WARWICK

Robust Multiview Multimodal Driver Monitoring System Using Masked Multi-Head Self-Attention

Yiming Ma [1] Victor Sanchez [1] Soodeh Nikan [2] Devesh Upadhyay [2]
Bhushan Atote [1] Tanaya Guha [3]

[1] University of Warwick

[2] Ford Motor Company

[3] University of Glasgow

Sun 18th Jun, 2023

Outline

WARWICK

1. Introduction
2. Related Work
3. Method
4. Experiments

WARWICK

Introduction

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

WARWICK

Modern *driver monitoring systems* (DMSs) in Level-2+ self-driving-enabled cars aim to enhance safety by estimating drivers' readiness levels for driving and enabling safe control handovers when necessary.

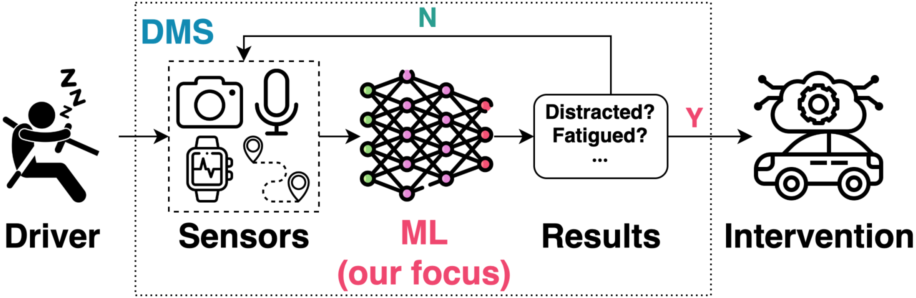


Fig. 1: A simplified illustration of a DMS.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

These systems usually rely various sensors, which may be deployed at different in-car locations, to comprehensively monitor drivers' states, e.g.,

- **RGB**: optical details.
- **Depth**: 3D information.
- **Infrared**: thermal information.
- **ECG**: heart rates.
- **Audio**: speech and sound.

Hence, modern DMSs are *multimodal* (and *multiview*).

Our work specifically focuses on *driver action recognition*, which involves classifying drivers' actions into *normal driving* and several *non-driving-related activities* (NDRAs), e.g., texting and drinking.

(a) Top IR **(b)** Top Depth **(c)** Front IR **(d)** Front Depth

Fig. 2: Sample frames from the DAD dataset [1].

Our contributions in this paper are as follows:

1. We propose a novel robust *multiview multimodal* DMS for driver action recognition that leverages feature-level fusion through masked *multi-head self-attention* (**MHSA**).
2. We manually annotated the anomalies in DAD dataset with 9 fine-grained classes of non-driving-related activities (NDRAs).
3. We conduct extensive experiments on the DAD dataset to compare different fusion strategies, assess the significance of individual views/modalities, and evaluate the efficacy of patch masking in enhancing MHSA's robustness against view/modality collapses. Results show that our MHSA-based DMS achieves state-of-the-art performance with an AUC-ROC score of 97.0%.

Related Work

- AUC-DD [2] is the first public dataset for DMSs. It was collected using an RGB camera from a single side view and thus have some limitations.

Fig. 3: A sample from the AUC-DD dataset [2] illustrating that RGB is not robust to illumination changes.

Related Work
Driver Monitoring Datasets

- Later databases [1], [3]–[5] have incorporated additional views and modalities to address these issues.
  - For example, top and front views have also been introduced to capture the driver's hand and head movements amongst other movements.
  - Regarding modalities, IR and depth have also become popular, as they can provide thermally based features and geometry information, which are complementary to the optical details from RGB.
- Among these datasets, we benchmark our models on DAD [1], the only one designed for SAE L2+ with open-set recognition: its test set contains extra classes of NDRAs in addition to those in the training split.

WARWICK

Various multiview multimodal DMSs have also been proposed with different emphases:

- Kopuklu *et al.* [1] proposed a novel learning framework based on contrastive learning.
- Ortega *et al.* [4] and Su *et al.* [6] proposed to leverage Conv-LSTM structures.
- Only Shan *et al* [7] proposed a feature-level modality fusion method, but it has several drawback:
    - Features are pooled before fusion, which leads to the loss of semantic information.
    - Its fusion module has the additional task of handling the temporal dimension.

Method

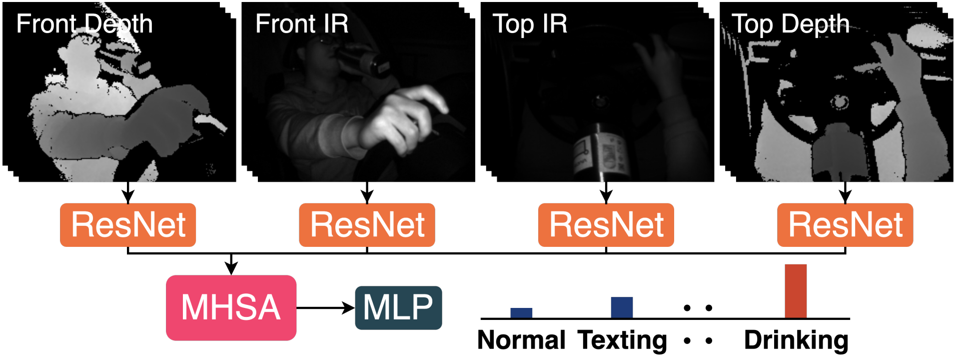


Fig. 4: An overview of our proposed DMS.

Method

Overview

1. We first use R3D-18 [8] to extract spatiotemporal features from the input multiview multimodal videos.
2. We the feed the feature maps to our masked multi-head self-attention module to interact and fuse the features.
3. We also used the supervised contrastive learning based on MoCo [9] to facilitate training.
4. The classifier is co-trained under the supervision of the cross-entropy loss.

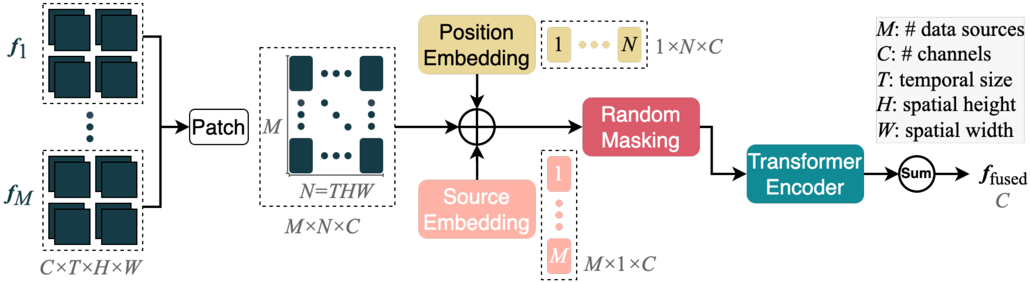
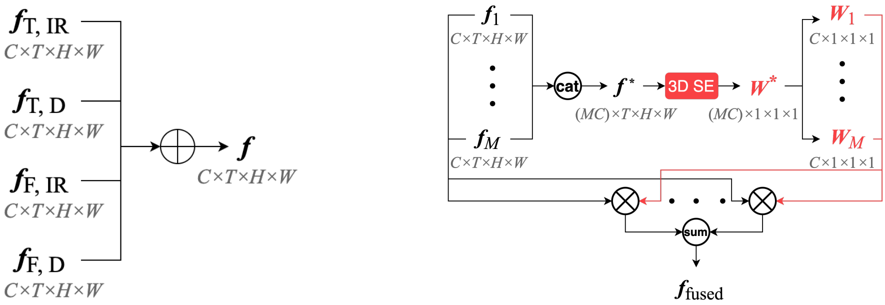


Fig. 5: The architecture of our masked multi-head self-attention module.

Method

Other Fusion Methods



(a) Sum. **(b)** SE.

Fig. 6: The architectures of the other fusion methods.

Experiments

Experiments

Detecting NDRAs

WARWICK

Sources	Decision [1]	Sum	Conv	SE	AFF	MHSA (our)
Top (D)	91.3	**92.9**				
Top (IR)	88.0	**91.3**				
Top (D+IR)	91.7	91.7	92.2	92.3	92.5	**92.9**
Front (D)	90.0	**91.7**				
Front (IR)	87.0	**90.2**				
Front (D+IR)	92.0	92.7	92.9	92.9	**93.1**	**93.1**
Top+Front (D)	96.1	94.8	05.8	05.9	96.5	**96.7**
Top+Front (IR)	93.1	94.5	94.6	94.9	95.0	**95.7**
Top+Front (D+IR)	96.6	96.3	96.2	96.4	96.7	**97.0**

Table 1: The AUC-ROC scores of different fusion methods on the NDRAs detection task on DAD. **D** and **IR** denote the depth and infrared modalities, respectively. The best scores for each view and modality are in **bold**.

Experiments

Classifiying Drivers' Actions

WARWICK

Source	Decision	Sum	Conv	SE	AFF	MHSA (ours)
Top (D)	84.3					
Top (IR)	83.7					
Top (D+IR)	84.5	85.0	85.4	85.4	85.4	**85.7**
Front (D)	87.7					
Front (IR)	83.7					
Front (D+IR)	87.9	87.7	88.1	88.2	88.5	**88.7**
Top+Front (D)	90.7	90.1	90.4	90.5	90.6	**90.9**
Top+Front (IR)	88.4	89.9	90.2	90.2	90.4	**90.6**
Top+Front (D+IR)	90.9	90.8	91.2	91.4	91.5	**91.6**

Table 2: The mAP scores for multi-classification of drivers' activities on DAD.

WARWICK

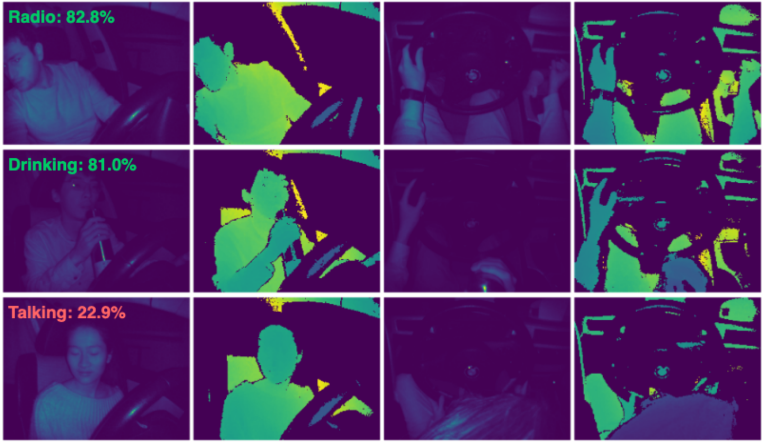
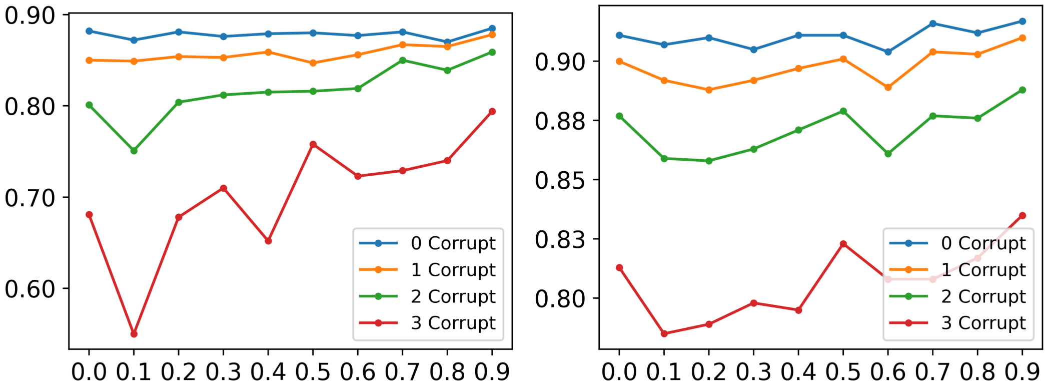


Fig. 7: Visualisation of the middle frames of four test samples from DAD.

Experiments

Robustness against Modality/View Collapses



(a) Mult. Acc. **(b)** Mult. mAP

Fig. 8: Masked training improves MHSA's robustness against corrupt views/modalities.

WARWICK

Thanks!

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

References I

WARWICK

[1] O. Köpüklü, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 91–100.

[2] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification," in *Neural Information Processing Systems (NIPS 2018), Workshop on Machine Learning for Intelligent Transportation Systems*, Dec. 2018.

[3] M. Martin, A. Roitberg, M. Haurilet, *et al.*, "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2801–2810.

[4] J. D. Ortega, N. Kose, P. Cañas, *et al.*, "Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *European Conference on Computer Vision*, Springer, 2020, pp. 387–405.

[5] I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub, "A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3mdad," *Signal Processing: Image Communication*, vol. 88, p. 115 960, 2020.

[6] L. Su, C. Sun, D. Cao, and A. Khajepour, "Efficient driver anomaly detection via conditional temporal proposal and classification network," *IEEE Transactions on Computational Social Systems*, 2022.

References II

[7] G. Shan, Q. Ji, and Y. Xie, "Multi-view vision transformer for driver action recognition," in *2021 6th International Conference on Intelligent Transportation Engineering (ICITE 2021)*, Springer, 2022, pp. 962–973.

[8] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.