Detection and Recognition of Arrow Traffic Signals using a Two-stage Neural Network Structure

Tien-Wen Yeh[1] and Huei-Yung Lin[2]

[1]*Department of Electrical Engineering, National Chung Cheng University, Chiayi 621, Taiwan*

[2]*Department of Electrical Engineering and Advanced Institute of Manufacturing with High-tech Innovations*
National Chung Cheng University, Chiayi 621, Taiwan

Keywords: Traffic Light Detection, Traffic Light Recognition, Arrow Traffic Signal.

Abstract: This paper develops a traffic light detection and recognition system based on convolutional neural networks for Taiwan road scenes. A two-stage approach is proposed with first detecting the traffic light position, followed by the light state recognition. It is specifically designed to identify the challenging arrow signal lights in many urban traffic scenes. In the detection stage, the map information and two cameras with different focal lengths are used to detect the traffic lights at different distances. In the recognition stage, a new method combining object detection and classification is proposed to deal with various light state classes in Taiwan road scenes. Furthermore, an end-to-end network with shared feature maps is implemented to reduce the computation time. Experiments are carried out on the public LISA dataset and our own dataset collected from two routes with urban traffic scenes.

1 INTRODUCTION

At present, the advanced driver assistance systems (ADAS) or autonomous driving vehicles have achieved fairly good results in simple environments such as highways and closed parks. To further improve the autonomous driving capability, dealing with more complex scenarios including the urban roads need to be addressed. In these cases, it is required to have the ability to sense the traffic conditions. One of the important issues is to detect the traffic lights and understand their states for driving instructions. Although the vehicle positioning by GPS combined with GIS mapping can provide the rough road junction information, the exact locations and states of the traffic lights are not guaranteed to be precisely marked in the HD maps. Thus, online detection and recognition of traffic signals are essential for driving assistance or autonomous vehicles.

Traffic light detection approaches are mainly divided into two categories. One is to make traffic lights have the capability to communicate with nearby vehicles through the V2I (Vehicle-to-Infrastructure) communication framework (Abboud et al., 2016). The other is to detect the positions and light states of the traffic lights by the onboard sensors of the vehicles. In general, these two approaches have both pros and cons. However, the former requires the installation and replacement of basic equipment and infrastructure, which is usually more expensive compared to the sensing based methods adopted by individual vehicles.

The traffic light detection and recognition techniques using the onboard sensors of vehicles have been investigated for many years (Jensen et al., 2016). Early methods mainly apply computer vision and image processing algorithms to the videos captured by the in-vehicle camera. In recent years, the learning based approaches have become more popular because of the public availability of the large collection of driving data (Waymo, 2019; Caesar et al., 2019; Ramanishka et al., 2018). However, the detection accuracy is still not satisfactory due to many disturbance factors in the outdoor environment. A few issues for the image-based traffic light detection methods are as followings.

- Images with hue shift and halo interference due to other light sources.
- (Partial) occlusion due to other objects or oblique viewing angles.
- Incomplete light shape due to sensing malfunction.
- False positives from reflection, billboard, pedes-

Yeh, T. and Lin, H.
Detection and Recognition of Arrow Traffic Signals using a Two-stage Neural Network Structure.
DOI: 10.5220/0009345203220330
In *Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2020)*, pages 322-330
ISBN: 978-989-758-419-0

trian crossing light, etc.

- Dark light state due to unsynchronized light duty cycle and camera shutter.

These problems are difficult to solve by computer vision algorithms one by one. But with the success of deep learning and the use of convolutional neural networks (CNN) for traffic light detection (Weber et al., 2016), it is expected to have more powerful feature extraction capability to deal with these issues.

In this work, we present a traffic light detection system for Taiwan road scenes based on deep neural networks. There are two technical challenges in our application scenarios. First, the public datasets currently available contain traffic lights arranged vertically, which are different from the horizontally arranged traffic lights we are dealing with. Second, the arrow signals are very common in Taiwan's traffic scenes, but most existing works use classifiers to recognize the circle lights only. These problems lead to the solution of creating a self-collecting dataset. Moreover, the data unbalance among multiple classes due to the lack of sufficient arrow light images further complicates the network training. Thus, a new method by combining the object detector and classifier for light state recognition is proposed. This two-stage approach first detects the light position, followed by the classification on the types of the arrow lights. Finally, the traffic light detection network is integrated to an end-to-end model with feature maps sharing. Experiments carried out on the public LISA dataset and our dataset report better results compared to the previous works.

2 RELATED WORK

2.1 Classical Traffic Light Detection

The early developments of image-based traffic light detection are mostly based on conventional computer vision techniques (Fregin et al., 2017a). The input images are converted to different color spaces, and various features such as color, shape, edge and gray-level intensity are used for detection. In the later machine learning based approaches (Kim et al., 2011), image features such as HoG or Harr-like operators are adopted for SVM or AdaBoost classification techniques. There also exist techniques using multiple sensor inputs. Fregin *et al.* presented a method to integrate the depth information obtained using a stereo camera for traffic light detection (Fregin et al., 2017b). Alternatively, Müller *et al.* presented a dual camera system to increase the range of traffic light detection with different focal length settings (Müller et al., 2017). They used a long focal length camera to detect the far away traffic lights, while a wide-angle lens camera is adopted to detect the close by traffic lights.

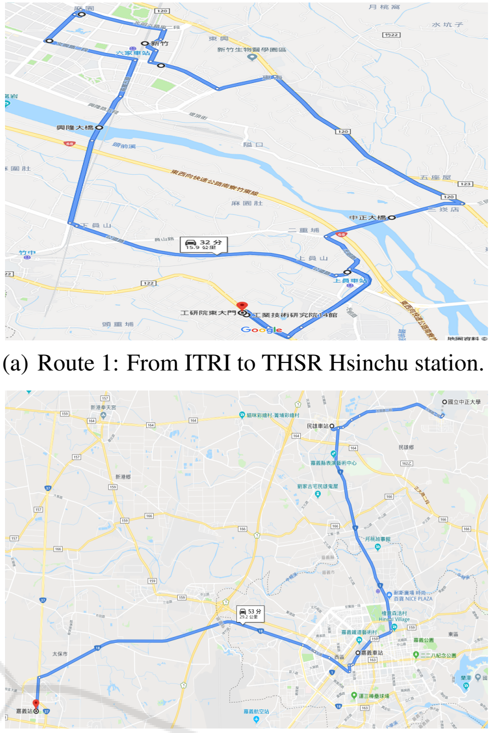
2.2 Traffic Light Detection using CNN

In the past few years, many techniques based on deep neural networks have been proposed to predict the positions of traffic lights. DeepTLR (Weber et al., 2016) and HDTLR (Weber et al., 2018) proposed by Weber *et al.* used convolutional neural networks for the detection and classification of traffic lights. Sermanet *et al.* presented an integrated framework, OverFeat, for classification, localization and detection (Sermanet et al., 2014). They have shown that the multi-scale and sliding window approach can be efficiently implemented in a convolutional network structure. Recently, general object detection networks are successfully adopted and specifically modified for traffic light detection. Behrendt *et al.* presented a deep learning approach to deal with the traffic lights using the YOLO framework (Behrendt et al., 2017; Redmon et al., 2016). Since the traffic lights might appear very small in some images, one common approach is to reduce the stride of the network architectures to preserve the features. Müller and Dietmayer adapted the single network SSD approach (Liu et al., 2016) and emphasized on the small traffic light detection (Müller and Dietmayer, 2018). Bach *et al.* presented a unified traffic light recognition system which is also capable of state classification (circle, straight, left, right) based on the Faster R-CNN structure (Bach et al., 2018; Ren et al., 2015).

For the traffic light recognition, recent approaches are divided into two categories. One is to detect the traffic light, crop the traffic light region, and send it to a classifier for the light state recognition (Behrendt et al., 2017). The other approach simultaneously detect the traffic light position and recognize the light state (Müller and Dietmayer, 2018; Bach et al., 2018). When the object location is predicted with a confidence and the bounding box, one more branch is used to predict the light state. For general traffic light recognition, except for the recognition of basic circular lights, it is also required to deal with various kinds of arrow lights in many countries. In the existing literature, this issue is only covered by a limited number of research (Weber et al., 2018; Bach et al., 2018). A two-stage approach is usually adopted with first the light color classification, followed by the arrow type classification.

2.3 Map Assisted Traffic Light Detection

One major drawback of image-based traffic light detection is the false positives caused by the similar features in the background. To reduce the incorrect detection, a simple method is to restrict the ROI in the image for traffic light search. Alternatively, the location of the traffic light on the map or from other input sources can also provide additional information for more accurate detection. This intends to improve, rather than replace the image-based methods. In this map-based traffic light detection approach, the idea is to utilize the fact that traffic lights are located at fixed locations in normal conditions. GPS or LiDAR are commonly used to establish the HD map and annotate the traffic light positions in the route (Fairfield and Urmson, 2011; Hirabayashi et al., 2019). When the vehicle is driving, the map and localization information is used to calculate the traffic light appeared in a small image region. Moreover, it can also provide the verification about which traffic light to follow if there exist more than one at a junction.



(a) Route 1: From ITRI to THSR Hsinchu station.

(b) Route 2: From CCU to THSR Chiayi station.

Figure 1: The routes for collecting data to create our own dataset. Route 1 is from ITRI to THSR Hsinchu Station, and Route 2 is from CCU to THSR Chiayi Station. The distances are 16 km and 39 km, respectively.

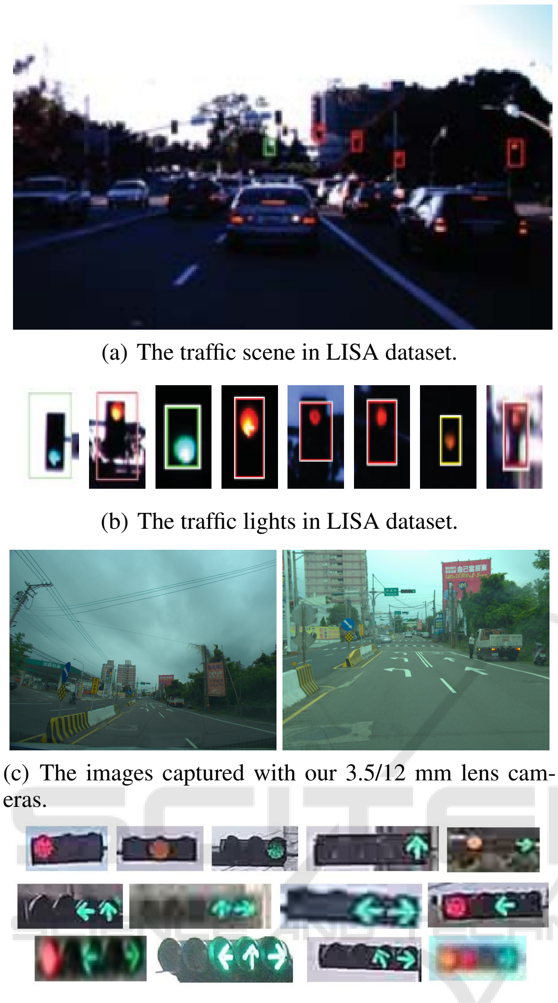
3 DATASET

Although several public datasets are available for traffic light detection and evaluation, they are not suitable for network training for Taiwan road scenes due to the different appearance. In this work, we cooperate with Industrial Technology Research Institute (ITRI) and collect our own dataset for both network training and performance evaluation.

Figure 1 shows the two commuter routes for data collection. One route is from the ITRI campus to Hsinchu High Speed Railway Station, and the other is from National Chung Cheng University to Chiayi High Speed Railway Station. These routes contain the driving distances of 16 km and 39 km, and the recording time of 40 minutes and 50 minutes, respectively. Two cameras with the focal length of 3.5 mm and 12 mm are mounted below the rear view mirror of a vehicle for image acquisition. The image sequences are captured at 36 fps with the resolution of 2048 × 1536. The LiDAR data are also recorded (by Velodyne Ultra Puck VLP-32C) and used for the segmentation of rough traffic light regions in the images.

The first route is recorded three times, and the second route is recorded once. We sample 5 images per second for processing, labeling the position of the traffic light and the class of the light state. The labeled data contain 26,868 image frames and 29,963 traffic lights. Only the traffic lights with clear light states are labeled, and there are totally 14 classes of light state combination in the dataset.

As shown in Figure 2, the traffic lights in LISA dataset (Jensen et al., 2016) are arranged vertically, which is very different from the ones arranged horizontally in Taiwan. Furthermore, the available light states are also different. Only a single light can be displayed at a time in LISA dataset (see Figure 2(b)), but there exist many combinations with various types of arrow lights in the Taiwan road scenes (see Figure 2(d)). Figure 3 shows our dataset collected using the cameras with 3.5 mm lens (in blue color) and 12 mm lens (in orange color) in terms of the cropped traffic light image size and the number of traffic lights in different classes. The dataset contains much more circular lights, but only a limited number of arrow lights. For the traffic light ROI size, LISA dataset mainly consists of the image regions in the range of 15 – 30 pixels. In our dataset, the images captured with 3.5 mm and 12 mm lenses have the traffic light regions in the ranges of 10 – 20 pixels and 15 – 50 pixels, respectively.



(a) The traffic scene in LISA dataset.

(b) The traffic lights in LISA dataset.

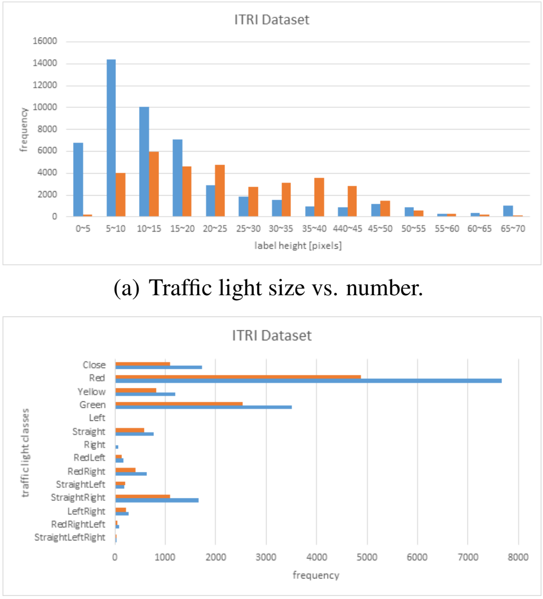
(c) The images captured with our 3.5/12 mm lens cameras.

(d) Examples of traffic lights collected in our dataset.

Figure 2: The traffic light images in the LISA dataset and our dataset captured using the cameras with 3.5mm and 12mm lenses.

4 APPROACH

This work integrates the map information for traffic light detection and recognition. We use the pre-established HD map with the traffic light annotation which contains ID, position, horizontal and vertical angles. The position between the vehicle and traffic lights can be determined by the LiDAR data and HD map during driving. This information is used to crop the image for a rough traffic light position. Due to the characteristics of the LiDAR data and the registration with images, it is not possible to identify the traffic lights accurately. The cropped ROI is then fed to the neural networks for precise location detection and light state recognition.



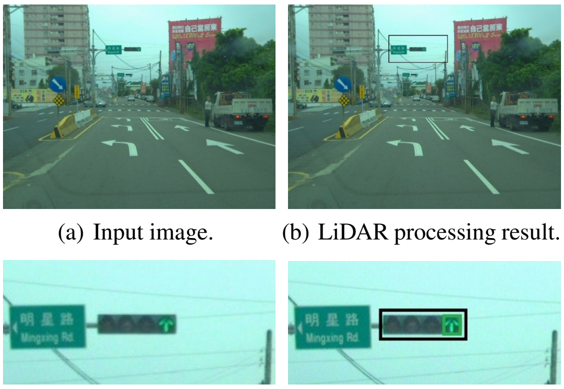
(a) Traffic light size vs. number.

(b) Traffic light class vs. number.

Figure 3: The statistics of our dataset in terms of the traffic light size and class. The blue and orange colors indicate the numbers from 3.5 mm and 12 mm lenses, respectively.

The proposed traffic light detection and recognition technique is a two-stage approach, with the first stage for the traffic light detection and the second stage for the light state recognition. In the first stage, several popular object detection networks including Faster R-CNN, SSD, YOLO have been tested. In the traffic light detection application, the computation load is one of the major concerns. YOLOv3 (Redmon and Farhadi, 2018) provides the best accuracy vs. processing time trade-off in the experiments, and is adopted for our detection framework. In the second stage, we propose a new method which combines the object detection and classification. The light states are detected by YOLOv3-tiny (Redmon and Farhadi, 2018) and classified to four classes: RedCircle, YellowCircle, GreenCircle and Arrow, followed by the classification of Arrows into LeftArrow, StraightArrow and RightArrow using LeNet (LeCun et al., 1998). As an example, if there is a Red-Left-Right light state, YOLOv3-tiny will detect one RedCircle and two Arrows, and LeNet will recognize the two Arrows as LeftArrow and RightArrow. The final traffic light state is then provided by combining the results of two networks.

This approach is expected to mitigate the unbalanced data problem. Furthermore, it is also flexible in that detecting less common light states can be performed with a slight modification to LeNet and prediction classes.



(a) Input image. (b) LiDAR processing result.

(c) The network input. (d) The network output.

Figure 4: The results of traffic light detection combining image and LiDAR data in the experiments.

4.1 Detection Network

4.1.1 Network Architecture

The proposed detection and classification technique consists of three cascaded network structures. It takes more time to train and inference if all three networks are independent. Thus, in the implementation, they are integrated to a single end-to-end network with shared feature maps. Better detection and classification results are obtained, and the speed of network training and inference is also improved. The unified network architecture is shown in Figure 5. Because the subnets share the same feature map, the architecture of the second and third subnets have changed. Feature extraction of the network is removed, which leaves only the prediction part. The subnet inputs are also changed from images to the feature maps coming from the FPN of the previous subnet.

4.1.2 Loss Function and ROI

The loss of the unified network is the error summation of the three subnets. It is expected that the network can back-propagate based on the overall task error. We use the original loss functions for YOLOv3 and YOLOv3-tiny, and the cross entropy loss for LeNet. The network training of the second and third subnets is not based on the detection results of the previous subnet, but directly from the groundtruth. This is due to the initial training of each subnet is not good enough to accurately identify the traffic lights or light states. Only when inferencing for evaluation, the network runs based on the detection results of the previous subnets.

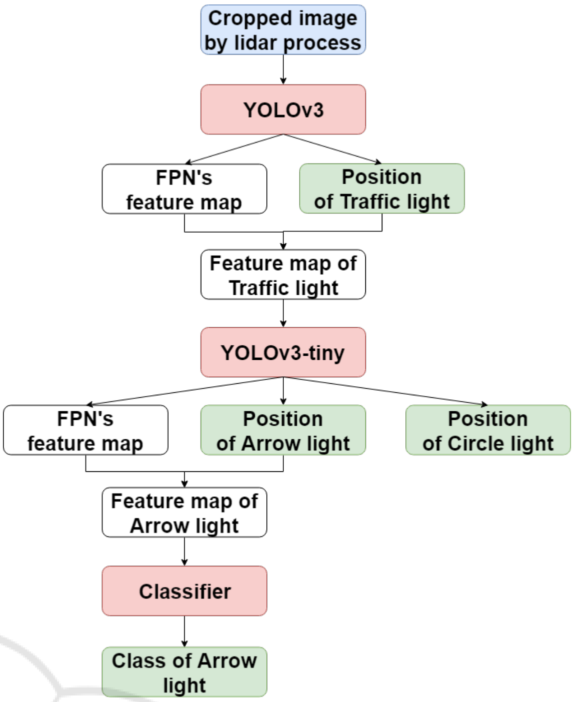


Figure 5: The flowchart of the proposed network architecture. The subnet inputs are the shared feature maps from the FPN (feature pyramid network).

4.1.3 Training Image and Data Augmentation

Our detection network is based on the cropped image, so the training data are also cropped to simulate the LiDAR processing, as shown in Figure 6. Each traffic light image is cropped 3 times, and the traffic light positions are randomly shown in the cropped image. The dataset mainly contains six classes: Red, Yellow, Green, Straight and StraightRight. Thus, data augmentation is carried out to generate more training data by rotating the Arrow light images (see Figures 6(d) and 6(e)).

5 EXPERIMENTS

5.1 Evaluation Criteria

We adopt three indicators for the evaluation of machine learning models: precision, recall, and F1-score.

- Precision: The ability to classify negative samples. If the precision is higher, the ability to classify negative samples is stronger, i.e.,

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Table 1: The LISA daytime dataset test result (mAP using IoU 0.5). The results of columns 1, 2, 3 are from (Jensen et al., 2016). The results of columns 4, 5, 6, 7 are from (Li et al., 2018).

Method	Stop	StopLeft	Go	GoLeft	Warning	WarningLeft	All
Color detector							0.04
Spot detector							0.0004
ACF detector							0.36
Faster R-CNN	0.14	0.01	0.19	0.001			0.09
SLD	0.08		0.10				0.09
ACF	0.63	0.13	0.40	0.37			0.38
Multi-detector	0.72	0.28	0.52	0.40			0.48
Ours	0.70	0.40	0.88	0.71	0.52	0.24	0.66

Table 2: The results obtained using our dataset. Three network structures are compared and one with data augmentation. The table shows the mAP and the computation time.

Network	YOLOv3 + AlexNet	YOLOv3 + YOLOv3-tiny+LeNet	Unified network	Unified network
Data augmentation	✗	✗	✗	✓
mAP	0.36	0.55	0.57	0.67
Speed (ms)	31	52	40	40

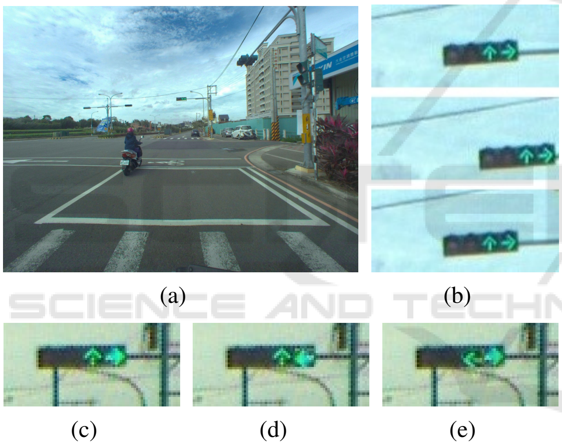

(a) (b)
(c) (d) (e)

Figure 6: Examples of the training images. (a) Original image, (b) cropped with LiDAR data, (c) training image, (d) data augmented 1, (e) data augmented 2.

- Recall: The ability to classify positive samples. If the recall is higher, the ability to classify positive samples is stronger, i.e.,

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

- F1-Score: A combination of precision and recall. If the F1-Score is higher, the classifier is more robust, i.e.,

$$F1-Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Two indicators are used to evaluate the object detection results: the precision-recall curve (PR curve) and the mean average precision (mAP).

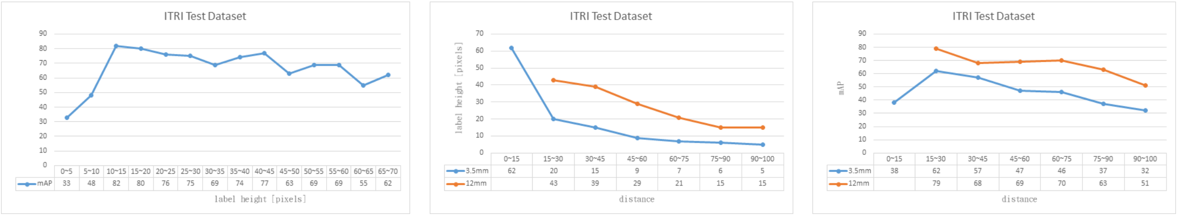
- PR curve: Different thresholds have different precision and recall. The precision-recall curve is given by drawing all the precision and recall. The trend of this curve represents the quality of a classifier.
- AP (average precision): It represents the area under the PR curve (AUC), which indicates the robustness of a classifier. If the area is larger, the classifier is stronger.
- mAP: AP only represents one class, but the models often detect many classes. So mAP is used to represent the average AP of all classes.

5.2 Training and Testing on LISA Dataset

For the performance comparison with other techniques, the LISA daytime dataset is used for training and testing. As shown in Table 1, the conventional detectors do not provide good results due to the complex scenes in the dataset. The low accuracy results obtained from Faster R-CNN are mainly due to the small traffic light regions, and it is difficult to detect after the layer-by-layer convolution (Li et al., 2018). The proposed method has achieved the best results as shown in the last row, which include the circular lights and arrow lights.

5.3 Training and Testing on Our Dataset

In Table 2, we compare the accuracy and computation speed of different network structures on our



(a) mAP vs. traffic light size

(b) Traffic light size vs. distance.

(c) mAP vs. distance.

Figure 7: The relationship among the mAP, the traffic light size in the image, and the distance of the traffic light in our dataset.

Table 3: The mAP of each stage of the network.

	Detection	State				Type		
Class	Traffic Light	Red	Yellow	Green	Arrow	Left	Left	Left
mAP	0.97	0.93	0.90	0.64	0.91	0.87	0.98	0.97

Table 4: The mAP for each class.

Class	Close	Red	Yellow	Green	Left	Straight	Right
mAP	0.43	0.78	0.79	0.76	No data	0.55	No data

Class	Red Left	Red Right	Straight Left	Straight Right	Left Right	Red Left Right	Straight Left Right
mAP	0.55	0.45	0.64	0.87	0.84	No data	0.69

dataset. The first one uses YOLOv3 to detect the traffic lights and AlexNet (Krizhevsky et al., 2012) to classify the light states. The second method is the 'YOLOv3+YOLOv3-tiny+LeNet' combination proposed in this work, but with three independent networks. The last one is the unified network structure containing the integration of the three subnets. All networks are trained and tested on the same dataset. As shown in the table, the mAPs of the proposed methods are better than 'YOLOv3+AlexNet' at the cost of more computation time. Comparing the first two network structures, the one with LeNet for arrow light classification has a better mAP but requires more computation time. For the proposed methods, the unified network (the last two columns in the table) has an improvement in mAP compared to the one without integration (the second column in the table). Furthermore, the computation speed is also faster due to the use of shared feature maps.

5.4 Comparison on Different Distance

The images taken from different distances contain the traffic lights with different ROI sizes. This will apparently affect the detection and recognition results. In the experiments, the cameras with 3.5 mm and 12 mm lenses are used for image acquisition. Figures 7(a), 7(b) and 7(c) show the mAP of different traffic light ROI size (height in pixel), the size of traffic lights at different distance, and the mAP of different distance, respectively. In Figures 7(a) and 7(b), the traffic lights taken by 12 mm lens are larger than taken by 3.5 mm lens. It is reasonable that larger traffic lights provide better detection results. Figure 7(c) shows the detection results for 12 mm lens are better than 3.5 mm for all distances. However, the short focal lens camera can still be used to cover the close range scenes. As shown in Figures 7(b) and 7(c), the traffic lights at the distance between 0 to 15 m can only be captured by the 3.5 mm lens due to the field of view of the cameras.

5.5 Detection Results in Each Stage

Our approach consists of three stages for traffic light detection, initial light state classification, and arrow type recognition. Table 3 shows the mAPs for each stage with light state and arrow type classification. The mAPs of the state and type are calculated based on the detection from the previous subnet. It shows that larger errors mainly appear in the second subnet, and the mAP of the Green class is the lowest. This is because the arrow light is similar to the green light for a far away distance. Table 4 shows the mAPs of all classes. The classes with sufficient training data have higher mAPs as expected. However, for the classes with insufficient samples, data augmentation is not able to improve the accuracy for all classes.

6 CONCLUSION

This paper presents a traffic light detection and recognition system based on convolutional neural networks for Taiwan road scenes. A two-stage approach is proposed with first detecting the traffic light position, followed by the light state recognition. It is specifically designed to handle the arrow signal lights. In the traffic light detection stage, the map information is used to facilitate the detection by restricting the ROI. Two cameras with different focal lengths are used to capture the near and far scenes. In the recognition stage, a method combining the object detection and classification is presented. It is used to cope with the problem of multiple light state classes in many urban traffic scenes. The proposed end-to-end unified network with shared feature maps has greatly reduced the training and inference computation. The experiments carried out using LISA dataset and our dataset have demonstrated the effectiveness of the proposed technique.

ACKNOWLEDGMENTS

The support of this work in part by the Ministry of Science and Technology of Taiwan under Grant MOST 106-2221-E-194-004 and the Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan is gratefully acknowledged.

REFERENCES

Abboud, K., Omar, H. A., and Zhuang, W. (2016). Interworking of dsrc and cellular network technologies for v2x communications: A survey. *IEEE Transactions on Vehicular Technology*, 65:9457–9470.

Bach, M., Stumper, D., and Dietmayer, K. C. J. (2018). Deep convolutional traffic light recognition for automated driving. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 851–858.

Behrendt, K., Novak, L., and Botros, R. (2017). A deep learning approach to traffic lights: Detection, tracking, and classification. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1370–1377.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2019). nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*.

Fairfield, N. and Urmson, C. (2011). Traffic light mapping and detection. *2011 IEEE International Conference on Robotics and Automation*, pages 5421–5426.

Fregin, A., Müller, J. M., and Dietmayer, K. C. J. (2017a). Feature detectors for traffic light recognition. *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 339–346.

Fregin, A., Müller, J. M., and Dietmayer, K. C. J. (2017b). Three ways of using stereo vision for traffic light recognition. *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 430–436.

Hirabayashi, M., Sujiwo, A., Monrroy, A., Kato, S., and Edahiro, M. (2019). Traffic light recognition using high-definition map features. *Robotics and Autonomous Systems*, 111:62–72.

Jensen, M. B., Philipsen, M. P., Møgelmose, A., Moeslund, T. B., and Trivedi, M. M. (2016). Vision for looking at traffic lights: Issues, survey, and perspectives. *IEEE Transactions on Intelligent Transportation Systems*, 17:1800–1815.

Kim, H.-K., Park, J. H., and Jung, H.-Y. (2011). Effective traffic lights recognition method for real time driving assistance systemin the daytime.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60:84–90.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, X., Ma, H., Wang, X., and Zhang, X. (2018). Traffic light recognition for complex scene with fusion detections. *IEEE Transactions on Intelligent Transportation Systems*, 19:199–208.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *ECCV*.

Müller, J. M. and Dietmayer, K. C. J. (2018). Detecting traffic lights by single shot detection. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 266–273.

Müller, J. M., Fregin, A., and Dietmayer, K. C. J. (2017). Multi-camera system for traffic light detection: About camera setup and mapping of detections. *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 165–172.

Ramanishka, V., Chen, Y.-T., Misu, T., and Saenko, K. (2018). Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Conference on Computer Vision and Pattern Recognition*.

Redmon, J., Divvala, S. K., Girshick, R. B., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *CoRR*, abs/1804.02767.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2014). Overfeat: Integrated recognition, localization and detection using convolutional networks. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Waymo (2019). Waymo open dataset: An autonomous driving dataset.

Weber, M., Huber, M., and Zöllner, J. M. (2018). Hdtlr: A cnn based hierarchical detector for traffic lights. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 255–260.

Weber, M., Wolf, P., and Zöllner, J. M. (2016). Deeptlr: A single deep convolutional network for detection and classification of traffic lights. *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 342–348.