Multi-modal Large Language Model for Training-free Vision-based Driver State Recognition

Chuanfei Hu
Southeast University
Nanjing, China
chuanfei_hu@ieee.org

Xinde Li
Southeast University
Nanjing, China
xindeli@seu.edu.cn

Hang Shao
Qingdao University
Qingdao, China
shaohang@qdu.edu.cn

Abstract

Recent advances focus on modeling a learning-based method to realize the driver monitoring system, benefiting from the powerful capability of data-driven feature extraction. Although the acceptable performances of these methods are achieved, the training procedure with massive data would significantly increase the labor costs. Thus, it is intuitive to explore a training-free vision-based driver state recognition in the era of large language model (LLM)/multi-modal large language model (MLLM). In this paper, we focus on a vision-based driver state monitoring method, where a novel training-free driver state recognition method via human-centric context and self-uncertainty-driven MLLM (HSUM). Extensive experiments are conducted on two public benchmarks, where the competitive performance of HSUM is demonstrated compared with the state-of-the-art training-based methods.

1. Introduction

Driver state is one of the vital factors which can significantly impact the vehicle operation. The positive states (such as concentration behavior and peace emotion) and negative states (such as distraction behavior and anxiety emotion) will strengthen the safety of driving and lead the growth of road traffic risks, respectively. Although the autonomous driving technology has gradually grown as the real-life applications [18], the Driving Automation Levels [1] ranging Level 0 to Level 3 still need the drivers to fully or partially engage in vehicle control [23]. At Level 0, the driver has complete control over the vehicle, and their state directly affects their ability to make safe driving decisions. Thus, driver monitoring system (DMS) plays one of the key components of guaranteeing the driving safety, which has attracted constant attention and interest from both the academic and industrial communities [20].

In the past few decades, the vision-based monitoring methods have emerged as the powerful technology [21] which is cost-efficient to perceive the richest information. The vision-based DMS can analyze the visual appearances (e.g. posture, gesture, facial expression, and action) to capture the potential negative driver states. Then, the driver will be alert to improve driving attention. Here, the learning-based methods [2] are the dominant tools to bridge the implicit gap between visual appearances and driver states, due to the superiority of learning-based methods for feature abstraction and inference [30].

Recently, the success of deep learning has been witnessed in many real-world applications [3, 9, 16, 17, 24, 25, 28]. Deep learning-based methods are capable of learning more robust and discriminative features from data automatically, which can avoid the cumbersome procedure of handcrafted feature extraction. For the vision-driven DMS, the deep learning-based methods [12, 22, 33] can be categorized into video-based and image-based methods. Since the driver videos convey the more contextual information than the images, we focus on exploring the video-based method to analyze the driver states. Although the learning-based driver state recognition methods have achieved the considerable performances, there is a fact that these superior methods rely on ***training procedure with massive data***, resulting in the significant increase in labor costs [8]. Meanwhile, the trained models might not be adaptive to the unseen classes sufficiently in real-world scenarios. Most recent years, the emergence of numerous large language models (LLMs) has attracted the significant attentions of cross-domain researchers [26], since the remarkable capability of LLMs has been achieved to analyze the human language via textual prompt and generate the understandable texts for natural language processing (NLP) tasks, such as text generation, sentiment analysis, and machine translation. The paradigm of LLM-based method is to first design a task-aware textual prompt, and then, assemble the source texts and textual prompt as the input of LLM to generate the expected texts. Even more recently, LLMs have been extended into multi-modal LLMs (MLLMs) [7, 31, 34, 35], in which the remarkable capability of LLMs is extended to deal with multi-modal sources, such as image, video, and audio information. These works "hug" the general reasoning capability

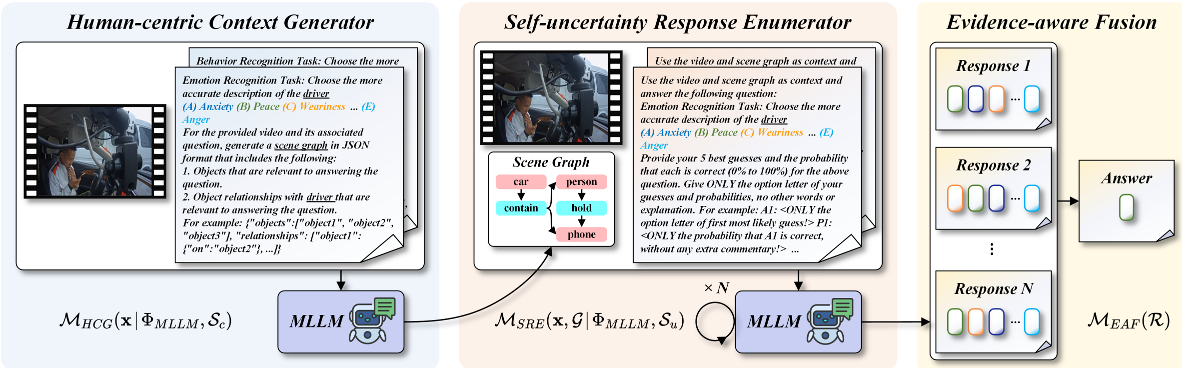

Figure 1. Overview of Human-centric context and self-uncertainty-driven MLLM (HSUM).

of MLLMs to achieve the competitive performances compared with supervised learning-based methods. Thus, there is an open question remains: "***Can we explore a method to reason the driver states from videos without training procedure in the era of LLM/MLLM?***"

To answer this question, we propose a novel training-free driver state recognition method via *human-centric context and self-uncertainty-driven MLLM* (HSUM). Specifically, a human-centric context generator (HCG) is first proposed based on a context-specific prompt. MLLM is guided to capture the human-centric contextual cues as a scene graph [6], which is powerful to represent the rich semantic relationships between objects, as well as the contextual interaction of objects with their surroundings, such as visual relationship detection. It would improve the MLLM capability of understanding the relationships between objects and their context. Then, a self-uncertainty response enumerator (SRE) is proposed to exploit the uncertainty of MLLM. The potential reasoning responses are enumerated repeatedly based on the assembly of the human-centric context and uncertainty-specific prompt. Furthermore, to reveal the precise reasoning result from the enumerated responses, we introduce the Dempster-Shafer evidence theory [27] (DST)-based combination rule to conduct an evidence-aware fusion (EAF). The enumerated responses are modeled as the evidences, while the fusion relationships among the evidences are analyzed via DST-based combination rule. The precise answer could be gathered theoretically, where the uncertainty of MLLM is mitigated relatively.

2. Methodology

The overall framework of HSUM is shown in Fig. 1, which consists of human-centric context generator (HCG), self-uncertainty response enumerator (SRE), and evidence-aware fusion (EAF). Let us denote a driver video with T frames $\mathbf{x} \in \mathbb{R}^{T\times C\times H\times W}$, annotated with a driver state label $\mathbf{y} \in \mathbb{H}^{K\times 1}$, where C, H, and W denote the number of color channels, height, and width, respectively. K denotes the number of driver state classes, and $\mathbb{H}$ is Hamming space. Specifically, the human-centric context $\mathcal{G}$ of $\mathbf{x}$ is first generated via HCG $\mathcal{M}_{HCG}$ as follows:

$$\mathcal{G} = \mathcal{M}_{HCG}(\mathbf{x}|\Phi_{MLLM}, \mathcal{S}_c), \tag{1}$$

where Φ_{MLLM} denotes the MLLM. $\mathcal{G}$ is the scene graph to present the human-centric context, which consists of objects and relationships. $\mathcal{S}_c$ denotes the string of context-specific prompt to guide the MLLM. Then, the potential responses are enumerated N times via SRE to explore the uncertainty of MLLM as follows:

$$\mathcal{R} = \mathcal{M}_{SRE}(\mathbf{x}, \mathcal{G}|\Phi_{MLLM}, \mathcal{S}_u), \tag{2}$$

where $\mathcal{S}_u$ denotes the string of uncertainty-specific prompt to guide the MLLM, and $\mathcal{R} = \{r_1, ..., r_N\}$ is a set including N potential responses from MLLM. Finally, EAF is conducted to model the enumerated responses as the evidences based on DST, while the precise "answer" e_* is revealed via DST-based combination rule as follows:

$$e_* = \mathcal{M}_{EAF}(\mathcal{R}). \tag{3}$$

Table 1. Comparison of HSUM with the state-of-the-art methods on AIDE and 3MDAD. ACC (%)↑, F1 (%)↑, CG-ACC (%)↑, and CG-F1 (%)↑ are utilized to evaluate the performance,where the best results are highlighted in **bold**.

Method	Backbone	AIDE $\mathcal{T}_{\mathrm{DDR}}$				3MDAD $\mathcal{T}_{\mathrm{DDR}}$				AIDE $\mathcal{T}_{\mathrm{DER}}$			
		ACC	F1	CG-ACC	CG-F1	ACC	F1	CG-ACC	CG-F1	ACC	F1	CG-ACC	CG-F1
Training-based													
VGG16 [29]	CNN	62.34	57.33	72.66	72.73	68.12	63.73	76.34	76.11	69.31	64.67	71.23	67.79
ResNet34 [15]	CNN	59.77	54.64	73.01	72.75	65.62	61.19	71.75	71.67	69.68	64.83	72.62	68.75
I3D [5]	CNN	**66.17**	61.35	74.38	74.36	**69.37**	64.63	76.93	76.37	70.94	65.99	71.43	68.05
SlowFast [11]	CNN	61.58	59.41	75.53	75.73	66.25	62.95	76.98	76.13	72.38	70.77	75.17	74.24
TimeSFormer [4]	ViT	65.18	63.24	73.73	73.91	68.75	66.39	77.31	77.53	74.87	72.56	76.52	74.92
DriveCLIP [14]	ViT	66.01	64.23	75.73	75.47	68.98	**66.73**	**78.67**	**78.53**	**75.56**	**73.63**	**78.78**	**76.15**
SRLF-Net [13]	ViT	**66.17**	**64.45**	**75.89**	**75.69**	69.11	66.56	78.53	78.13	75.20	73.31	78.65	75.91
Training-free													
mPLUG-Owl3 [34]	MLLM	53.03	47.24	61.96	61.38	52.17	49.48	60.34	60.01	56.90	54.34	61.69	59.83
Qwen2-VL [31]	MLLM	55.48	49.93	64.50	64.26	56.62	52.04	62.91	62.53	58.45	56.82	64.14	62.40
LLaVA-Video [35]	MLLM	54.92	49.37	63.94	63.71	56.06	51.48	62.45	62.07	58.12	56.26	63.58	61.83
VideoLLaMA2 [7]	MLLM	55.87	48.42	64.89	62.65	57.01	50.53	63.37	61.14	59.08	57.30	64.54	62.78
HSUM (Ours)	MLLM	**61.74**	**57.60**	**71.59**	**71.75**	**63.87**	**59.11**	**70.11**	**69.95**	**69.12**	**64.83**	**71.23**	**68.80**

3. Experiments

Dataset: The experiments are conducted on the two public benchmarks for the driver state monitoring task, where the driver distraction recognition ($\mathcal{T}_{\mathrm{DDR}}$) and driver emotion recognition ($\mathcal{T}_{\mathrm{DER}}$) are introduced as the evaluation tasks. **AIDE** [32] consists of 2898 video samples with 521.64K frames. Each sample of subject is captured via an in-car camera, annotated with bounding boxes (body and face) and states (7 behavior classes and 5 emotion classes). The dataset is split into the training, validation and testing sets with 65%, 15% and 20%, respectively. **3MDAD** [19] collects 1120 video samples with 574.13K frames during the daytime and the night, where the samples are annotated with driver behaviors (16 classes) and head positions. Here, we introduce the daytime samples which are split into the training and testing sets with 80% and 20%, respectively.

Evaluation Metric: The evaluation experiments are conducted to the driver state recognition, including the driver emotion recognition and driver behavior recognition tasks. Similar to [32], the classification accuracy (ACC), weighted F1 score (F1), coarse-grained accuracy (CG-ACC), and F1 score (CG-F1) are utilized to evaluate the performance of recognition. CG-ACC and CG-F1 are designed based on polarity emotions and anomaly behaviors, which consider the demand for practicality in DMS.

Comparisons with Other Methods: We compare the state-of-the-art methods categorized as general methods (VGG16 [29], ResNet34 [15], I3D [5], SlowFast [11], and TimeSFormer [4]) and specific methods (DriveCLIP [14] and SRLF-Net [13]), where the backbones involve CNN [5, 15, 29] and ViT [10]. Here, these methods are trained based on the paradigm of supervised learning. Meanwhile, these MLLMs used for HSUM had not prior access to labeled data of AIDE and 3MDAD, which is relatively fair to the other supervised learning-based methods. As reported in Tab. 1, we can observe that the performance of HSUM with VideoLLaMA2 [7] is competitive to the other methods. Furthermore, compared with the MLLM-based training-free methods [7, 31, 34, 35], the superior performances of HSUM are achieved on both $\mathcal{T}_{\mathrm{DDR}}$ and $\mathcal{T}_{\mathrm{DER}}$. Since HSUM is a training-free method, these results argue its potential superiorities as follows. **First**, the training procedure is not essential, where the significant labor costs of annotations could be avoided. HSUM could be more easily applied to different situations, such as different vehicles, viewpoints and driver state recognition tasks, without training and dataset collection. **Second**, HSUM would not suffer from the fixed classes in the training procedure, the unseen classes could be "known" adaptively. Intuitively, the performances of training-based methods would degenerate significantly for the unseen classes, since they are heavily reliant on the patterns and features present in the training data, and any deviation might lead to performance degradation.

4. Conclusion

In this paper, we propose a novel training-free driver state recognition method via human-centric context and self-uncertainty-driven MLLM (HSUM), in which the issues of understanding the contextual cues and alleviating the inherent uncertainty are addressed. Experimental results demonstrate that HSUM achieves the competitive performances in terms of driver distraction recognition and driver emotion recognition.

References

[1] Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *Society of Automotive Engineers (SAE) International Standard J3016_202104*, 2021. 1

[2] Andrei Aksjonov, Pavel Nedoma, Valery Vodovozov, Eduard Petlenkov, and Martin Herrmann. Detection and evaluation of driver distraction using machine learning and fuzzy logic. *IEEE Transactions on Intelligent Transportation Systems*, 20 (6):2048–2059, 2019. 1

[3] Zain Anwar Ali, Xinde Li, and Muhammad Ahsan Tanveer. Controlling and stabilizing the position of remotely operated underwater vehicle equipped with a gripper. *Wireless Personal Communications*, 116(2):1107–1122, 2021. 1

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 2, 3

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 3

[6] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1–26, 2021. 2

[7] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 2, 3

[8] Shi Dong, Ping Wang, and Khushnood Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021. 1

[9] Yilin Dong, Xinde Li, Jean Dezert, Rigui Zhou, Changming Zhu, Lei Cao, Mohammad Omar Khyam, and Shuzhi Sam Ge. Multisource weighted domain adaptation with evidential reasoning for activity recognition. *IEEE Transactions on Industrial Informatics*, 19(4):5530–5542, 2023. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2, 3

[12] Biswarup Ganguly, Debangshu Dey, and Sugata Munshi. An attention deep learning framework-based drowsiness detection model for intelligent transportation system. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–11, 2025. 1

[13] Ross Greer, Mathias Viborg Andersen, Andreas Møgelmose, and Mohan Trivedi. Driver activity classification using generalizable representations from vision-language models. *arXiv preprint arXiv:2404.14906*, 2024. 2, 3

[14] Md. Zahid Hasan, Jiajing Chen, Jiyang Wang, Mohammed Shaiqur Rahman, Ameya Joshi, Senem Velipasalar, Chinmay Hegde, Anuj Sharma, and Soumik Sarkar. Vision-language models can identify distracted driver behavior from naturalistic videos. *IEEE Transactions on Intelligent Transportation Systems*, 25(9):11602–11616, 2024. 2, 3

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3

[16] Chuanfei Hu and Yongxiong Wang. An efficient convolutional neural network model based on object-level attention mechanism for casting defect detection on radiography images. *IEEE Transactions on Industrial Electronics*, 67(12): 10922–10930, 2020. 1

[17] Chuanfei Hu, Chenyang Zhao, Hang Shao, Jin Deng, and Yongxiong Wang. Tmff: Trustworthy multi-focus fusion framework for multi-label sewer defect classification in sewer inspection videos. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024. 1

[18] Matthew Hutson. People don't trust driverless cars. researchers are trying to change that. *Science*, 14, 2017. 1

[19] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3mdad. *Signal Processing: Image Communication*, 88: 115960, 2020. 3

[20] Arief Koesdwiady, Ridha Soua, Fakhreddine Karray, and Mohamed S Kamel. Recent trends in driver safety monitoring systems: State of the art and challenges. *IEEE transactions on vehicular technology*, 66(6):4550–4563, 2016. 1

[21] Iuliia Kotseruba and John K. Tsotsos. Attention for vision-based assistive and automated driving: A review of algorithms and datasets. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):19907–19928, 2022. 1

[22] Kunyoung Lee, Hyunsoo Seo, Seunghyun Kim, Byeong Seon An, Shinwi Park, Yonggwon Jeon, and Eui Chul Lee. Quality-based rppg compensation with temporal difference transformer for camera-based driver monitoring. *IEEE Transactions on Intelligent Transportation Systems*, 26(2): 1951–1963, 2025. 1

[23] Zhuoren Li, Jia Hu, Bo Leng, Lu Xiong, and Zhiqiang Fu. An integrated of decision making and motion planning framework for enhanced oscillation-free capability. *IEEE Transactions on Intelligent Transportation Systems*, 25(6): 5718–5732, 2024. 1

[24] Khan Muhammad, Tanveer Hussain, Hayat Ullah, Javier Del Ser, Mahdi Rezaei, Neeraj Kumar, Mohammad Hijji, Paolo Bellavista, and Victor Hugo C. de Albuquerque. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):22694–22715, 2022. 1

[25] Gang Peng, Zhenyu Ren, Hao Wang, Xinde Li, and Mohammad Omar Khyam. A self-supervised learning-based 6-dof grasp planning method for manipulator. *IEEE Transactions on Automation Science and Engineering*, 19(4):3639–3648, 2021. 1

[26] Saed Rezayi, Zhengliang Liu, Zihao Wu, Chandra Dhakal, Bao Ge, Haixing Dai, Gengchen Mai, Ninghao Liu, Chen Zhen, Tianming Liu, et al. Exploring new frontiers in agricultural nlp: Investigating the potential of large language models for food applications. *IEEE Transactions on Big Data*, 2024. 1

[27] Glenn Shafer. *A mathematical theory of evidence*. Princeton university press, 1976. 2

[28] Hang Shao, Lei Luo, Jianjun Qian, Shuo Chen, Chuanfei Hu, and Jian Yang. Tranpulse: Remote photoplethysmography estimation with time-varying supervision to disentangle multiphysiologically interference. *IEEE Transactions on Instrumentation and Measurement*, 73:1–11, 2024. 1

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 2, 3

[30] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. 1

[31] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 3

[32] Dingkang Yang, Shuai Huang, Zhi Xu, Zhenpeng Li, Shunli Wang, Mingcheng Li, Yuzheng Wang, Yang Liu, Kun Yang, Zhaoyu Chen, et al. Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20459–20470, 2023. 3

[33] Haohan Yang, Haochen Liu, Zhongxu Hu, Anh-Tu Nguyen, Thierry-Marie Guerra, and Chen Lv. Quantitative identification of driver distraction: A weakly supervised contrastive learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 25(2):2034–2045, 2024. 1

[34] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 1, 2, 3

[35] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. 2024. 1, 2, 3