Semi-supervised Learning: Fusion of Self-supervised, Supervised Learning, and Multimodal Cues for Tactical Driver Behavior Detection

Athma Narayanan
Honda Research Institute, USA
anarayanan@honda-ri.com

Yi-Ting Chen
Honda Research Institute, USA
ychen@honda-ri.com

Srikanth Malla
Honda Research Institute, USA
smalla@honda-ri.com

Abstract

In this paper, we presented a preliminary study for tactical driver behavior detection from untrimmed naturalistic driving recordings. While supervised learning based detection is a common approach, it suffers when labeled data is scarce. Manual annotation is both time-consuming and expensive. To emphasize this problem, we experimented on a 104-hour real-world naturalistic driving dataset with a set of predefined driving behaviors annotated. There are three challenges in the dataset. First, predefined driving behaviors are sparse in a naturalistic driving setting. Second, the distribution of driving behaviors is long-tail. Third, a huge intra-class variation is observed. To address these issues, recent self-supervised and supervised learning and fusion of multimodal cues are leveraged into our architecture design. Preliminary experiments and discussions are reported.

1. Introduction

Intelligent transportation systems require interdisciplinary efforts including computer vision, machine learning, robotics, psychology, and control theory. It is challenging to drive in real world because decisions need to be made with incomplete information and diverse situations. Moreover, modeling uncertain behaviors of road users is still unsolved.

Towards this goal, we collected a naturalistic driving dataset, which will appear in CVPR'18 main conference [1]The total size of the dataset is 104 video hours with the predefined driving behaviors annotated. We defined a 4-layer scheme to annotate driver behaviors including tactical driver behaviors and interactive behaviors between the drivers and traffic participants. More details of the annotation and definition of driver behaviors will be provided in the supplementary material. Note that driving behaviors are a combination of driver behaviors, the interactive behaviors between driver and traffic participants, and traffic participants' behaviors. In this paper, we focus on detecting tactical driver behaviors as in [1].

Manual annotation of driving behaviors is time-consuming and expensive. To minimize human efforts, automatic detection mechanism is necessary. While supervised learning is a common approach to address the problem, it suffers from when labeled data is scarce. This issue is presented in the collected dataset. Specifically, the dataset has the following three challenges. First, predefined driving behaviors are sparse. Only 15% of data is labeled. Most of the time, drivers are doing "going straight," "stopping for red light," and "parking." Second, the distribution of driving behaviors is long-tail. For example, we observe more "turning" than "U-turn." Third, a huge intra-class variation is observed. For instance, a "turning right" is different from a "turning right while yielding a group of pedestrians."

We leverage recent advances in self-supervised learning for structure from motion [10], supervised learning for semantic segmentation [6, 2], imbalance class distribution handling [7], and multimodal fusion [4] to address aforementioned issues. The proposed algorithm is presented.

2. Methodology and Experiments

We hypothesize that semantic context, 3D scene structure and vehicle motion are crucial tactical driver behavior detection. We intended to leverage features extracted from these cues than using features trained by supervision. This section gives the details of the architecture design and the preliminary results.

Given a synchronized images and Controller Area Network (CAN bus) sensors data, the baseline model [1] sampled input frames from video streams and values from CAN
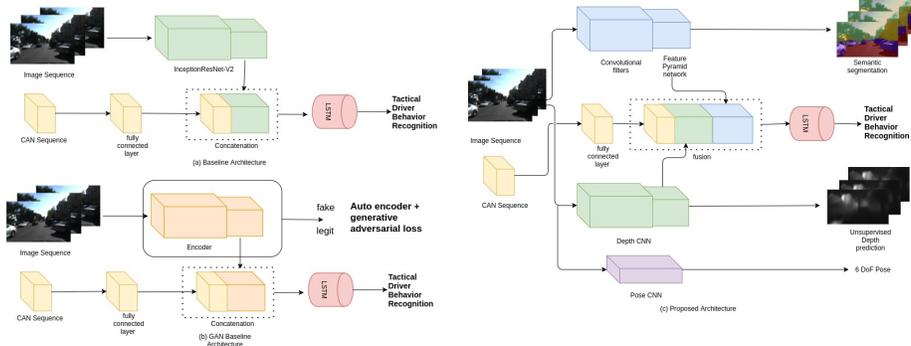
Figure 1: Figure (a) The baseline architecture combining InceptionResNet-V2 Image features and CAN sensor data [1]. Figure (b) A baseline architecture using features obtained by [8]. Figure (c) The proposed architecture combining self-supervised learning, semi-supervised learning and multimodal cues is presented for tactical driver behavior recognition.

bus sensors at 3 Hz. The frame representation is extracted from the *Conv2d_7b_1x1* layer of InceptionResnet-V2 [9] pretrained on ImageNet [3]. The features are convolved with a 1×1 convolution to reduce dimensionality from $8 \times 8 \times 1536$ to $8 \times 8 \times 20$. Raw sensor values are passed through a fully-connected layer to obtain a one dimensional feature vector which is further concatenated with image features. The concatenated features are fed into a Long-Short Term Memory (LSTM) [5] to encode the necessary history of past measurements. During training, we formed batches of sequence segments by sequentially iterating over driving sessions. The last LSTM hidden state from the previous batch is used to initialize the LSTM hidden state on the next step. The training is performed using truncated backpropagation through time. Addtionally, the data imbalance between foreground and background frames are handled using the recently proposed technique for modifying cross-entropy loss to deal with the class imbalance [7]. Note that the details of the training protocol will be provided in the supplementary material.

Five different experiments were conducted as shown in Table 1. Note that we adopted the same architecture as in [1] to detect tactical driver behaviors, but with different image features. First, we presented the baseline as in [1]. Second, we trained an auto-encoder with adversarial loss to generate image features as in [8]. We expect the reconstruction of images can learn the scene composition of the dataset. To reduce the reliance on direct supervision, we leveraged image reconstruction features to serve as a substitute. We took the intermediate encoder feature representation to the LSTM. Third, as 3D scene structure is crucial, we leveraged the unsupervised learning based structure from motion [10]. Fourth, for semantic context, we modify Deeplab [2] to incorporate Feature Pyramid Network [6] to enrich features at higher resolution features. Finally, features from 3D scene structure and semantic context are fused with CAN bus features by concatenation and batch normalization, similar to

Table 1: Experimental results on a set of 104-hour data. All number are in %.

Driver behavior class	[1]	[8]	Depth+CAN	Seg+CAN	Our
left lane change	35.72	14.06	**38.96**	37.13	34.45
right lane change	25.48	6.42	25.32	23.25	**28.06**
railroad passing	**7.27**	0.14	0.82	3.06	5.40
left lane branch	20.00	4.09	26.68	35.94	**43.05**
right lane branch	0.74	0.59	1.58	**2.97**	2.10
left turn	73.52	66.00	74.21	**77.78**	75.07
right turn	73.95	73.52	**76.20**	75.63	75.82
U-turn	15.78	26.77	**32.54**	27.77	26.40
intersection passing	74.12	29.45	69.98	**79.69**	77.70
crosswalk passing	4.04	0.63	6.65	**14.02**	13.14
merge	6.35	0.40	9.17	14.83	**16.42**
mean	30.63	20.19	32.92	35.64	**36.15**

the work done in [4]. The aforementioned architectures are shown in Figure 1.

3. Discussion

Our experiments indicate that robust fusion of image features from auxiliary tasks such as 3D scene structure and semantic context help the driver behavior detection tasks as demonstrated in Table 1. With semantic context and 3D scene structure, we see improvements in classes such as intersection passing, cross-walk passing, U-turn and merge class. The proposed architecture improves the performance of [8] by 16 %. This demonstrates the effectiveness of the proposed features over features obtained by reconstruction.

However, we expected a significant performance boost in those behaviors with strong correlations to semantic context (e.g., lane change due to the existence of lane markers in semantic context). The fusion of semantic context with CAN (i.e., 4th column results) does not reflect this hypothesis. Note that **railroad** is not trained in the current semantic context algorithm. A better architecture design in the multimodal fusion, imbalanced distribution, and temporal modeling is necessary for further improvement.

References

[1] Anonymous. Anonymous. In *CVPR*, 2018. 1, 2

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *PAMI)*, 40(4):834 – 848, 2018. 1, 2

[3] J. Deng, W. Dong, R. Socher, L. jia Li, K. Li, and L. Fei-fei. ImageNet: A large-scale Hierarchical Image Database. In *CVPR*, 2009. 2

[4] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusenet: Incorporating Depth Into Semantic Segmentation via Fusion-based CNN Architecture. In *ACCV*, 2016. 1, 2

[5] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[6] T.-Y. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature Pyramid Networks for Object Detection. In *CVPR*, 2017. 1, 2

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal Loss for Dense Object Detection. In *ICCV*, 2017. 1, 2

[8] E. Santana and G. Hotz. Learning a Driving Simulator. *arXiv preprint arXiv:1608.01230*, 2016. 2

[9] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *CoRR*, abs/1602.07261, 2016. 2

[10] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. In *CVPR*, 2017. 1, 2