Two-stream Network for Moving Object Detection

Dhammatorn Wisan, Naoshi Kaneko, Seiya Ito, and Kazuhiko Sumi

Aoyama Gakuin University, Japan

Abstract. Object detection is one of the fundamental challenges in computer vision. In the real world, however, problems such as occlusion by other objects, background or foreground variations, and low contrast of the target object arise. Attempting to detect the targets with low similarity to the object model increases the probability of false detection. If the target appears in front of the background, the detection threshold of the target can be lowered in the foreground region while keeping the probability of false detection low. Using this idea, we propose a new method that, in addition to frame-by-frame object detection, can detect missed target objects by performing object classification on newly appearing objects in regions where no objects are detected. In our proposed network, the first stream is standard object detection and the second stream performs background subtraction for non-object regions detected by the first stream. Then the second stream performs object classification for the detected foreground regions. Finally, those two streams are merged and object class and regions are output. We applied this method to the detection of vehicles on the road and were able to detect even low-contrast vehicles that could not be detected by frame-by-frame detection.

Keywords: Object detection · Image segmentation · Vehicle detection · Background subtraction · deep neural network.

1 Introduction

Object detection is a fundamental problem in computer vision tasks. In real environments, even state-of-the-art object detection methods suffer from a variety of conditions such as low contrast, occlusion, similar backgrounds, and uneven illumination. Especially in vehicle detection on highways, physical and economic reasons limit the conditions under which cameras can be installed. As a result, the following difficulties arise. Vehicles are hidden by other vehicles, vehicles have low contrast against the background, visibility is low, and the size of the vehicle in the image varies with distance from the camera. Until the mid-2000s, background subtraction was the primary method used for vehicle detection. However, background subtraction does not work well when the background changes significantly. Subsequently, object modeling such as HOG [2] and SIFT [9] features, which are robust to background changes, have been preferred

over background subtraction methods. In particular, object modeling using neural networks has become popular in recent years. However, state-of-the-art object detection methods are unable to detect low-contrast objects, which are often found in traffic scenes. Therefore, we propose to improve the detection performance of low-contrast objects by combining a state-of-the-art object detection network with a background subtraction method. In our proposal, objects are detected in two streams. The first stream performs frame-by-frame object detection. The output of the first stream includes an object mask. In the second stream, background subtraction is performed. The foreground region is then masked by the object mask of the first stream. The image then contains regions where some moving object that was not detected in the first stream appears. These regions are segmented and object classification is applied to each region. This classification lowers the detection threshold while keeping the false positive rate low.

In summary, our contributions in this work are:

1. A method to separate areas for object detection and classification later.
   We modified an existing method [5] for multi-object semantic segmentation that is robust to low contrast image. The segmented area will be used later to compare undetected foreground from next Step.
2. A method for object detection and segmentation.
   In this work, we employ state-of-the-art object detection to generate reliable object bounding boxes. We modified some methods of YOLOv7 [10] and combined them with the object classification method to improve model efficiency under low contrast.
3. A method for object classification.
   We apply a pre-trained a model for object classification using GluonCV trained with the CIFAR-10 dataset. GluonCV [14] is a toolkit which provides various implementations of the state-of-the-art based on using deep learning models in computer vision.

2 Related Work

Before performing object detection and segmentation, one of the most challenging tasks is to segment foreground and background from video sequences. Although many methods have now been proposed that do not require background subtraction methods in object detection and tracking, this pre-processing step is still important under heavy occlusion cases. Currently, applying deep convolutional neural networks (CNNs) shows apparent achievements in the field of computer vision. Its outstanding performance provides an efficient ability to extract object features and object masks that can be used in multiple related tasks including our task in vehicle detection. We describe some details of each related component in the following sections.

2.1 Object Detection and segmentation

Object detection is one of the classic and interesting challenges in the field of computer vision. It is divided into two methods, traditional neural network methods, and recent complex deep learning methods. Generally, object detection methods can be divided into two categories, One-stage detection and Two-stage detection. You Only Look Once (YOLO) [7] and Single Shot Multibox Detector(SSD) [12] are well-known in one-stage methods. Faster R-CNN [8] and Mask R-CNN [3] are frequently referred to two-stage methods. However, when the some object parts are occluded by other obstructions, these methods are not able to detect some of those objects. A number of approaches have recently been proposed to address occlusion problem. Yuan et al. [13] presented a method for multi-object instance segmentation which can locate occluders and classify objects based on non-occluded parts in order to detect an object that is occluded by other objects.

2.2 Background Subtraction

So far, background subtraction is a significant processing part of object detection and object tracking performed with video sequences. In the past, the most widely known method was a classical background subtraction method that used a previous frame or a static image to be a background for performing subtraction. This technique can be used to a certain degree, but it is very sensitive to various conditions, such as dynamic changes, noise, illumination, and so on. Due to those issues, many approaches have been proposed to accurately segment between foreground object, and background. At present, the use of deep learning shows amazing success and have been approved its efficiency by many researchers. Long and Hacer [4] proposed a method that creates a mask to separate the object from the background from an input image trained from a video sequence. The advantage of this method is that it doesn't need to use a lot of data for training, but it provides good results. However, this method still has a limitation: the input image that will be tested for segmentation must be an image with the same background as the data used.

3 Methodology

In this section, we introduce our main structure for object detection using vehicle model under various conditions in traffic videos. An algorithm is explained by following:

Step 1: We first obtained the image input by an RGB camera and defined it as I^{xy} when the (x, y) is coordinate.

Step 2: Masks of sub ROI [5] areas are extracted from background subtraction approach based on deep learning technique in order to classify undetected objects later. Foreground F can be obtained after performing an

input image I processed by GluonCV:

$$F(x,y) = \begin{cases} 1, & \text{if} \quad I(x,y) \text{ is foreground} \\ 0, & \text{if} \quad I(x,y) \text{ is background} \end{cases} \tag{1}$$

Step 3: The YOLOv7 [10] with pre-trained weight is used to detect the vehicles of input image.

Step 4: Semantic segmentation of each vehicle area is applied to create masks for comparing with masks extracted from previous background subtraction in Step 2. Foreground G can be obtained after performing an input image I processed by semantic segmentation:

$$G(x,y) = \begin{cases} 1, & \text{if} \quad I(x,y) \text{ is segmented area} \\ 0, & \text{if} \quad I(x,y) \text{ is unsegmented area} \end{cases} \tag{2}$$

Step 5: We compare different regions, between the mask from Step 2 and the mask from Step 4 then the objects in different regions are classified as to whether they become a vehicle category. The salient is undetected area and need to be classified the category which can be explained by:

$$S(x,y) = F(x,y) - G(x,y) \tag{3}$$

where S is area for classification.

Step 6: S is combined with RGB image input. Then, to classify whether the object in the area S is a vehicle, we used a pre-trained model named CIFAR_ResNet110_v1 of image classification network from the GluonCV [14] framework.

To summarize all methods, a flowchart of the proposed framework is shown in Fig. 1.

4 Experiment and Discussion

In this section, we evaluate the effectiveness of our proposed method by using two different datasets, CDnet2014 [11] and a custom dataset constructed from the CARLA simulator[1]. Each of the datasets contains challenging conditions such as dynamic background movement, shadow, noise, and so on to evaluate our method. An example of output result is shown in Fig. 2.

4.1 Vehicle Dataset

Video sequences from the CDnet2014 dataset [11] are used for training with spatial resolution 320×240 that contain various scenarios such as dynamic background motion, illumination changes, multiple objects, contrast, etc. Moreover,

[1] https://carla.org

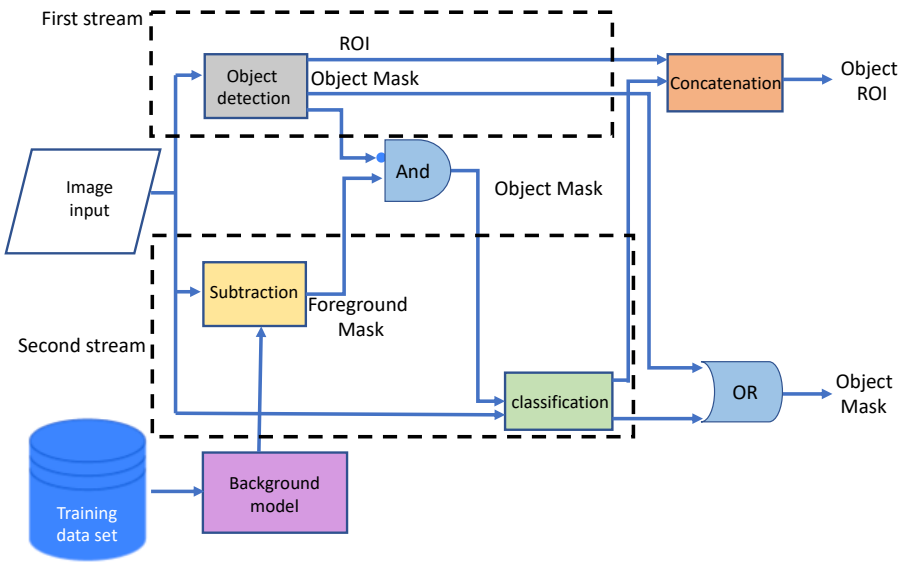

Fig. 1. Overall flow of the proposed method.

in order to increase the flexibility of the datasets and various aspect adjustments not available in the downloaded datasets, the CARLA simulator is applied to construct datasets under various challenging conditions. The CARLA is an open-source simulator widely used for urban driving research.

4.2 Evaluation Measure

To ensure that our proposed framework can be certainly used in various conditions, we have trained a model using different image sizes of each dataset. However, for a fair comparison, test images used for assessment of the performance of the system are performed with the same object detection and classification methods.

We used F-Measure, precision, and recall value to assess the performance of our model. The formulas of accuracy, precision, and recall we use to evaluate and describe the quality of our model are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

where TP, FP, and FN mean true positive, false positive, and false negative, respectively. In our experiment, we randomly select images for testing from CDnet2014 20 images and randomly select from our custom datasets 10 images. The result of experiment is shown in Table 1.

Table 1. Accuracy of the network model.

Dataset	Precision	Recall	F-Measure
CDnet2014	0.9917	0.9523	0.9716
Our test set	1.0000	0.7586	0.8627

5 Conclusion and Future Work

In this paper, we presented an efficient method for object detection using vehicle model under several conditions from the perspective of surveillance cameras for various surveillance video scenes with the pipeline shown in Fig. 1. First, image or video sequence is used as input of system. Next, vehicle detection is performed by YOLOv7 with pre-trained weight. To address the problem of the small or occluded vehicles, which cannot be detected. A more effective sub-area was obtained by extracting mask of foreground segmentation of each vehicle shape from road surface areas. Then, each object in sub-ROI areas of each frame detected to obtain better vehicle detection results from object classification method. Finally, bounding boxes are constructed in case object's category is classified as vehicle. If not, nothing performed. In addition, we also conducted some pre-processing and post-processing of traditional methods to enhance the performance and results of our framework further. Our proposed method clearly shows better results in vehicle detection compared with traditional methods. For future work, key-point detection and tracking from the optical flow method can be aggregated in the pipeline for improving efficiency of framework.

Acknowledgements

Part of this research was operated as the project of Center for Advanced Information technology Research (CAIR), Aoyama Gakuin University.

References

1. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3464–3468 (2016). https://doi.org/10.1109/ICIP.2016.7533003

2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) **1**, 886–893 (2005)
3. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. arXiv (2017). https://doi.org/10.48550/ARXIV.1703.06870
4. Long, A. L., Hacer, Y. K.: Foreground segmentation using convolutional neural networks for multiscale feature encoding, vol. 112, pp. 256–262. Elsevier (2018). https://doi.org/10.1016/j.patrec.2018.08.002
5. Long, A. L., Hacer Y. K.: Learning multi-scale features for foreground segmentation. Pattern Analysis and Applications, vol. 23, pp. 1369–1380. Springer Science and Business Media (2019). https://doi.org/10.1007/s10044-019-00845-9
6. Nicolai, W., Alex B., Dietrich P.: Simple Online and Realtime Tracking with a Deep Association Metric. arXiv (2017) https://doi.org/10.48550/ARXIV.1703.07402
7. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
8. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv (2015). https://doi.org/10.48550/ARXIV.1506.01497
9. Lindeberg, T.: Scale Invariant Feature Transform. Scholarpedia **7**, 10491 (2012)
10. Wang, C., Bochkovskiy, A., Liao, H.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv (2022) https://doi.org/10.48550/ARXIV.2207.02696
11. Wang, Y., Jodoin, P., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P.: CDnet 2014: An Expanded Change Detection Benchmark Dataset. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 393–400 (2014). https://doi.org/10.1109/CVPRW.2014.126
12. Wei, L., Dragomir, A., Dumitru, E., Christian, S., Scott, R., Cheng, Y., Alexander, C. B.: SSD: Single Shot MultiBox Detector: Computer Vision – ECCV 2016, pp. 21–37. Springer International Publishing (2016)
13. Xiaoding, Y., Adam, K., Yihong, S., Alan, Y.: Robust Instance Segmentation through Reasoning about Multi-Object Occlusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11136–11145 (2021)
14. GluonCV Homepage, http://cv.gluon.ai/. Last accessed 16 Jan 2023

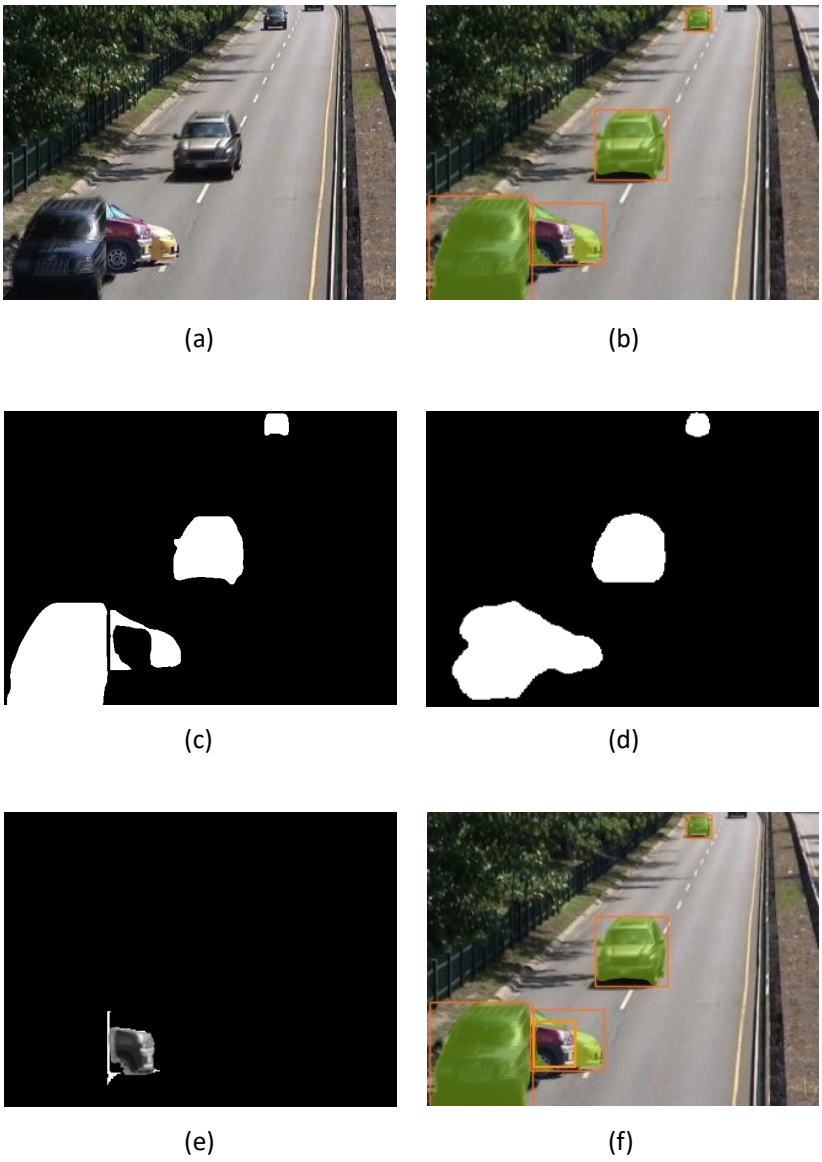

Fig. 2. Single-frame video object detection result. (a) RGB image input; (b) object detection and segmentation from YOLOv7; (c) mask segmentation from YOLOv7; (d) mask segmentation from [4]; (e) different region between (c) and (d); (f) result of full image detection method.