Vehicle Semantic Understanding for Automated Driving in Multiple Lane Urban Roads using Deep Vision-based Features

Vijay John , Seiichi Mita
Toyota Technological Institute, Japan
{vijayjohn, smita}@toyota-ti.ac.jp

Abstract

Vehicle semantic understanding is a key task in fully automated driving, where an assessment of which vehicles to follow, and which vehicles to ignore is made. In this paper, we obtain a semantic understanding of the vehicle status using their image-based features and a rule-based system. The image-based features represent the vehicle spatial and temporal information. vehicle spatial feature is obtained using a fine-tuned YOLO-3 network. The vehicle temporal information is obtained using a novel semantic segmentation framework. Using these preliminary perception information, a semantic understanding of the neighbouring vehicles is obtained using rule-based system. The status of the neighbouring vehicles are categorized as "safe-to-follow", "safe-to-ignore" and "ignore-with-caution". We validate our proposed framework 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.

1 Introduction

Automated driving research has gained prominence in the industry as well as the academia in recent years [John *et al.*2018]. In autonomous driving, environment perception, situation assessment and decision making play an important role. The intelligent processing of information from the vehicle sensors results in perceiving the environment. Following environment perception, vehicle semantic understanding is used for effective decision making [van Veen *et al.*2017].

In the complicated driving scene such as urban area, the semantic understanding of the neighbouring vehicles plays an important role for realizing the fully automated driving. Fig 1 depicts typical driving scene in the urban area, where the autonomous vehicle should not only detect and classify the surrounding vehicles, but should also assess their status.

In this research, a vision-based vehicle semantic understanding framework is proposed for automated driving in urban roads with multiple lanes using deep learning and rule-based system. In this framework, we firstly estimate the spatial-temporal information of all the vehicles in a video using deep learning-based environment perception [Krizhevsky *et al.*2012, Noh *et al.*2015, Sermanet *et al.*2014]. To estimate the spatial information for the vehicles, the YOLO-3, is utilized [Redmon and Farhadi2018]. A fine-tuned YOLO-3 is used to detect, localize and categorize all the vehicles in a given image according to their spatial location with respect to the autonomous vehicle.

The temporal information or binary motion status of all the vehicles in the road are estimated from a sequence of images using a novel multi-frame semantic segmentation framework, termed as the vehicle motion estimator (VME). The VME is able to estimate the motion status of all the vehicles across multiple frames without the need for tracking. The estimated vehicle spatial and motion information are then used for the semantic understanding of the neighbouring vehicles in a multiple lane urban road. The neighbouring vehicles are categorized as "**safe-to-follow**", "**safe-to-ignore**" or "**ignore-with-caution**" using a rule-based system.

Safe-to-Follow: In Fig 1, the front vehicle (red box) has either stopped for traffic, traffic signal or, alternatively, is "safe-to-follow". Subsequently, the automated vehicle should follow the front vehicle in these situations.

Safe-to-Ignore: Vehicles in the left, right and opposite lanes (green boxes), which have stopped for traffic, traffic light, temporary parking or turning, can be "ignored" by the automated vehicle, while it follows its predefined route.

Ignore-to-Caution: Vehicles in the left, right and opposite lanes (blue boxes) which are moving have to be "ignored-with-caution", as the possibility of the these vehicle entering our lanes, as part of lane change (left-right lane vehicles) or overtaking (opposite lane vehicles) needs to be considered for decision making.

Our main contribution to literature are as follows:

- Spatial vehicle localization with respect to the autonomous vehicle using fine-tuned YOLO-3.
- Novel vehicle motion estimation without the need for vehicle tracking.
- Vehicle semantic understanding framework using deep learning-based features and rule-based system.

The remainder of the paper is structured as follows. In Section 2, we survey the literature for situation assessment. Our proposed algorithm is presented in Section 3, and the

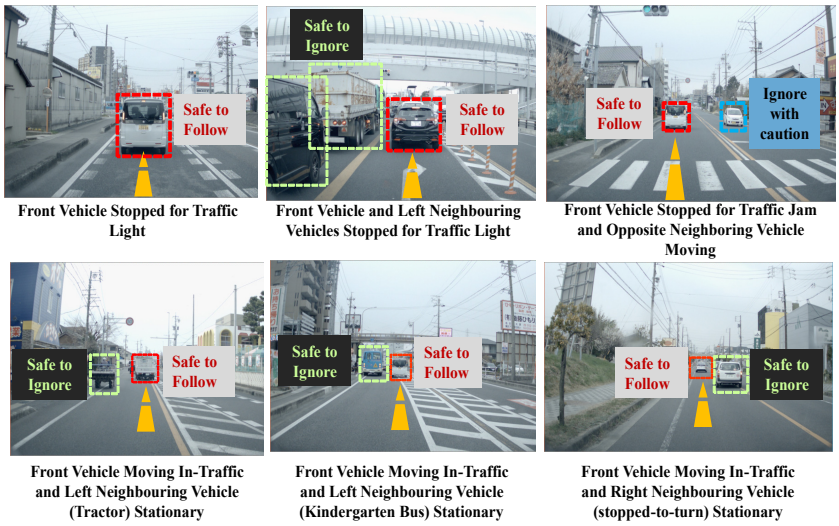


Figure 1: Typical driving scene for urban driving scene

results obtained are summarized in Section 4. Finally, we summarize our contributions and present directions for our future work in Section 5.

2 Related Work

To assess the situation, the environment information obtained from the perception module is often used, and modeled [Hermann and Desel2008]. Typically, the environment information is represented using techniques such as ontology [Zhao *et al.*2016, Bagschik *et al.*2018]. In ontology, the knowledge of the environment are represented using the concepts (classes) and the relationships (properties) between them. This representation is then used by the autonomous vehicle for situation assessment and decision making [Zhao *et al.*2016]. Alternatively, researchers have also used probabilistic methods for situation assessment [McAree *et al.*2017, Barbier *et al.*2018, Hillenbrand *et al.*2005]. In the works by Coue et al. [Rummelhard *et al.*2014, Laugier *et al.*2011], the authors use dynamic probabilistic grids and Bayesian occupancy filters for situation modeling and assessment, these are then used to make decisions for collision avoidance. Compared to the ontology methods, the probabilistic models also take into account the uncertainty involved in perceiving the environment and assessing the situation. Some researchers have combined ontology and probabilistic methods for situation assessment [Geng *et al.*2017]. In the work by Geng et al. [Geng *et al.*2017], the hidden Markov model is used to learn continuous features of driving behavior of target vehicles in the environment, which are then represented using ontology. This representation is used to predict or assess the situation of the target vehicles future behavior.

Compared to these works, we propose a rule-based situation assessment framework using deep learning-based vehicle spatial and temporal information. In our work, the temporal information is estimated using a novel multi-frame semantic segmentation framework, without the need for tracking, which is another contribution of our work. In literature, to estimate the temporal information, vehicle detection and tracking across frames [Hadi *et al.*2014], using techniques such as HMM [Jazayeri *et al.*2011], CAMShift [Xia *et al.*2013] etc, is used.

3 Algorithm

The vehicle spatial and motion information is estimated using deep learning framework. These estimated vehicle features are then used to categorize the vehicles using rule-based logic. We first provide a brief overview of the different modules in the proposed algorithm, followed by a detailed overview. An overview of the algorithm is presented in Fig 2.

Vehicle Spatial Information: Given a sequence of images, $\mathbf{z}(1:k)$, a fine-tuned YOLO-3 network is used to detect, localize and categorize all the vehicle's in each k-th frame. The detected vehicles are categorized into 5 classes by the fine-tuned YOLO-3 network based on their spatial location with respect to the autonomous vehicle, given as **front vehicle**, **left vehicle**, **right vehicle**, **opposite vehicle** and **other vehicle**. The **other vehicle** class includes vehicles which are either parked outside the road or travelling perpendicularly to the autonomous vehicle.

Vehicle Motion Information: The vehicle motion information is estimated using a novel semantic segmentation framework, termed as the vehicle motion estimator. Here, we generate a novel motion representation image template termed as the Vehicle Motion History Image (VHMI). The VHMI along with their corresponding pixel-level motion labels, are used as training pairs, to train a U-Net-based semantic segmentation framework, termed as the Vehicle Motion Estimator (VME), to estimate the motion information.

Vehicle Semantic Understanding: Using the estimated vehicle spatial and motion information, the vehicle situation is assessed using a rule-based logic. The assessed vehicles categories are given as "**safe-to-follow**", "**safe-to-ignore**" or "**ignore-with-caution**". We next present the detailed overview of the algorithm.

3.1 Detailed Overview

Vehicle Spatial Information

In our fine-tuned tiny YOLO-3 network, we use the same architecture as the original architecture, apart from modifying the 80 object classes in YOLO layers on the network to 5 object classes.

Vehicle Motion Information

As a precursor to the binary motion estimation, the bounding boxes of the vehicle detected, by the YOLO-3 network in the previous step, are accumulated across multiple frames into a single image template (VHMI). The VHMI is generated from a sequence of K frames, $\mathbf{z}(1:k)$, using the output of the fine-tuned YOLO-3 network, as follows,

Frame 1 (Initialization):

- Initialize a VHMI template, V, with same size as the image $z(1)$ and zero pixel values.

Iterate for Frame $(1:k)$:

1. Consider the output of the fine-tuned YOLO-3 vehicle detector, with N objects, $\{\lambda(1)_n\}_{n=1}^{N}$. Each detection λ=[$\mathbf{b}, c, p(c)$] contains the bounding box coordinates, $\mathbf{b}$, the object class, c, and their probability $p(c)$.
2. Omitting the **other vehicle** class detection, for each remaining detection, the detection bounding box region in the image $z(1)[(b)]$, are transferred to the VHMI template V. This is given as, $V[(b)]$= $z(1)[(b)$. (Fig 2-c).
3. Image regions from the latter frames overwrite the image regions corresponding to the initial frames in the VHMI template V.

An illustration of the VHMI generation is given in Fig 2-c. Following, the generation of V, its corresponding ground-truth label image G is obtained, manually. In G, the pixels corresponding to the moving and stationary bounding boxes in V are annotated as **moving vehicle** or **stationary vehicle**, while the other pixels are labeled as **background**. An illustration of the ground-truth annotation of G is shown in Fig 2-c.

The input-output pair V-G is used to train the multi-class VME which is formulated using a 3-class semantic segmentation framework. The VME architecture is based on the U-Net-based semantic segmentation [Ronneberger *et al.*2015, Badrinarayanan *et al.*2015]. However, unlike the original U-Net which is a binary segmentation framework, we utilize a multiclass segmentation framework. The VME is trained with a Adam optimizer with learning rate of 0.01, β of 0.9 with no decay. The VME is trained with categorical cross-entropy error function.

3.2 Vehicle Semantic Understanding

Given the estimated vehicle spatial and motion information, the vehicle situation is estimated using a rule-based system. The rule-based system is presented in Table 1. The rule-based system is designing considering the specific scenario of a urban road with multiple lanes (at-least one lane for automated vehicle and one lane for oncoming traffic).

Table 1: Deep Learning Spatial and Motion Feature-based Situation Assessment

Spatial Feature	Motion Feature	Situation Assessment
Ego Lane	Stationary	Safe-to-follow
Ego Lane	Motion	Safe-to-follow
Opposing Lane	Stationary	Safe-to-ignore
Opposing Lane	Motion	Ignore-with-caution
Left Lane	Stationary	Safe-to-ignore
Left Lane	Motion	Ignore-with-caution
Right Lane	Stationary	Safe-to-ignore
Right Lane	Motion	Ignore-with-caution

3.3 Algorithm: Training and Testing

In the training phase, firstly, the pre-trained YOLO-3 is fine-tuned for vehicle detection using our dataset. Secondly, the VME network is trained using the input-output pairs of generated VHMI templates V and ground truth annotations G.

In the testing phase, for a given sequence of k-test frames, the fine-tuned YOLO-3 is used to generate the N objects for each k-th frame $\{\lambda(k)_n\}_{n=1}^{N}$. This detection result corresponds to the vehicle spatial information.

The N objects detected over K frames are used to generate the test VHMI V. The test V is given as input to the trained VME, and the output G is obtained. Note that each G is obtained for an input sequence of k-test frames.

Subsequently, for the K-th or final frame of the test sequence, the N detected objects, $\{\lambda(K)_n\}_{n=1}^{N}$, are retrieved, and the **other vehicle** objects are omitted.

For each remaining object, the output region in G corresponding to the object bounding box coordinate (b) is retrieved. From the retrieved output region, the motion label with highest class probability is identified. Thus, the vehicle motion information is estimated.

Given the vehicle spatial and vehicle motion information, for every **non-other vehicle** objects in the K-th or final frame of the sequence, the high-level vehicle situation information is estimated using the rule-based system.

4 Experiments

The proposed algorithm is validated on acquired dataset with 7 sequences of urban road with multiple lanes, with at least one lane for the automated driving and one lane for the oncoming traffic. Each sequence has $50-300$ training and testing frames each. Samples of the dataset are shown in Fig **??**,4 and 1. The vehicle spatial class, motion class and vehicle situation assessment class bounding boxes where manually annotated. The dataset was acquired with an in-house camera. The algorithm was implemented on a Nvidia Titan X Ubuntu 16.04 machine using Keras-Theano backend [Chollet and others2015]. Some results of the algorithm for urban road of Kariya city in Japan are shown in Fig 3.

We also perform a comparative analysis of our algorithm as well as a parametric analysis. We report the precision and recall to validate the algorithms. We evaluate the different modules of our proposed framework with baseline algorithms.

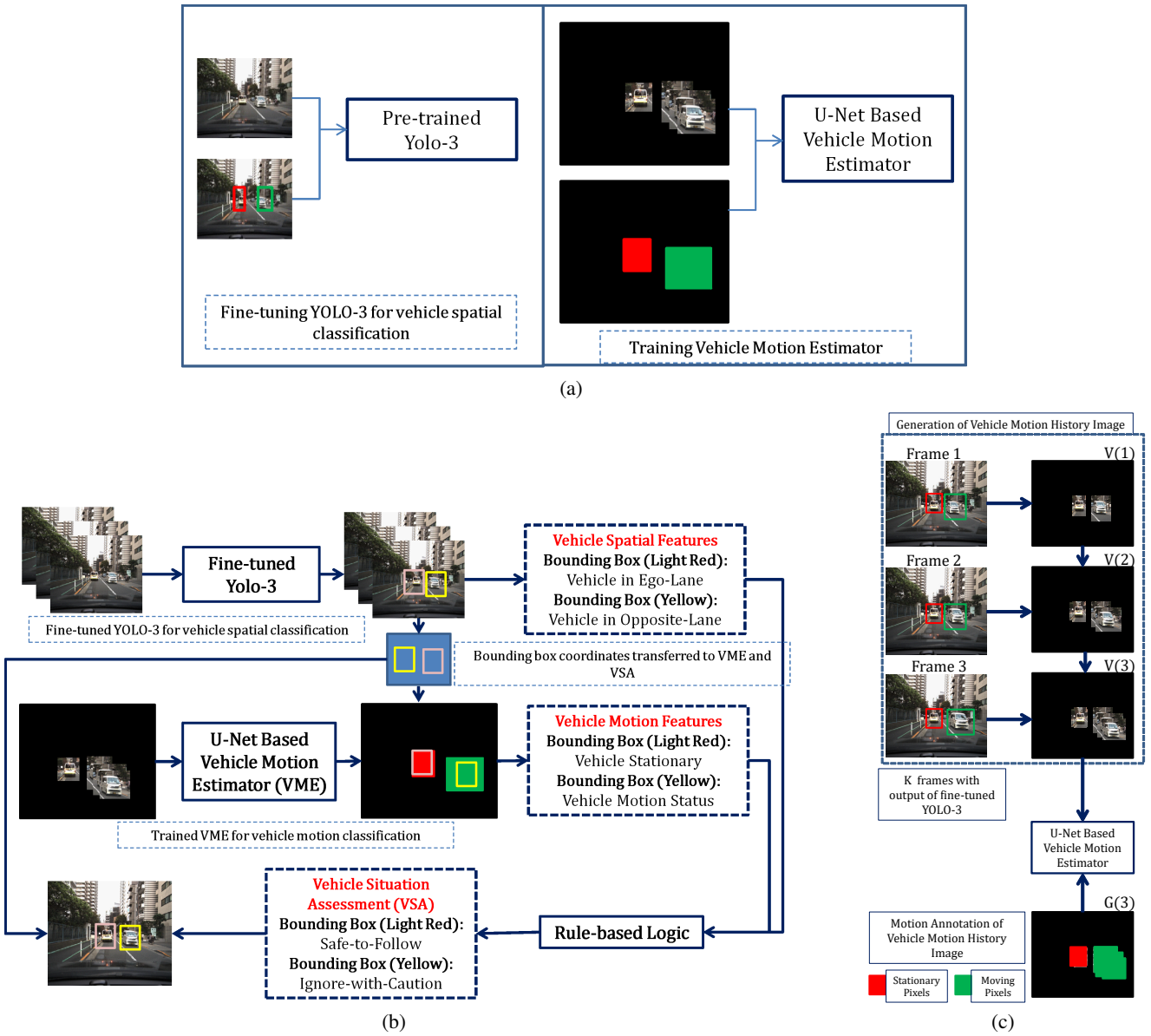


Figure 2: A detailed overview of the algorithm (a) training, (b) testing phase and (c) vmhi generation.

4.1 Vehicle Detector

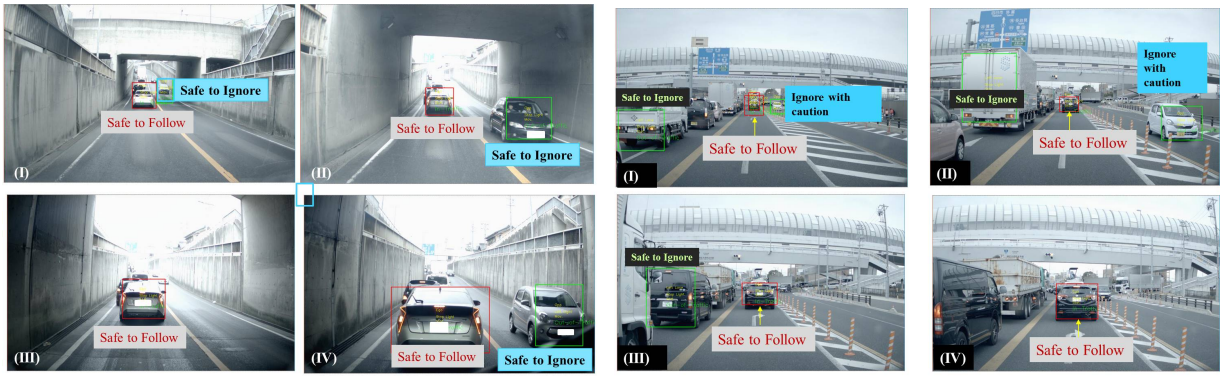
Firstly, we evaluate the performance of the fine-tuned tiny YOLO-3 network which reports a detection accuracy of 91.4% . On closer inspection we observe that the vehicle detector has high detection accuracy. However, there are indeed a few false positives, and few vehicles at distances greater than 50m which are missed has reported by the precision and recall of 0.95 and 0.92, respectively.

4.2 Vehicle Motion Estimation

To validate the proposed VME, we perform a comparative analysis with an optical flow-based deep learning semantic segmentation framework [Sevilla-Lara *et al.*2016]. In the baseline network which is based on U-Net, there are two input branches, and one output branch. The intensity image of the K-th frame is given as the input to the first input branch. For the second input branch, the derived optical flow components between the k-th and k-1 th frame is used as the input. Additionally, we have skip connections between the convolutional layers in encoding side and the convolutional layers in the decoding side. Similar to the VME, a categorical cross entropy is used to train the network. The results tabulated in Table 3 show that the proposed framework is better than the baseline algorithm.

4.3 rule-based system

The rule-based system is validated with several input variations, given as,



(a) (b)

Figure 3: Some samples of assessment of the detected vehicles in the urban area.

Table 2: Validation of the Vehicle Motion Estimation

Algo.	Precision	Recall	Det Accuracy
VME	0.94	0.91	91.06%
Opt Flow-based U-Net	0.93	0.89	70.25%

Vehicle Spatial Feature Only: Only the deep learning-based vehicle spatial feature is given as input to the rule-based system.

Vehicle Motion Feature Only: Only the deep learning-based vehicle motion feature is given as input to the rule-based system.

Lane ROI-Vehicle Spatial Feature: Here, the deep learning-based motion feature and a non-deep learning-based spatial feature is given as input to the rule-based system. More specifically, a region-of-interest (ROI) based spatial feature is given as input using fixed lane ROIs in the image

To estimate the lane ROI-based vehicle spatial feature, the pre-trained YOLO-3 network without fine-tuning is used to detect all vehicles in the road. Subsequently, the vehicle spatial feature is estimated by matching the bounding box centroid with the closest lane ROI.

Table 3: Validation of the rule-based system

Algo.	Precision	Recall	Detection Accuracy
Proposed	0.94	0.91	91.06%
Spatial Feat. only	0.92	0.89	89.79%
Motion Feat. only	0.90	0.77	77.45%
Lane ROI	0.90	0.81	81.25%

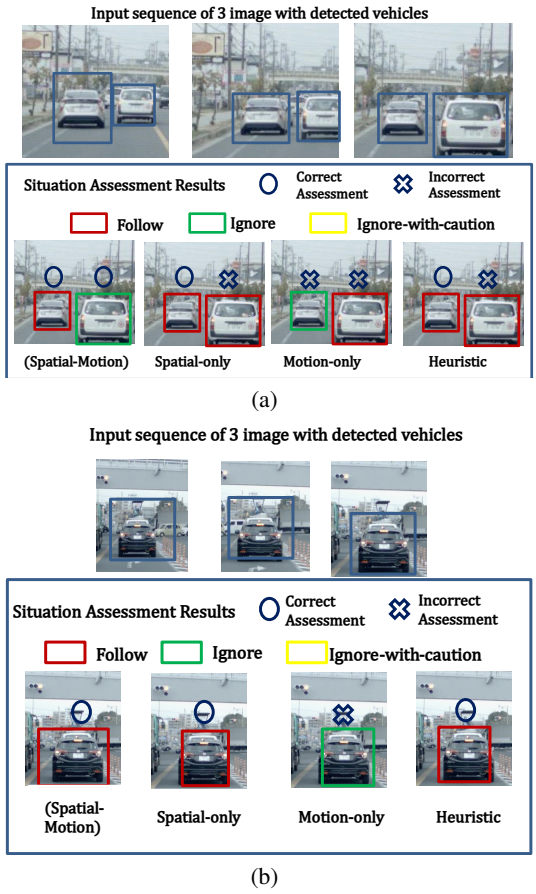


Figure 4: Result of the different models for (a) complex and (b) simple scene.

4.4 Discussion

On closer observation of the results, we see that the spatial feature alone and motion feature alone-based vehicle semantic understanding are inferior. In Fig 4, we can observe the performance of the different models for complex and simple scenes. In case of a simple scene (Fig 4-b), the performance of the proposed framework, spatial-only and lane ROI spatial models are similar. While, the performance of the motion-only model is inferior, as the motion alone is not sufficient to categorize the vehicle as "safe-to-follow" or "safe-to-ignore". The advantage of the proposed framework with spatial and motion model is more clearer in case of a complex scene

(Fig 4-a)

The computational time of the different modules are as follows, the vehicle spatial categorization takes 25ms per frame, or 75ms for the three frame sequence. The vehicle motion estimation takes 27 ms for the three frame sequence.

The results show that the proposed rule-based system with deep learning-based spatial and motion features obtain nearly 90% accuracy for vehicle semantic understanding for urban roads with multiple lanes.

To further enhance the system for autonomous driving in various different road scenarios, such as rural and shared single lane roads, further image-based features such as vehicle light information, lane marker information and road surface information are required. The performance of the rule-based system under this scenario will be evaluated in our future work.

5 Summary and Conclusion

In this research, a vision-based vehicle semantic understanding framework is proposed for an autonomous vehicle. To obtain the semantic understanding of neighbouring vehicles, a vision-based vehicle spatial and temporal estimation framework is proposed using a fine-tuned YOLO-3 and a novel multi-frame semantic segmentation framework. Using the estimated vehicle spatial and motion information, a rule-based system is used for the semantic understanding of the neighbouring vehicles. The experimental results show that the proposed framework is better than the baseline algorithms.

References

[Badrinarayanan *et al.*, 2015] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[Bagschik *et al.*, 2018] Gerrit Bagschik, Till Menzel, and Markus Maurer. Ontology based scene creation for the development of automated vehicles. In *Intelligent Vehicles Symposium*, pages 1813–1820, 2018.

[Barbier *et al.*, 2018] Mathieu Barbier, Christian Laugier, Olivier Simonin, and Javier Ibanez-Guzmin. Probabilistic decision-making at road intersections: Formulation and quantitative evaluation. In *15th International Conference on Control, Automation, Robotics and Vision, ICARCV*, pages 795–802, 2018.

[Chollet and others, 2015] François Chollet et al. Keras. https://keras.io, 2015.

[Geng *et al.*, 2017] Xinli Geng, Huawei Liang, Biao Yu, Pan Zhao, Liuwei He, and Rulin Huang. A scenario-adaptive driving behavior prediction approach to urban autonomous driving. *Applied Sciences*, 7(4), 2017.

[Hadi *et al.*, 2014] Raad Ahmed Hadi, Ghazali Sulong, and Loay Edwar George. Vehicle detection and tracking techniques : A concise review. *Signal Image Processing : An International Journal*, 5(1):1–12, 2014.

[Hermann and Desel, 2008] A. Hermann and J. Desel. Driving situation analysis in automotive environment. In *2008 IEEE International Conference on Vehicular Electronics and Safety*, pages 216–221, 2008.

[Hillenbrand *et al.*, 2005] J. Hillenbrand, K. Kroschel, and V. Schmid. Situation assessment algorithm for a collision prevention assistant. In *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, pages 459–465, 2005.

[Jazayeri *et al.*, 2011] A. Jazayeri, H. Cai, J. Y. Zheng, and M. Tuceryan. Vehicle detection and tracking in car video based on motion model. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):583–595, 2011.

[John *et al.*, 2018] Vijay John, Nithilan Meenakshi Karunakaran, Chunzhao Guo, Kiyosumi Kidono, and Seiichi Mita. Free space, visible and missing lane marker estimation using the psinet and extra trees regression. In *24th International Conference on Pattern Recognition*, pages 189–194, 2018.

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Laugier *et al.*, 2011] Christian Laugier, Igor E. Paromtchik, Mathias Perrollaz, Yong Mao, John-David Yoder, Christopher Tay, Kamel Mekhnacha, and Amaury Nègre. Probabilistic analysis of dynamic scenes and collision risks assessment to improve driving safety. *IEEE Intell. Transport. Syst. Mag.*, 3(4):4–19, 2011.

[McAree *et al.*, 2017] Owen McAree, Jonathan M Aitken, and Sandor M Veres. Towards artificial situation awareness by autonomous vehicles. *IFAC-PapersOnLine*, 50:7038–7043, 2017.

[Noh *et al.*, 2015] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. *CoRR*, abs/1505.04366, 2015.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[Ronneberger *et al.*, 2015] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[Rummelhard *et al.*, 2014] Lukas Rummelhard, Amaury Nègre, Mathias Perrollaz, and Christian Laugier. Probabilistic grid-based collision risk prediction for driving application. In *Experimental Robotics - The 14th International Symposium on Experimental Robotics, ISER*, pages 821–834, 2014.

[Sermanet *et al.*, 2014] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proceedings of the International Conference on Learning Representations*, 2014.

[Sevilla-Lara *et al.*, 2016] Laura Sevilla-Lara, Deqing Sun, Varun Jampani, and Michael J. Black. Optical flow with semantic segmentation and localized layers. *2016 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 3889–3898, 2016.

[van Veen *et al.*, 2017] Tom van Veen, Juffrizal Karjanto, and Jacques Terken. Situation awareness in automated vehicles through proximal peripheral light signals. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '17, pages 287–292, 2017.

[Xia *et al.*, 2013] Jingxin Xia, Wenming Rao, Wei Huang, and Zhenbo Lu. Automatic multi-vehicle tracking using video cameras: An improved camshift approach. *KSCE Journal of Civil Engineering*, 17(6):1462–1470, Sep 2013.

[Zhao *et al.*, 2016] Lihua Zhao, Ryutaro Ichise, Yutaka Sasaki, Zheng Liu, and Tatsuya Yoshikawa. Fast decision making using ontology-based knowledge base. In *Intelligent Vehicles Symposium*, pages 173–178, 2016.