One Stage Detection Network with an Auxiliary Classifier for Real-Time Road Marks Detection

Guan-Ting Lin†1, Patrisia Sherryl Santoso*1, Che-Tsung Lin*ǂ, Chia-Chi Tsai† and Jiun-In Guo†

†National Chiao Tung University, Hsinchu, Taiwan
E-mail:{ ilovevictor0424, apple.35932003, jiguoccu}@gmail.com
* Industrial Technology Research Institute, Hsinchu, Taiwan
ǂ National Tsing Hua University, Hsinchu, Taiwan
E-mail: {sherrylsantoso, AlexLin}@itri.org.tw
1 =equal contribution

ABSTRACT

We construct a robust road mark detector that achieves high accuracy with real-time processing performance (32 fps) under nVidia Titan-X GPU. We combine one stage deep learning detector with auxiliary CNN classifiers as a robust road marks detector. We found out that one stage detector not only detects multiple objects via single inference efficiently, but also remains a good accuracy in performance perspective. However, to make it better, we add an extra CNN classifier as the back part of the proposed architecture to reduce false positive and get better accuracy. The proposed detector can achieve 86.8% mAP in our in-house six-class road mark database.

***Index Terms*— Road-mark detection, Real-time, CNNs**

1. INTRODUCTION

Autonomous and unmanned vehicles have been of great interests in recent years. However, researchers still encounter difficulties because this application involves many detection and recognition tasks. Perception of road marks is one-of-a-kind that plays an important role to recognize the driving environment. Road marks provide information to drivers that ensure safer circulation traffic flows. Detecting the road marks can reduce the possibility of traffic collisions. Therefore, we tackle and focus on this road-mark detection sub-problem in the research field of autonomous vehicles.

Most researchers nowadays decompose objects detection pipeline into two tasks: localization and classification. Moreover, they improve this pipeline performance with the use of convolutional neural networks to extract robust features. R-CNN [1] used Selective Search [2] as region proposals, crop fixed size images, then fed it into CNNs to extract its features and use SVM as a classifier. Fast R-CNN [3] made improvement by proposing ROI-pooling to eliminate SVM classifier and made the pipeline able to update the whole CNNs weights. Then, Faster R-CNN [4] was proposed to speedup previous two methods, by using Region Proposal Networks (RPN) as Selective Search replacement, which adapts CNN to generate region proposals. However, the speed of these complex pipelines still far from real-time performance.

Then, another work [5] improved the speed furthermore by combining Cascade-Adaboost classifier [6] as region proposals with CNN and achieved 100 fps performance. Although its succeed made improvement on speed, that pipelines are limited with fixed rotation angle instances only and has difficulties in detecting multi-class objects.

Yolo [7] is one of fast object detectors that successfully integrated detection problems into single regression problems to produce bounding boxes associated with class probabilities. Its unified architecture achieves not only fast, but also able to detect objects from different angles in inference stage. Nonetheless, we found out that this kind of detectors have difficulties to differentiate “Chinese-words” road mark sign and frequently produced false positives in this case.

Deep CNN based classifiers have significantly improved classification performance, such as AlexNet [8], GoogLeNet [9] and the powerful ResNet [10]. Then, smaller CNNs, SqueezeNet [11] was designed with less parameters yet achieved AlexNet-level accuracy. We adapt these classifiers on top of the detection backbone to overcome the false prediction issues. The proposed frameworks offer competitive trade-off between speed and accuracy.

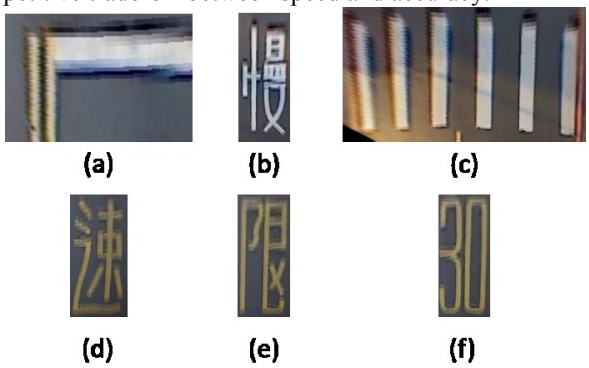

Figure 1: Examples of our six-class road mark objects

Table 1: One stage detection model results on our road mark data

Detector Settings		Stop line	Slow sign	Zebra Cross	Speed Limit			mAP	fps
					速	限	30		
GoogleNet	Original settings without BN	0.818	0.817	0.727	0.725	0.541	0.545	0.695	38.4
	B=3, Default grid size	0.907	0.727	0.726	0.815	0.633	0.808	0.769	38.2
	Grid size 14	Less than 0.5							-
	Original settings with BN ReLU	0.907	0.886	0.727	0.861	0.778	0.882	**0.840**	38.1
	Original settings with BN PReLU	0.906	0.814	0.725	0.634	0.540	0.726	0.724	37.9
	B = 3 , Grid size 14, BN ReLU	0.904	0.811	0.700	0.872	0.707	0.607	0.767	37.5
ResNet50	Original settings with BN PReLU	0.908	0.896	0.817	0.798	0.773	0.877	**0.845**	33.1

Table 2: Our system results on our road mark data

Detector Settings	CNNs Settings	Stop line	Slow sign	Zebra Cross	Speed Limit			mAP	fps
					速	限	30		
Googlenet Original with BN ReLU	S- 4cls	0.909	0.795	0.795	0.767	0.754	0.868	0.814	37.6
	G-4cls	0.907	0.779	0.793	0.694	0.789	0.771	0.789	34.7
	R-4cls	0.909	0.801	0.795	0.809	0.782	0.860	0.826	30.8
Googlenet with B = 3 , Grid size 14, BN ReLU	S- 7cls	0.893	0.765	0.756	0.807	0.774	0.780	0.796	37.1
	G-7cls	0.904	0.807	0.790	0.906	0.804	0.885	0.849	34.3
	R-7cls	0.902	0.799	0.756	0.900	0.788	0.767	0.819	32.4
ResNet 50 with Default boxes, BN PReLU	S- 4cls	0.909	0.879	0.811	0.847	0.784	0.883	0.852	32.8
	G-4cls	0.909	0.890	0.811	0.891	0.786	0.868	0.859	32.5
	R-4cls	0.909	0.890	0.811	0.891	0.786	0.868	0.859	32.2
	S- 7cls	0.897	0.856	0.785	0.892	0.793	0.885	0.851	32.7
	G-7cls	0.906	0.889	0.787	0.890	0.779	0.875	0.854	32.7
	R-7cls	0.907	0.900	0.805	0.903	0.797	0.898	**0.868**	32.0

2. OUR DATASET

In this work, we collected total 120k frames from real driving environments by a FOV 130-degree camera and convert the captured images into images with bird's-eye view perspective. Figure 2 shows the examples of our driving environments and bird's-eye transformed images. We randomly split it into training and testing parts; 70k images for training and 50k images for testing. Our dataset consists of frames annotated with bounding boxes of six commonly seen road mark classes, including stop-line Figure 1, slow sign 慢 Figure 1, zebra cross Figure 1, speed limit 速 Figure 1, speed limit 限 Figure 1, speed limit 30 Figure 1. All models and statistical results in this paper are trained and calculated on this database.

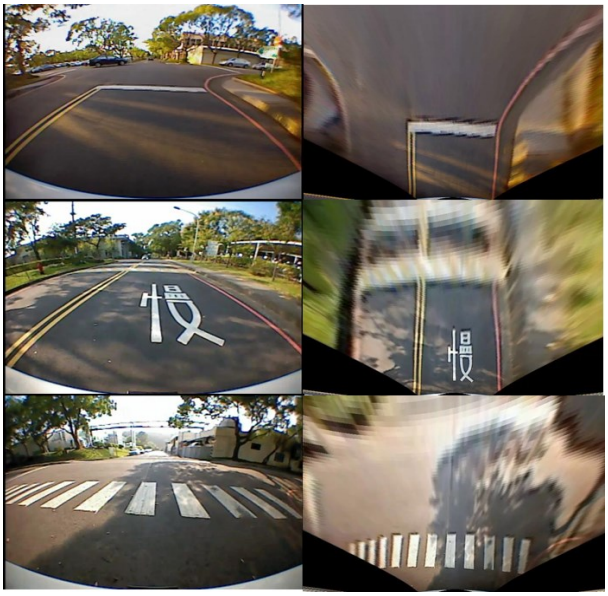

Figure 2: Example images of our database

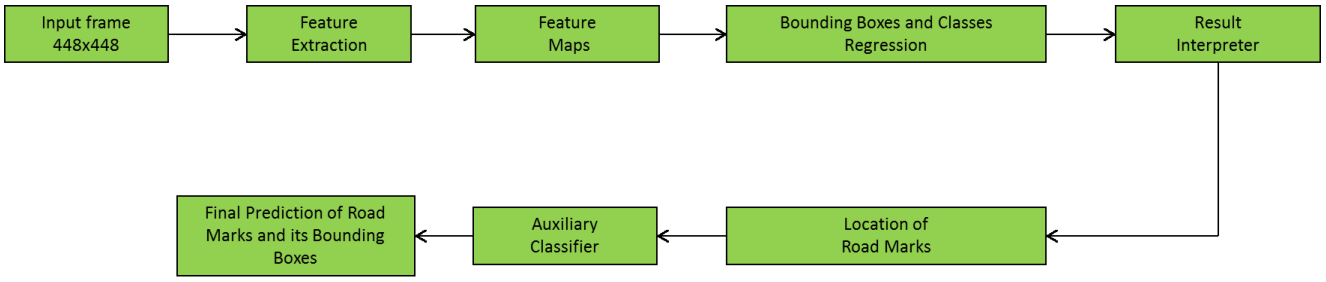

Figure 3: The system flow of the proposed design

3. DETECTION MODEL SETTINGS

We follow the YOLO regression mechanism for the proposed detection model. We used Googlenet and ResNet-50 as the base feature extractor. Beside the original settings, we also follow the setting from [12], such as set a grid size equal to 14; each cell predicts 3 bounding boxes and stack two more additional convolutional layers followed by batch normalization [13] with PReLU [14].

Table 1 shows the experimental results on the proposed detection model. It shows that PReLU activation setting is not working well on the base of GoogLeNet. The highest mAP is achieved by using the original settings followed with ReLU and batch normalization. However, when replacing GoogLeNet with the powerful ResNet-50, we can get the best result of mAP with PReLU and batch normalization followed by. All timing information is measured on NVIDIA DIGITS DevBox and Titan X GPU

4. AUXILLIARY CLASSIFIER

We adopted and compared few models for the auxiliary classifier in our frameworks, such as SqueezeNet, ResNet-18 and Googlenet(denoted as S, R and G in Table 2). Here, we tried two different training schemes for CNNs: (1) we use 4 classes "Chinese-words" sign {慢, 速, 限, 30} to train CNNs, (2) we use all 6 classes and we added another class as a negative class (total 7 classes).

5. DETECTION-CLASSIFIER SYSTEM

In this section, we combine one stage detector with an auxiliary classifier as backbone part of the proposed system. For the detection front part, we use two highest mAP models as mentioned in Section 3. For the first model, the results show that CNNs backbone did not give any improvement on mAP. We argue that it comes from the failed generation of boxes on some particular frame (insufficient proposal boxes). As the result, we further experimented with GoogLeNet base (B = 3, Grid size 14, BN ReLU), since it produces more boxes than those two with highest mAP. Table 2 shows experimental results of the proposed detectors.

Next, we tried with the second detection model. This architecture empowers the ability of the proposed detector to propose more box candidates than the first model. By combining it with CNNs, it provides significant improvement from its base model.

Then, we tried another larger model. This model is able to produce accurate boxes and achieved high mAP as described in Section 3. After we combined it with auxiliary classifier, we can get the increased mAP which is 2.3 % higher than the original one. The overall best result is achieved when it is combined together with ResNet-18 base (7 class scheme) classifier and scored 86.8% in mAP metrics. The overall system flow of the proposed design is graphically shown in Figure 3.

6. CONCLUSION

In summary, we have proposed a deep learning detector combining a detection-classification flow for real-time detecting road marks with Chinese words. We have compared different architectures for both detection {Googlenet and ResNet-50} and auxiliary classification back part {SqueezeNet, GoogLeNet and ResNet18} to achieve a best combination on the proposed design. The proposed system achieves high accuracy with 86.8% mAP in our in-house six-class road mark datasets.

7. REFERENCES

[1] R. B. Girshick, J. Donahue, T. Darrel and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 580-587, 2014.

[2] J. R. Uijilings, K. van de Sande, T. Gever and A. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision,* vol. 104, no. 2, pp. 154-171, 2013.

[3] R. Girshick, "Fast R-CNN," *International Conference on Computer Vision (ICCV),* 2015.

[4] S. Ren, K. He, R. Girschick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Neural Information Processing Systems (NIPS),* 2015.

[5] G.-T. Lin, P. S. Santoso, C.-T. Lin, C.-C. Tsai and J.-I. Guo, "Stop Line Detection and Distance Measurement for Road Intersection based on Deep Learning Neural Network," in *Asia-Pacific Signal and Information*

Processing Association Annual Summit and Conference (APSIPA ASC), 2017.

[6] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR),* 2001.

[7] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: unified, real-time object detection," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR),* 2016.

[8] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems (NIPS),* 2012.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR),* 2015.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR),* 2016.

[11] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally and K. Keutzer, "Squeezenet: AlexNet-level accuracy with 50x fewer parameters and model size," *CoRR,* vol. abs/1602.07360, 2016.

[12] C.-T. Lin, P. S. Santoso, S.-P. Chen, H.-J. Lin and S.-H. Lai, "Fast Vehicle Detector for Autonomous Driving," *International Conference on Computer Vision Workshop (ICCVW),* pp. 222-229, 2017.

[13] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning (ICML),* 2013.

[14] K. He, X. Zhang, S. Ren and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," *International Conference on Computer Vision (ICCV),* 2015.

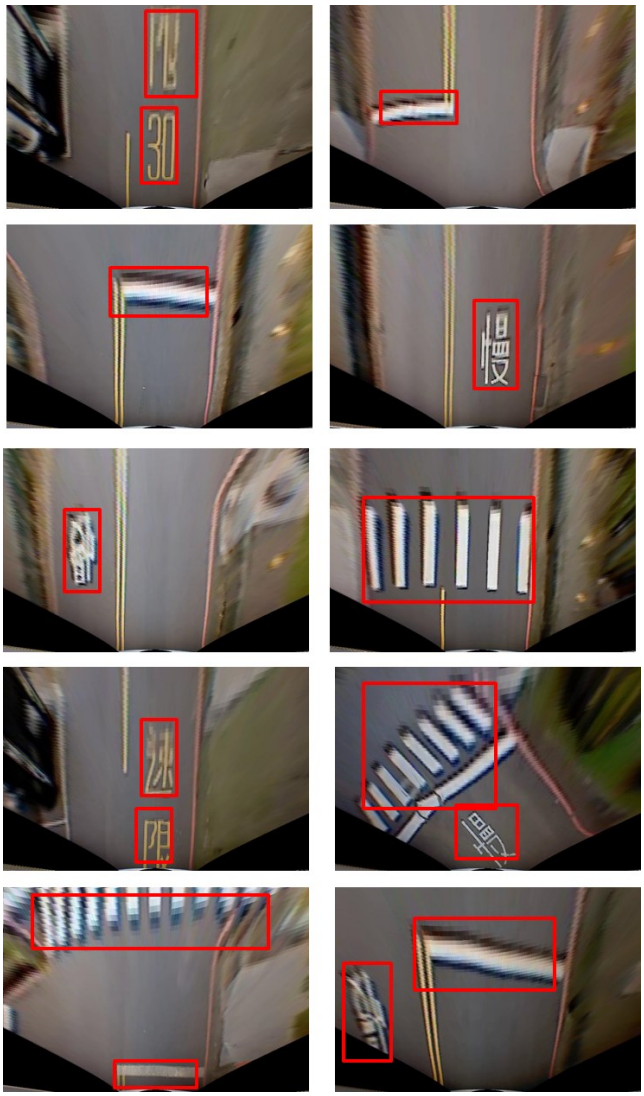

Figure 4: Detections results of road marks in the proposed design