Mixed Spatiotemporal Modeling for Vehicle Behavior Recognition

Ying Zhang[1], Yaochen Li[1]*, Wei Guo[1], Gaojie Li[1], Shaohan Yang[2], Yuehu Liu[3]

[1]School of Software Engineering, Xi'an Jiaotong University, China

[2]Department of Information and Computing Science, Xi'an Jiaotong-Liverpool University, China

[3]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

Vehicle behavior recognition is a crucial step in perceiving the driving environment for autonomous vehicles. We propose a novel mixed spatiotemporal modeling network based on 2D CNN for vehicle behavior recognition. The network is constructed by alternating temporal and spatial modeling to help generate deep and effective spatiotemporal features.

➢ For temporal relationships capturing, a Complementary Temporal Extraction (CTE) block is proposed to capture global temporal evolution and motion information in videos, including a Global Spatiotemporal Modeling (GSTM) module, a Motion Excitation (ME) module and a fusion approach designed to merge long and short term features.

➢ For spatial information learning, a Channel-enhanced Spatial Extraction (CSE) block is designed, including a modified channel attention module called Spatial Efficient Channel Attention (Spatial-ECA) to learn the relationship between channels.

➢ By integrating these modules into the standard residual block with ResNet-50 as the backbone, our method achieves the highest accuracy on vehicle behavior datasets and has a lower model complexity than other 2D CNN-based methods.

Figure 1. The visualization results of vehicle behavior recognition.

Methods

The architecture of our mixed spatiotemporal modeling network is shown in Figure 2.

➢ Using 2D ResNet-50 as the backbone, we retain the Conv1 and Conv2_x layers to extract low-level spatial features. Then the CTE block and the CSE block are deployed alternately in the Conv3_x, Conv4_x, and Conv5_x layers to generate deep and effective spatiotemporal feature maps for vehicle behavior recognition.

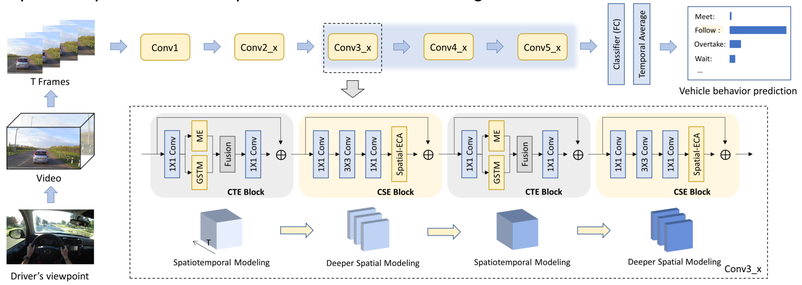

Figure 2. The architecture of our mixed spatiotemporal modeling network.

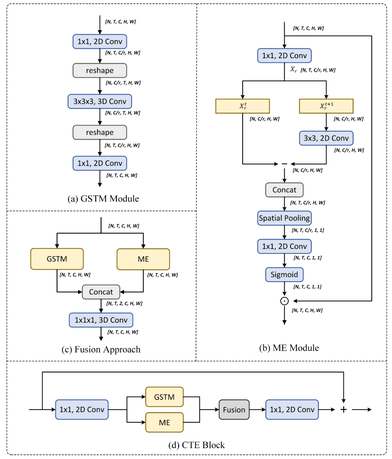

Figure 3. The CTE block.

➢ The CTE block is proposed to capture the long-term spatiotemporal relationship and short-term motion information.

➢ The CSE block is designed for spatial modeling, including a Spatial-ECA module to learn the relationship between channels.

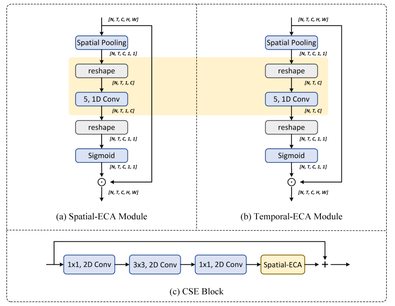

Figure 4. The CSE block.

Results

The experiments are conducted on the HDD+ dataset and the BDD100K dataset annotated in our previous work[1].

➢ We compare our method with three representative 2D CNN-based methods TSN, TSM and ACTION-Net. Our method has the highest performance and lowest complexity, which is very cost-effective.

Method	Backbone	FLOPs	△FLOPs	Params	△Params	BDD100K	
						Top-1	Top-2
TSN	ResNet-50	33G	-	23.56M	-	77.86	93.73
TSM	ResNet-50	33G	-	23.56M	-	92.37	93.78
ACITON-Net	ResNet-50	34.67G	+1.67G (+5.10%)	27.74M	+4.18M (+17.74%)	95.94	98.52
Ours	ResNet-50	**27.53G**	**-5.47G (-16.58%)**	**19.28M**	**-4.28M (-18.17%)**	**97.79**	**99.26**

Table 1. Comparisons with the 2D CNN-based methods.

➢ We compare our method with the state-of-the-art approaches on two datasets. Our method achieves the highest accuracy on both the BDD100K and HDD+ datasets, and has significant advantages on computational complexity and parameter number, with an impressive performance on the vehicle behavior recognition task.

Method	Backbone	Pretrain	Frames	FLOPs	Params	BDD100K		HDD+	
						Top-1	Top-2	Top-1	Top-2
C3D	-	-	16	38.55G	78.38M	81.92	86.35	87.25	90.49
R3D	3D ResNet-18	-	16	40.89G	33.18M	87.82	92.25	93.75	94.10
SELayer-3DCNN	-	-	16	**10.77G**	**3.85M**	90.04	95.94	92.44	93.44
TSM	ResNet-50	ImageNet	8	33G	23.56M	92.62	98.52	96.39	98.63
TEA	ResNet-50	ImageNet	8	34.74G	24.12M	92.99	97.05	96.56	98.97
TDN	ResNet-50	ImageNet	8	36G	24.07M	94.83	98.16	96.22	98.80
ACITON-Net	ResNet-50	ImageNet	8	34.67G	27.74M	95.94	98.52	97.77	98.97
Ours	ResNet-50	ImageNet	8	27.53G	19.28M	**97.79**	**99.26**	**97.94**	**99.66**

Table 2. Comparisons with the state-of-the-arts.

Conclusion

In this paper, we propose a new mixed spatiotemporal modeling network for vehicle behavior recognition. By mixed spatiotemporal modeling, the network generates deep and effective spatiotemporal features with smaller fusion complexity to help classify vehicle behavior.

Contact

*Corresponding Author, E-mail: yaochenli@mail.xjtu.edu.cn

References

[1] Yaochen Li, Haochuan Hou, et al. Spatiotemporal Analysis of Static and Dynamic Traffic Elements From Road Scenes. IEEE Transactions on Intelligent Transportation Systems, 2022.