Iraqi Journal of Information and Communications Technology(**IJICT**)
Vol. 4, Issue 3, December 2021

ISSN:2222-758X
e-ISSN: 2789-7362

REAL OBJECT DETECTION SYSTEM FOR IRAQ TRAFFIC SIGNS BASED ON MASK R-CNN

Ammar A. Aggar [1], **Mohammed J. Zaiter** [2], **Abdalrazak T. Raheem** [3]
[1,2,3] Electrical Engineering Technical College, Middle Technical University, Baghdad, Iraq
{bbc0029, mjzaiter}@mtu.edu.iq [1,2], abdtareq@gmail.com [3]
Received:18/2/2021, Accepted:22/4/2021

***Abstract*- Traffic signs object detection has gained high interest in recent years, as one of the most significant object detector applications. Where the development of deep learning technologies gives support to traffic signs detector which it offers several advantages, including the benefit of high detection precision and the timely response to condition changes of traffic signs. Therefore, this paper shows an efficient method for detecting traffic signs in real-time. Hence, it implements a new Iraqi Traffic Sign Detection Benchmark (IQTSDB) dataset based on Mask Region-based Convolutional Neural Network (Mask R-CNN) algorithm. The experimental results show that the implementation of the IQTSDB dataset with Mask R-CNN has high efficiency in different conditions such as sunny, cloudy, weak light, and rainy conditions. Besides, real images were captured for the traffic signs in Baghdad and then compared to the German traffic signs (GTSDB). In addition, the results show that the IQTSDB dataset has a better performance than the GTSDB dataset based on the performance parameters training loss and mean Average Precision (mAP).**

I. INTRODUCTION

Object detection has brought the attention of many researchers in both the academic and real-world applications fields. For example, security monitoring, drone scene analysis, robotic vision, autonomic driving, and transportation surveillance. Besides, autonomic driving and Traffic Signs Detection (TSD), are considered one of the most important of these applications, because it provides an accurate inventory of traffic signs and timely with minimal human effort. Also, Automobiles have become a vital part of our daily life, due to the fast technological development. So, this makes the road traffic further complicated, and that leads to more traffic accidents yearly. Hence, about 20-50 million are wounded or incapable, and about 1,300,000 million persons die (including 1,600 kids under 15 ages!) every year due to traffic accidents. According to the reports of the Association for Safe International Road Travel (ASIRT) [1] [2]. For this reason, many academics institutions and big companies paid more attention to autonomous vehicles and self-driving cars such as (Tesla, Google, Toyota, Ford, Audi, Uber, Mercedes-Benz, etc.). On the other hand, Deep learning technology has given traffic signs detection the benefit of high detection precision and the timely response to condition changes of traffic signs. However, in traditional machine learning, this task is executed manually. Where traffic signs have captured using a camera installed on the vehicle and recognition of them manually are performed offline by a human operator to check for concurrence with the existing database. Moreover, many issues face the drivers and make them have difficulty recognizing the traffic signs correctly, such as weather and light conditions. All these issues have made traffic signs detection is the focus of much research recently. Where, there are several deep learning algorithms used for traffic signs detection, such as Single Shot MultiBox Detector (SSD) [3], You Only Look Once (YOLO) [4], Faster Region-based Convolutional Neural Network (Faster R-CNN) [5], and Mask R-CNN [6]. Furthermore, many benchmarks have played an important role in traffic signs detection so far, such as PASCAL Visual Object Classes (PASCAL VOC) [7] [8], Microsoft Common Objects

ISSN:2222-758X
e-ISSN: 2789-7362

in Context (MS COCO) [9], and ImageNet [10] [11], where, models depend on pre-trained weights of these datasets in training.

II. Related Work

There are many standard Traffic Signs (TS) datasets used in this field, such as German traffic signs benchmark (GTSRB / GTSDB) [12], Sweden Traffic Sings (STS) [13], Belgium traffic signs (KUL) [14], the Netherlands traffic sings (RUG)[14], France traffic signs (Stereopolis), and the United States traffic signs (LISA)[14]. Moreover, our contribution to this paper proposed a new Iraqi Traffic Sings Detection Benchmark (IQTSDB), based on the Mask R-CNN algorithm. In this contribution, Mask RCNN uses to force various layers in a neural network to learn features with different scales, just like the ROIAlign and anchors, instead of treating layers as a black box. Then generates segmentation masks and bounding boxes for every instance of an object in the image. The research of traffic signs detection and recognition began as a European High-Performance traffic Project. This project has been financed by car companies such as (Mercedes Benz) to study traffic sign detection and recognition system [15]. Furthermore, popular methods of traffic sign detection involve a wide variety of algorithms and different representations. Hence, L. Wu et al. Proposed a German Traffic Signs Detection Benchmark (GTSDB) model based on deep learning to extract features automatically by using the Faster R-CNN algorithm [16]. While P. S. Zaki et al. Proposed a German Traffic Signs Detection Benchmark (GTSDB) model based on F-RCNN Inception v2 [1]. Also, C. G. Serna et al. Proposed a traffic signs recognition and detection system based on Mask R-CNN for European urban environments. This system includes symbol signs, and text signs, Categories not found in the original dataset. Moreover, GTSRB and GTSDB are used with the system [17].

III. Mask R-CNN Algorithm

Mask R-CNN is considered the expanded predecessor of Faster R-CNN, which is proposed by K. He et al. Mask R-CNN consists of a two-stage for detection. In the first stage, Mask R-CNN scans the input image and generates region proposals. That stage represents the Region Proposal Network (RPN). While the second stage classifies the region proposals and produces bounding boxes and a binary mask for every Region of Interest (ROI) [6]. Fig. 1 shows the illustration of Mask R-CNN.Mask R-CNN detects objects effectively in an image whereas produces for every instance a segmentation mask, at the same time. Moreover, Mask R-CNN is simple to train and appends only a tiny overhead to Faster R-CNN, operating at 5 fps [6]. Where, in this section, the main components of Mask R-CNN will be explained in detail.

A. Backbone

This stage is a convolutional neural network such as a residual neural network (ResNet 50, and ResNet 101) that works as a feature extractor. Hence, this network is dividing into two parts. The first part is the early layers, which detect low-level features (corners and edges). The second part represents the following layers that extract the higher-level features (traffic signs, human, car, animal). When the input image passed through the main network, it will be converted from (RGB) format $1024 * 1024 * 3$ to a feature map $32 * 32 * 2048$. In the next step, this feature map will become the entry of the following stage [6]. In addition, ResNet 50, ResNet 101 is more commonly used with Mask R-CNN. Where:

ISSN:2222-758X
e-ISSN: 2789-7362

- **ResNet 50:** is a residual neural network that is 50 layers deep.
- **ResNet 101:** is a residual neural network that is 101 layers deep.
- Feature Pyramid Network (FPN)

  FPN proposed by [19]. It is a fundamental component of object detection systems at different scales. Besides, it improves the basic feature extraction pyramid by appending a new step represented by another pyramid that brings high-level features from the main pyramid and transfers them to the lower layers. This step allows the features in each level to reach the features from the upper and lower levels. As shown in Fig. 2 Feature pyramid network.

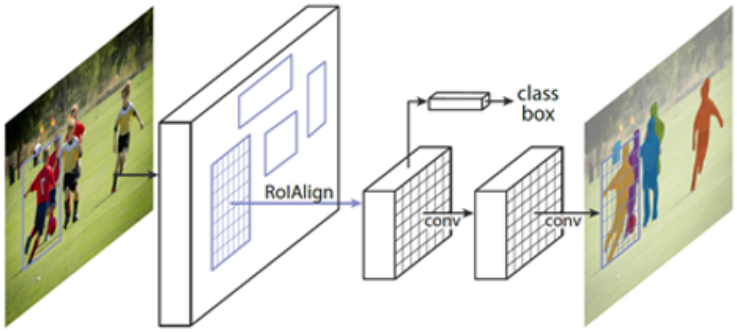

Figure 1: The illustration of mask R-CNN [6]

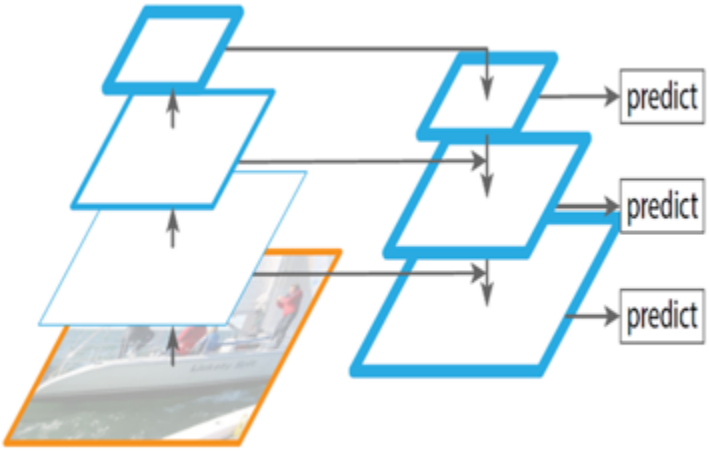

Figure 2: Illustration of feature pyramid network [19]

B. Proposal Network (RPN)

RPN aims to improve effectively the prediction of regional proposals in a huge range of sizes and aspect ratios. Furthermore, it is also called (sliding window proposers) [5] [19]. Besides, RPN increases the speed of generating proposal

ISSN:2222-758X
e-ISSN: 2789-7362

regions, it shares completely-image convolutional features and a Mutual group of (Conv) layers with a detection network [20]. The region that is scanned by RPN is called anchors [5] [19], where it is illustrated in Fig. 3.

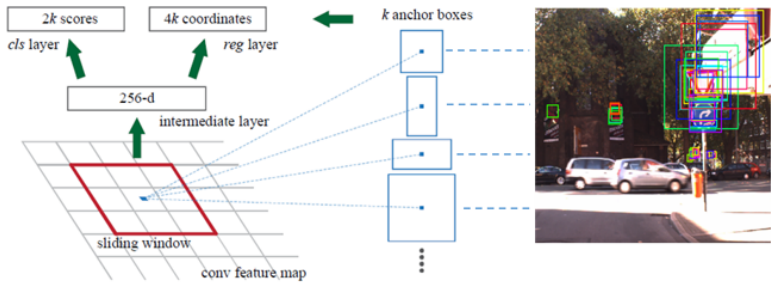

Figure 3: Illustration of region proposal network (RPN) [5]

1) Anchors

   RPN predicts multiple regional proposals simultaneously for each sliding window region as shown in Fig. 3. Moreover, the number of maximum potential proposals for every position is labeled as K. Therefore, the regression (reg) layer has 4k coordinates of k boxes. While the classification (cls) layer has 2k scores that measure the probability that each proposition has an object or not [5].

2) Region of interest Classifier and Bounding Box Regressor

   This step is based on the ROIs proposed by the RPN, it has the same function as RPN, which produces two outputs to every ROI [5] [6].

   - Class: ROI is a deeper network and can classify regions according to specific classes (people, vehicle, animal, etc.) as opposed to the RPN stage, which has two categories: foreground and background (FG/BG).
   - Bounding Box Improvement: It is the same step in RPN. Its goal is to improve the size and location of the bounding box to encapsulate the object.

C. RoIAlign

RoIAlign is one of Mask R-CNN features. It has proposed to fix the misalignment problem that was present in ROI pooling [6]. Consequently, to address this issue, RoIAlign follows two steps to bypasses the quantization of the ROI boxes. Where in the first step, RoIAlign calculates the floating number of the exact location of every ROI feature map track by a binary insertion process. Then, RoIAlign estimates the correct amounts of the features at four orderly Sampling sites in every ROI box. In the second step, RoIAlign gathers the results by using average or max-pooling to take the values of every box [20].

D. Segmentation Masks

The Mask branch is a tiny Fully Convolution Network (FCN). It is applied to every ROI. Hence, it predicts a pixel-to-pixel segmentation mask. Where it takes positive areas identified by the ROI classifier and then produces masks for them

ISSN:2222-758X
e-ISSN: 2789-7362

[6].

IV. DATASETS

This research uses the GTSDB dataset, and the experimental results show that the method has efficiency in deference conditions such as sunny, cloudy, weak light, and rainy conditions. As well as the new contribution of this research uses real captured images for traffic signs in Baghdad has been taken and compared to (GTSDB) dataset. These datasets have been trained by the Mask R-CNN algorithm. There is no standard dataset for the Iraqi traffic signs. For this reason, a new Iraqi Traffic Signs Detection (IQTSD) dataset has been proposed. It consists of 30 classes and 1400 high-resolution images, 1000 images for training, and 400 images for testing. Furthermore, it consists of about 1600 annotation objects. These images were gathered, at various times, in diverse environments, and different conditions. On the other hand, these images are collected by using a camera installed on a car. In addition, the proposed model was tested with another dataset (GTSDB). It consists of 43 classes and 900 high-resolution images, 600 images for training, and 300 images for testing. Moreover, it consists of 1200 annotation objects. Fig. 4 shows Cloudy condition (IQTSD), Sunny condition (IQTSD), Sunny condition (GTSDB), and Cloudy condition (GTSDB).

Figure 4: Traffic sings with variant conditions

V. LABEL IMAGES ANNOTATION

VGG Image Annotator (VIA) is a simple manual labelling and annotation tool for video, audio, and images. This tool is a standalone, lightweight, and offline software that works without any setup or installation and runs singly in a web browser. VIA tool allows human annotators to label and defines spatial regions in video or image frames and temporal segments in video or audio. VIA tool can have one of the following six shapes: polygon, ellipse, circle, rectangle, polyline, and point. For example, Polygon-shaped regions are used to labeling and annotate the edges of complex objects. Then can be exported this annotation in one of these formats, such as CSV and JSON [21]. Fig. 5 shows an example of this utility.

ISSN:2222-758X
e-ISSN: 2789-7362

Figure 5: User interface of VIA

VI. THE PROPOSED WORK

In this section, a proposed traffic signs detection system based on mask R-CNN for detect traffic signs of Iraq. This system includes symbol signs classes, and text signs not found in the original dataset. In the first step, our model extracts the traffic signs from the images by following all the steps in section III and then applied their corresponding annotation masks for each one can be seen in Fig. 6(a). Finally, this system generates class predicts, boundary box predicts, and mask predicts for each sign. Shows in Fig. 6 (b). This proposed system has been tested by used two datasets GTSD and IQTSD.

VII. RESULTS

In our research based on Mask R-CNN, ResNet-101 and FPN were used as a backbone network. However, Mask R-CNN is not desirable when based on the ResNet-101 network as a backbone network. Because it is slow and needs more time during training. On the other hand, ResNet 101 has high performance compared to the other networks, such as ResNet-50. The workflow for our model is illustrated in Fig. 7. In this research, the experimental results using Google Colab as a work environment and python language version of the TensorFlow GPU and Keras deep learning system. Moreover, using Windows 10 pro with Anaconda environment in Intel(R) Core (TM) i7-7500U CPU @ 2.70GHz and NVIDIA GeForce 940MX GPU. Also, the proposed system uses a learning rate of 0.001, and a weight decay of 0.0001 is used for Mask R-CNN. This approach also uses a momentum of 0.9.

A. Training Loss

The proposed model was trained for 100 epochs with pre-trained COCO wights for both datasets IQTSDB and GTSDB. Each epoch consists of 1000 steps. When trained for the head layers, at the final epochs of both datasets we got the final

ISSN:2222-758X
e-ISSN: 2789-7362

loss value for IQTSDB is 0.0566, while for GTSDB is 0.0716. These results show that the IQTSDB dataset has less loss than the GTSDB dataset on the final epoch of the training as shown in Fig. 8.

Figure 6: (a) Some samples of traffic signs in the GTSD and IQTSD traffic-sign dataset with their corresponding annotation masks showing the accuracy of the annotations (b) Final step of the proposed system generates class predicts, boundary box predicts, and mask predicts

B. Mean Average Precision (mAP)

mAP [22] is used as the norm for object detection in this search. Where the greater value of mAP leads to a larger value of object detection accuracy of the framework. Also, mAP can calculate as the following equation:

$$mAP = \frac{1}{n}\sum_{i}^{n} AP_i \tag{1}$$

Where i is the label of a class and n is the number of object classes. Besides,AP_i is the average precision of i and also it represents the region under the Precision-Recall Curve (PRC). Furthermore, Precision is used to measure the ratio of all instances up that rank which are of the positive class. While Recall is used to measure the ratio of all positive instances [20]. Fig. 9 shows that generally the Mask R-CNN more efficient than Faster R-CNN inception v2 and the traditional

ISSN:2222-758X
e-ISSN: 2789-7362

Faster R-CNN based on the mAP parameter. Also, the comparison of the two values of Mask R-CNN reveals that the IQTSDB dataset has better performance than the GTSDB dataset. Based on these results of mAP, it can be concluded that with large databases the IQTSDB dataset gives more accuracy than the GTSDB dataset. Where, IQTSDB dataset has more images take in different conditions such as sunny, cloudy, weak light, and rainy conditions when compared to GTSDB dataset. While Fig. 10 (a) and (b) show the implementation of the proposed model based on Mask R-CNN for (GTSDB) and (IQTSDB) datasets. As well, in this paper to evaluate the value of mAP, the threshold value of IoU (Intersection over Union) set at 0.5 (or 50%). Finally, Fig. 11(a) and (b) show the best values of mAP at IoU threshold = 0.5 (or 50%) for IQTSDB and GTSDB datasets using the proposed model based on Mask R-CNN.

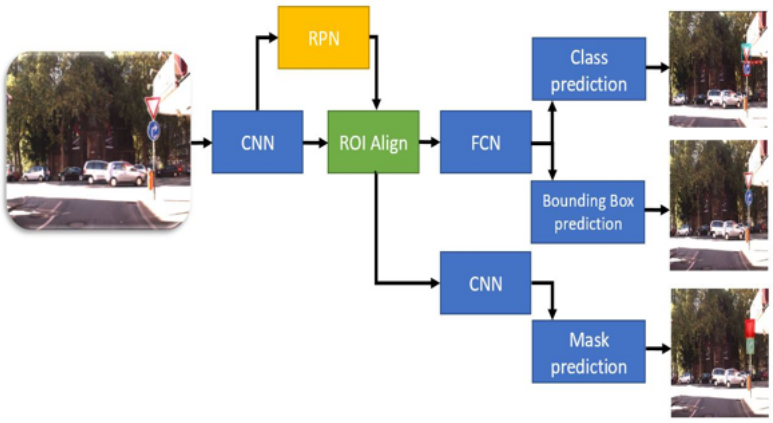

Figure 7: Mask R-CNN workflow

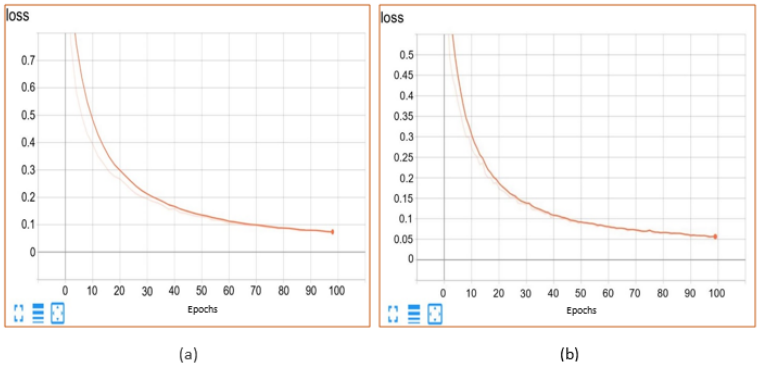

Figure 8: Final training loss value for (a) GTSDB (b) IQTSDB

ISSN:2222-758X
e-ISSN: 2789-7362

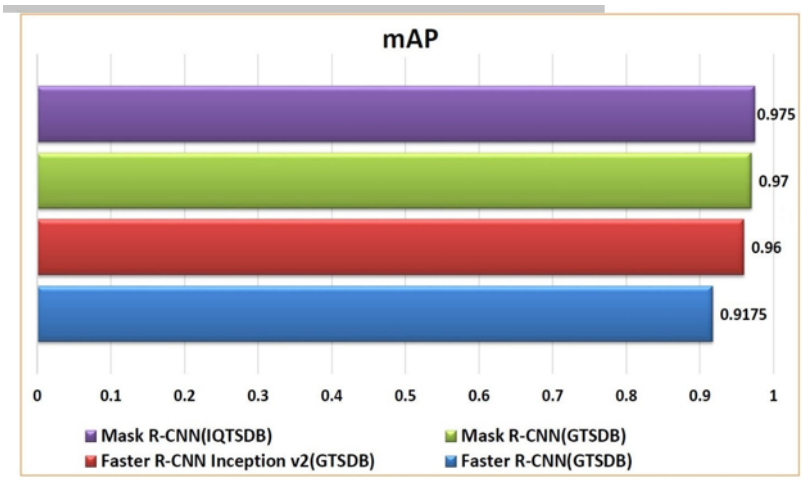

Figure 9: A comparison between mask R-CNN and (faster R-CNN Inception v2, faster R-CNN

(a)

(b)

Figure 10: The implementation of the proposed model based on mask R-CNN (a) GTSDB and (b) IQTSDB

ISSN:2222-758X
e-ISSN: 2789-7362

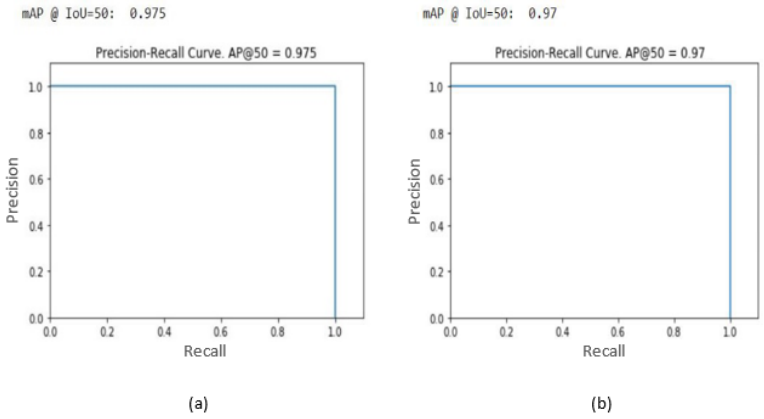

Figure 11: mAP value using the proposed model based on mask R-CNN (a) IQTSDB and (b) GTSDB

VIII. CONCLUSION

This research has applied traffic signs detection based on the Mask R-CNN algorithm with Benchmark GTSDB and IQTSDB. Iraqi dataset of traffic signs was gathered in Baghdad city. Where the real images of the traffic signs have been taken in different conditions such as (sunny, cloudy, and raining conditions). The results of the implementation of the proposed module with the Mask R-CNN algorithm show that the IQTSDB has better performance than GTSDB. The comparison is based on the performance parameters which are training loss and mean Average Precision (mAP). On the other hand, the most important contribution of this research is real images are captured for traffic signs in Baghdad has been taken and compared with the GTSDB dataset. Finally, the result shows that the mAp is 97.5% with IQTSDB while, the mAP is 97% with GTSDB.

ISSN:2222-758X
e-ISSN: 2789-7362

REFERENCES

[1] P. S. Zaki, M. M. William, B. K. Soliman, K. G. Alexsan, K. Khalil, and M. El-Moursy, "Traffic Signs Detection and Recognition System using Deep Learning" , arXiv Prepr. arXiv2003.03256, 2020.
[2] E. Nasr, E. Kfoury, and D. Khoury, "An IoT Approach to Vehicle Accident Detection, Reporting, and Navigation" , in 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET) , pp. 231-236, 2016.
[3] W. Liu et al. , "Ssd: Single Shot Multibox Detector" , in European conference on computer vision, pp. 21-37, 2016.
[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection" , in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788, 2016.
[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-Cnn: Towards Real-Time Object Detection with Region Proposal Networks" , in Advances in neural information processing systems, pp. 91-99, 2015.
[6] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-Cnn" , in Proceedings of the IEEE international conference on computer vision, pp. 2961-2969, 2017.
[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge" , Int. J. Comput. Vis. , Vol. 88, No. 2, pp. 303-338, 2010.
[8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective" , Int. J. Comput. Vis. , Vol. 111, No. 1, pp. 98-136, 2015.
[9] T. Y. Lin et al. , "Microsoft COCO: Common Objects in Context " , arXiv Prepr. arXiv1405.0312, 2019.
[10] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A Large-Scale Hierarchical Image Database" , in 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255, 2009.
[11] O. Russakovsky et al. , "Imagenet Large Scale Visual Recognition Challenge" m Int. J. Comput. Vis. , Vol. 115, No. 3, pp. 211-252, 2015.
[12] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark" , in The 2013 international joint conference on neural networks (IJCNN) , pp. 1-8, 2013.
[13] F. Larsson and M. Felsberg, "Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition" , in Scandinavian conference on image analysis, pp. 238-249, 2011.
[14] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-Based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey" , IEEE Trans. Intell. Transp. Syst. , Vol. 13, No. 4, pp. 1484-1497, 2012.
[15] C. Wang, "Research and Application of Traffic Sign Detection and Recognition Based on Deep Learning" , in 2018 International Conference on Robots & Intelligent System (ICRIS) , pp. 150-152, 2018.
[16] L. Wu, H. Li, J. He, and X. Chen, "Traffic Sign Detection Method Based on Faster R-CNN" , in Journal of Physics: Conference Series, Vol. 1176, No. 3, p. 32045, 2019.
[17] C. G. Serna and Y. Ruichek, "Traffic Signs Detection and Classification for European Urban Environments" , IEEE Trans. Intell. Transp. Syst. , Vol. 21, No. 10, pp. 4388-4399, 2019.
[18] P. Singh, R. Manikandan, N. Matiyali, and V. P. Namboodiri, "Multi-Layer Pruning Framework for Compressing Single Shot Multibox Detector" , IEEE Winter Conference on Applications of Computer Vision (WACV) , pp. 1318-1327, 2019.
[19] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection" , in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125, 2017.
[20] L. Jiao et al. , "A survey of Deep Learning-Based Object Detection" , IEEE Access, Vol. 7, pp. 128837-128868, 2019.
[21] A. Dutta and A. Zisserman, "The VIA Annotation Software for Images, Audio and Video" , in Proceedings of the 27th ACM International Conference on Multimedia, pp. 2276-2279, 2019.
[22] P. Henderson and V. Ferrari, "End-to-End Training of Object Class Detectors for Mean Average Precision" , in Asian Conference on Computer Vision, pp. 198-213, 2016.