МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

Национальный исследовательский Томский государственный университет
Томский государственный университет систем управления
и радиоэлектроники
Болгарская Академия наук
Академия инженерных наук им. А.М. Прохорова
Международная научно-техническая организация «Лазерная ассоциация»
Всероссийское общество изобретателей и рационализаторов

ИННОВАТИКА-2021

СБОРНИК МАТЕРИАЛОВ

**XVII Международной школы-конференции студентов,
аспирантов и молодых ученых
22–23 апреля 2021 г.
г. Томск, Россия**

Под редакцией А.Н. Солдатова, С.Л. Минькова

Scientific & Technical Translations
STT
ИЗДАТЕЛЬСТВО
Томск – 2021

AUTONOMOUS DRIVING OBJECT DETECTION USING ACF

M.J. Mohammed, S.V. Shidlovskiy
National Research Tomsk State University
maryamjasim80@gmail.com

In this work, we obtain a vehicle semantic understanding using their image features and a rule-based system. These features provide the vehicle spatial and temporal information. Vehicle spatial feature is obtained using an ACF network. The vehicle temporal information is obtained using a novel semantic segmentation framework. The statuses of the neighboring vehicles are categorized as «Car-follow» and «Car-avoid» also. We validate our proposed frame-work with multiple acquired sequences. Our experimental results show that the proposed framework can estimate the status of the different vehicles in the urban road environment in near real-time.

Keywords: autonomous vehicle, aggregate channel feature (ACF), deep learning, object detection, classification, semantic segmentation.

Introduction

Automated driving research has gained prominence in the industry as well as in academia in recent years. In self-driving, awareness of the environment, assessment of the situation, and decision-making play an important role. Intelligent processing of information from vehicle sensors leads to environmental awareness. After being aware of the environment, the semantic understanding of the vehicle is used to make effective decisions [1]. In a complex driving scene such as an urban area, semantic understanding of neighboring vehicles plays an important role in achieving fully automatic driving. As an autonomous vehicle must not only detect and classify surrounding vehicles, it must also assess their condition. In this paper, a vision-based semantic comprehension framework for automated driving in multi-lane urban roads using deep learning and rules based system is proposed. Within this framework, we first assess the spatial and temporal information of all vehicles in a video clip using deep learning-based environment awareness. To estimate the spatial information of the vehicles, accurate ACF was used to discover, locate and classify all vehicles in a given image according to their spatial location with respect to the autonomous vehicle. The temporal information or binary movement status of all vehicles on the road is estimated from a series of images using a new multi-frame semantic segmentation framework, where the movement status of all vehicles is estimated across multiple tires without the need for tracking. Then the estimated vehicle spatial and kinematic information is used for the semantic understanding of nearby vehicles in a multi-lane urban road.

The main task of this work is the ability to detect cars, pedestrians and cyclists, in a single-camera image captured from a front-end camera of the

traffic scene from the front. The detection was to be achieved by predicting the bounding box around the target object in addition to estimating the range also of the object. The following sections explain the dataset and the structure of the model Implementation details related to the disclosure mission. To perform all the afore mentioned tasks in real-time, it requires high performing hardware. In this paper image segmentation is employed to track without any errors using MATLAB automated tool box.

Aggregate channel features (ACF)

ACF is an object detection method in computer version. A group of channels would be export from an input image. Channels are defined by a block of pixels values. Features are obtained by pixels in a channel according to an interested rectangular region. Boost trees help to decide to distinguish objects. ACF has proven to be one of the fastest existing detectors with source code available online [1]. ACF is the descendant of the ICF detector [2] with the main difference that instead of using Haar like features, it divides channels into blocks, where the pixels in each block are summed/aggregated. Every value in these aggregated channels represents a feature, more explicitly, features are single pixel lookups. Furthermore, ACF uses 10 channels: normalized gradient magnitude (1 channel), histogram of oriented gradients (6 channels), and LUV color (3 channels). With a detection window size of and the aggregation, ACF creates proposals of the size (1280 features). For the detection, 2048 trees with depth 2 are applied to every detection window in a constant soft cascade manner. In order to make our results easily reproducible, we use the pre-trained model.

Implementation details

The motivation for the approach to detection originated from the implementations in [3]. Since the feature extractor down-sampled the image by a factor of 32, an up sampling operation through tiling was introduced such that the effective stride of the end to end classifier is 4. This means that the classifier was capable of producing predictions of resolution 4×4 on the input image. The pre-trained ResNet model was trained on the ImageNet data set. Hence, the images to be used for training or testing from the KITTI data set, had to be normalized using the same normalization parameters of the pre-trained network. For ResNet-34, the mean to be subtracted from the input RGB channels is {0.485, 0.456, 0.406} and the standard deviation to be adjusted for {0.229, 0.224, 0.225}.

Training

The approach to training consisted of constructing masks corresponding to the resolution of dense predictions that the network was designed to make, and train in an end-to-end manner. Crops of size 224 × 224 were extracted from the large KITTI images and masks of size 56 × 56 were created corresponding to effective stride of the network. The masks for training were constructed by considering the object bounding box labels, and shrinking them to 20% of their original height and width. Some examples of the crops and corresponding masks are presented in Figure 1.

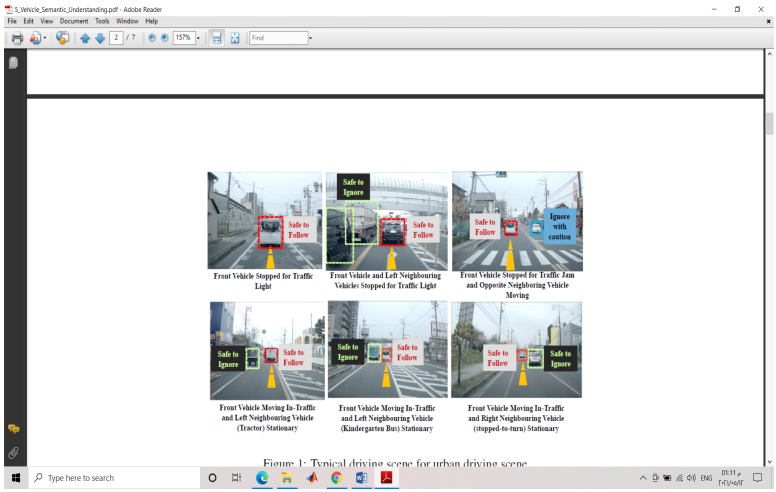

Fig. 1. Typical driving scene

When there was an object occluded by another object, then the mask label was constructed to contain the object class of the object closer to the ego vehicle. The crops were constructed such that 70% of them belonged to centered objects and 30% were random patches extracted from the image, which could in turn contain objects or be pure background. The classes of the object-centered crops were drawn such that the ratios of «Car», «Pedestrian» and «Cyclist» were roughly equal. Furthermore, the object-centered crops were created to have slight offsets from the center of the crop by adding or subtracting a random number of pixels drawn uniformly in the range -20 to 20 pixels, to the coordinates of the crop to be extracted. This is a data augmentation technique adopted to improve the robustness of the network. The KITTI data provides the occlusion status of objects, and the object-centered crops were always ensured to be non-occluded. Another aspect of the KITTI data is that it contains a «Grey-zone» class which consists of far-

away objects, and such regions in the test set are ignored while evaluating. Hence, by assigning those regions in the training set with a «Grey-zone» class, hard negative training on those ambiguous regions was avoided.

Result

Figure 2 presents a sample of images from the test set with the classification score of the «Car» class projected on top and the resulting bounding boxes after suppression and merging using non-maximum suppression. The figure shows the network is successfully able to detect objects of varying scale, lighting conditions and orientation. The network also manages to detect partially occluded objects, resulting in overlapping bounding boxes.

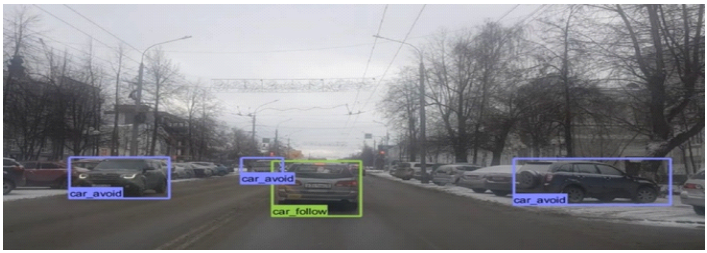

Fig. 2. Illustrate of how bounding box coordinates in the bounding box mask

Conclusion

The algorithm applied to these four tests does not show a great result. It has two main limitations. The one of important limitation is the windy and snow shower weather in testing day. Image qualities are not satisfied as the previous test. The detector is only able to detect objects which close to the testing vehicle. Increasing threshold distance and sensitivity is unable to increase detection abilities. In addition, some frame was missing detection due to low image qualities. Therefore, mass manual work needed by ground truth labeler app.

References

1. Ciberlin J., Grbic R., Teslić N. et al. Object detection and object tracking in front of the vehicle using front view camera // 2019 Zooming Innovation in Consumer Technologies Conference (ZINC). 2019. P. 27–32.
2. Lin-Bo Luo, In-Sung Koh, Kyeong-Yuk Min at al. Low-cost implementation of bird's-eye view system for camera-on-vehicle // 2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE). 2010. P. 311–312.
3. Object Detection Evaluation [Electronic Resource]. – URL http://www.cvlibs.net/datasets/kitti/eval_3dobject.php (date: 15.03.2021).