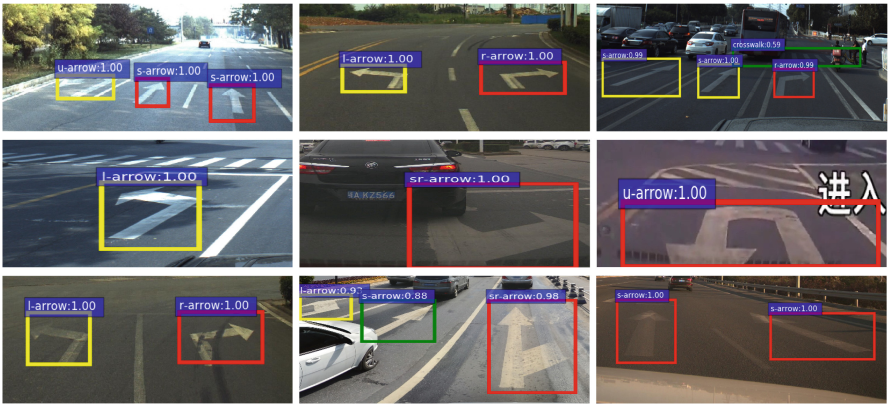

Fig. 3. Visualization of test results of our model

4 Conclusion

In this paper, in order to the deal with the problem of insufficient and imbalance data, we first built a dataset containing seven common classes of road markings. Secondly, we constructed a road marking model based on R-FCN structure with ResNet-18. We further improved the accuracy of the model through median frequency balancing and SE blocks. The experimental results show that our model has a good result on our self-built dataset, and achieves a good balance between testing speed and accuracy.

Acknowledgement. This work was supported by National Natural Science Foundation of China (61801441) and the Fundamental Research Funds for the Central Universities.

References

1. Touqeer, A., David, I., Ebrahim, E., George, B.: Symbolic road marking recognition using convolutional neural networks. In: Intelligent Vehicles Symposium (IV), pp. 1428–1433. IEEE, Los Angeles (2017)
2. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. **32**(11), 1231–1237 (2013)
3. Xinyu, H., et al.: The ApolloScape dataset for autonomous driving. In: Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 954–960. IEEE, Salt Lake City (2018)
4. Seokju, L., et al.: VPGNet: vanishing point guided network for lane and road marking detection and recognition. In: International Conference on Computer Vision (ICCV), pp. 1947–1955. IEEE, Venice (2017)
5. Zhe, Z., Dun, L., Songhai, Z., Xiaolei, H., Baoli, L., Shimin, H.: Traffic-sign detection and classification in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2110–2118. IEEE, Las Vegas (2016)
6. Kaiming, H., Georgia, G., Piotr, D., Ross, G.: Mask R-CNN. In: The IEEE International Conference on Computer Vision (CVPR), pp. 2961–2969. IEEE, Hawaii (2017)
7. Ross, G., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587. IEEE, Columbus (2014)
8. Ross, G.: Fast RCNN. In: The IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448. IEEE, Santiago Chile (2015)
9. Shaoqing, R., Kaiming, H., Ross, G., Jian, S.: Faster RCNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems(NIPS), Curran Associates, Montreal (2015)
10. Joseph, R., Santosh, D., Ross, G., Ali, F.: You only look once: unified, real-time object detection. In: The IEEE International Conference on Computer Vision (ICCV), pp. 779–788. IEEE, Las Vegas (2016)
11. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
12. Cheng-Yang, F., Wei, L., Ananth, R., Ambrish, T., Alexander, C.B.: DSSD: deconvolutional single shot detector. In: The IEEE International Conference on Computer Vision (ICCV), arXiv preprint arXiv:1701.06659 (2017)
13. Jifeng, D., Yi, L., Kaiming, H., Jian, S.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 379–387. Curran Associates, Barcelona (2016)
14. Jonathan, L., Evan, S., Trevor, D.: Fully convolutional networks for semantic segmentation. In: The IEEE International Conference on Computer Vision (ICCV), pp. 3431–3440. IEEE, Boston (2015)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. IEEE, Hawaii (2018)
16. Kaiming, H., Xiangyu, Z., Shaoqing, R., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, Las Vegas (2016)
17. Vijay, B., Alex, K., Roberto, C.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 (2015)