Posture and Appearance Fusion Network for Driver Distraction Recognition

Hao Yu, Chong Zhao, Xing Wei, Yan Zhai, Zhen Chen, Guangling Sun and Yang Lu

November 1, 2022

Posture and Appearance Fusion Network for Driver Distraction Recognition

Hao Yu[1], Chong Zhao[2,3(✉)], Xing Wei[1,2,4], Yan Zhai[1], Zhen Chen[5], Guangling Sun[6], and Yang Lu[1,4]

[1] School of Computer and Information, Hefei University of Technology, China
[2] Intelligent Manufacturing Institute of Hefei University of Technology, China
[3] Engineering Quality Education Center of Undergraduate School, Hefei University of Technology, China
[4] Engineering Research Center of Safety Critical Industrial Measurement and Control Technology, Ministry of Education, China
[5] School of Computer Science and Technology, Anhui University, China
[6] School of Electronic and Information Engineering, Anhui Jianzhu University, China
`zhaochong@hfut.edu.cn`

Abstract. Distracted driving is the act of driving while engaged in other activities, such as using a cell phone, texting, eating, or reading, which takes the driver' attention away from the road. Nowadays, the distracted driving detection models based on deep learning can extract critical information from video data to characterize the driving behavior process. But the distraction driving method based solely on appearance features cannot essentially eliminate the noise impact of the complex environment on the model, and the distracted driving recognition method based solely on skeletal information is unable to recognize the joint action of the human body and the objects. Therefore,the development of an accurate distracted driving detection model has become challenging. In this paper, we propose a distracted driving recognition model MFD-former based on the fusion of posture and appearance. First, a feature extraction module is proposed to extract skeleton data(i.e., posture) and appearance features(i.e., descriptors), which are merged by a graph neural network. Then, the two kinds of information are input into the MFD-former encoder module, and the self-attention mechanism quickly extracts the sparse data. Finally, the classification results of distracted driving are obtained by extracting the classification labels through the MLP Head. The MFD-former model outperforms existing models. It achieved 95.1% accuracy on the State Farm dataset and 90.24% accuracy on the self-built Train Drivers dataset.

Keywords: Driver distraction recognition · Attention mechanism · Graph neural network· Heterogeneous information fusion

1 Introduction

Distracted driving is a phenomenon in which the driver's attention is directed to activities unrelated to normal driving (calling, smoking, etc.), which leads to a decline in driving ability. Distractions can multiply driving safety risks and even lead to traffic accidents. In recent years, using computer vision technology to monitor driver behavior and dynamically identify distracted driving has become a research hotspot [9].

The distracted driving recognition method, based on the object detection network VGG and so on [6], roughly divides the human body into different categories of targets, such as head, hand, shoulder, etc., and judges human behavior by its relative position, validated on public datasets such as Kaggle [4] for distracted driving. However, in a specific application environment, it will be affected by conditions such as background and lighting. For example, when a person's clothing color is similar to the background color, it may not be recognized correctly.

Due to the method relying solely on image features, it is still unable to accurately describe the spatiotemporal features of human behavior. Distracted driving recognition methods based on skeletal information, such as ST-GCN [14] and NAS-GCN [10], can separate human gesture recognition from the influence of light, and background, and have strong robustness. However, the appearance features cannot be integrated since the skeleton data is obtained. For example, when the driver uses the same hand to make a phone call or adjust his glasses, the movement of joint points is almost the same, so the two kinds of actions cannot be distinguished correctly.

In order to solve the above problems, this paper proposes **M**ulti-information **F**usion **D**river Activity Recognition network based on Trans**former** [13](MFD-former). First, the MFD-former uses OpenPose [3] to extract the coordinate information of the posture and then uses the Visual Descriptors Extraction module to extract the appearance features of the joint points, that is, the descriptors of the joint points. Then, the graph neural network is introduced to fuse the two types of information, input the fused information into the MFD-former Encoder, and use the self-attention mechanism to quickly extract the sparse data. Finally, the classification labels are extracted by Multilayer Perceptron Head(MLP Head) to obtain the classification results of distracted driving.

We compare MFD-former with other state-of-the-art baselines to highlight the advantages of the proposed approach. Moreover, we study the model at different scales to investigate the impact of the number of parameters and attention heads. To sum up, our contributions are as follows:

1. An attention-based MFD-former model is proposed, demonstrating that self-attentional architectures can outperform existing convolutional and graph convolutional distracted driving recognition models.

2. The graph neural network is used to fuse posture and descriptors, fully mining the driving behavior under the joint description of appearance features and skeletal data.

3. A visual descriptors extraction module is proposed to extract the descriptors in the images.

2 Related work

2.1 Graph Neural Network

The graph contains rich information, and many studies have begun to use deep neural network models to learn the feature representation of nodes in the network. Extending deep neural network models to non-Euclidean data, that is, graph convolutional neural networks(GCN), has become an emerging research hotspot [1]. HetGNN [15] obtains different types of neighbor nodes of each node by adopting random walk and restart sampling strategy, and then use different neural network modules to aggregate different characteristics of nodes, the same type of neighbor nodes, and different types of neighbor information, respectively. Finally, the representation of the node is formed. With the ability of these methods to represent graph data, graph heterogeneous graph neural models are also gradually applied to different types of information aggregation. This paper fuses the skeleton data and appearance features by graph neural networks.

2.2 Driver Behavior Recognition

Existing deep learning methods for driver distraction recognition mainly focus on two data patterns: Appearance features and Skeleton data.

Appearance features: As Koesdwiady proposed an end-to-end deep learning solution for distracted driving image recognition. The framework utilizes a pre-trained convolutional neural network VGG-19 [6] for feature extraction, and adds two fully connected layers to fine-tune VGG-19. Moslemi proposed to utilize 3D convolutional neural network and optical flow method [8] to improve driver distraction detection tasks to obtain helpful information from temporal information.

Skeleton data: This type of method uses the coordinates of the human skeleton to determine actions and processes the skeleton coordinates to obtain feature values that can be used to perform action recognition. There are two main methods for extracting skeleton data: (1) Using a camera (such as Controller-Pose [2]), (2) using various tools to extract key points from RGB images (such as OpenPose [3]). Maosen Li proposed an action recognition network ST-GCN [14], which uses motion structure graph convolution and temporal convolution as basic building blocks to learn the spatiotemporal characteristics of actions.

3 Proposed Method

In this section, we will introduce the architecture of the MFD-former(Fig. 1). We define $V = \{\mathrm{I}_t \in R^{H\times W\times 3}\}_{t=1}^{T}$ as the input, H and W are the height and width of the video, T is the frame number of the input video, and I_t is the video frame. First, the OpenPose extracts the skeleton data in the picture, and the Visual Descriptors Extraction module extracts descriptors. Two kinds of information are fused by graph neural network, that is, a Patch in Fig. 1

obtained in a frame of video. And a trainable vector Class[token] is added to the input sequence. This vector forces the self-attention in the MFD-former Encoder module to aggregate the information into a compact high-dimensional representation to separate different driver behavior categories. Then we need to input $T+1$ Patches into the MFD-former Encoder module, use the self-attention mechanism to extract the sparse data quickly, and finally use the MLP Head to extract the classification labels.

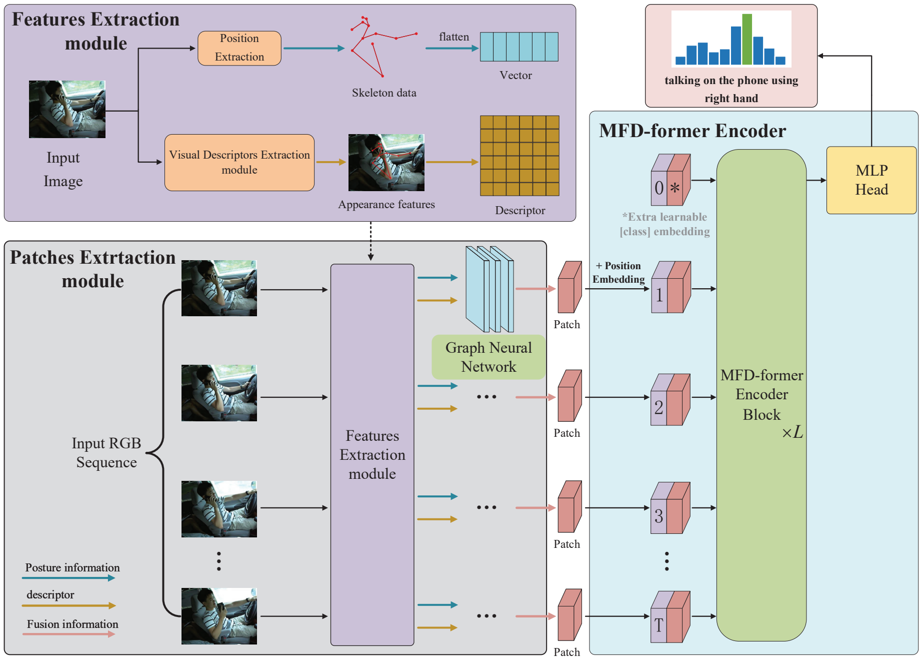


Fig. 1. Schematic of our MFD-former. The MFD-former consists of two network models: the Patches Extraction module(gray part) and MFD-former Encoder(blue part). The Patches Extraction module extracts the Patch on each frame of video, that is, the fusion information of the posture and appearance. $T+1$ Patches are input into the MFD-former Encoder module, and sparse data are extracted quickly by the self-attention mechanism, to obtain classification results through MLP Head. T is the frame number of the input video.

3.1 Patch Extraction module

The purpose of the module is to extract the Patch of each frame of the video, which is mainly composed of the Features Extraction module and the Graph Neural Network. Features Extraction module consists of the Position Extraction and Visual Descriptors Extraction module.

Position Extraction module In this article, the Position Extraction module is based on OpenPose [3]. The obtained posture $p_i := (x_i, y_i)$ of the driver represents the abscissa and ordinate of the i-th joint point. Distracted driving has nothing to do with the lower body, considering the specific driving scenario. To avoid and reduce the impact of lower limb movements on distracting behavior recognition, we will only consider the driver's upper limb nodes(1. Left ear 2. Left eye 3. Nose 4. Right eye 5. Right ear 6. Left wrist 7. Left elbow 8. Left shoulder 9. Neck 10. Right shoulder 11. Right elbow 12. Right wrist).

Visual Descriptors Extraction module We propose the Visual Descriptors Extraction module to extract descriptors for each joint. It mainly includes two parts: encoder blocker and decoder blocker. However, we get the descriptor of the whole image here, which is not all we need. We compare the p_i obtained by Position Extraction to find the descriptor d_i we need.

Graph Neural Network The p_i and d_i obtained by the Features Extraction module are two different types of information, so we use the GNN to fusion(Eq 1). In this paper, g_i is used to represent the fusion result of the posture feature of the i-th joint point and the descriptor, and ${}^{(0)}g_i$ represents the initial feature after fusion. Moreover, we use Multilayer Perceptron (MLP) to embed joint point information into a high-dimensional vector; The formula is as Eq 1.

$$ {}^{(0)}g_i = d_i + MLP_{enc}(p_i) \tag{1} $$

3.2 MFD-former Encoder

Patch Embedded What we get through the Graph Neural Network is a two-dimensional matrix. Before entering the MFD-former Encoder, we need to add [class]token and Position Embedding(Eq 2).

$$ z_0 = [g_{class}; g_i^1 E; g_i^1 E; \cdots; g_i^T E] + E_{pos},\ E \in R^{256\times D}, E_{pos} \in R^{(T+1)\times D} \tag{2} $$

We insert g_{class} for classification into the T tokens we just got, which is a trainable parameter. E_{pos} is the position code added to the original feature, which is related to the frame number in the video.

MFD-former Encoder architecture MFD-former is to stack the MFD-forme Encoder Block L times repeatedly. The MFD-forme Encoder Block consists of an alternation of Multi-Head Self-Attention (MSA) and MLP blocks.

$$ z'_l = MSA(LN(z_{l-1})) + z_{l-1},\ l = 1 \ldots L \tag{3} $$

Eq 3 is the MSA part, including multi-head self-attention, skip connection (Add) and layer normalization (Norm), which can repeat L times.

$$z_l = MLP(LN({z'}_l)) + {z'}_l,\ l = 1 \ldots L \tag{4}$$

Eq 4 is the MLP Block part, including feedforward network (FFN), skip connection (Add) and layer normalization (Norm), and can also repeat L times.

$$C = LN(z_L^0) \tag{5}$$

Eq 5 is layer normalization. An MLP Head is used as the classification head lastly, where is the output logit vector of the model, i.e. the classification labels. The other tokens are the only inputs to the module, but the supervision signal only comes from the [CLS] token.

4 Experiments

4.1 Datasets settings

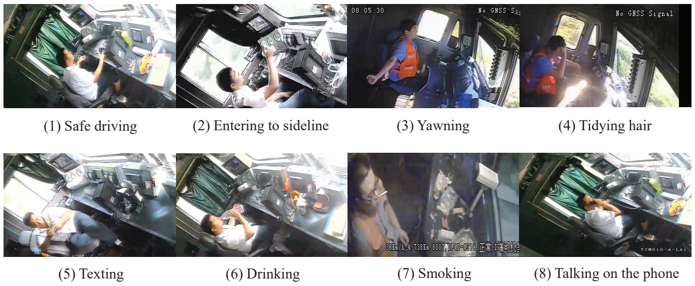
State Farm dataset The State Farm insurance company released a dataset available that classifies images into 10 categories [4]. Although the source dataset is still images, we reconstructed the time-series relationship from the CSV files in the original dataset. We obtained 20094 ten-frame sequences and 20094 ten-frame sequences, and appropriate labels were applied to each sequence.

Train Drivers dataset We built a specific dataset for train driver driving situations, divided into eight categories, which include two categories of normal driving videos and six categories of distracted driving videos. As shown in Fig. 2. We cut each video into about 3 seconds, and obtained a total of 9362 instances. To ensure the diversity of our data, we select participants with different heights, weights, and driving styles, wearing different uniforms, as shown in (1) and (2) in Fig. 2, and record videos with different brightness during the day and night, as shown in (1) and (7) in Fig. 2.

4.2 Implementation Details

We used the State Farm dataset and the self-built Train Drivers dataset in the following experiments. Moreover, we divide each action video into T frames, and input different versions of MFD-former on the experimental results. The hyperparameter analysis of the three versions of MFD-former is summarized in Table 1.

In experiments, we used the Adam Optimization Algorithm [5] for all training, $\beta_1 = 0.9, \beta_2 = 0.999$, a batch size of 1 and apply a high weight decay of 0.1. $T = 10$ when using the State Farm dataset, $T = 40$ when using the Train Drivers dataset. The ratio of training to test for both datasets is $6 : 4$.


(1) Safe driving (2) Entering to sideline (3) Yawning (4) Tidying hair

(5) Texting (6) Drinking (7) Smoking (8) Talking on the phone

Fig. 2. The Train Drivers dataset contains eight classes, including Category 2 Normal Driving and Category 6 Distracted Driving. (2) in the diagram is normal driving, the thumb and little finger of one hand stand up, indicating that the train enters the sideline. The dataset was recorded by multiple drivers in various environments. The video resolution is 1280 × 720 at 13Hz.

Table 1. Details of MFD-former model variants. Layers are the stacking times of the encoder block; Hidden size is the length of the token vector; MLP size is the number of nodes in the first fully connected layer of the MLP block in the Encoder block; Heads are Multi-head Attention the number of heads in.

Model	Layers	Hidden size D	MLP size	Heads	Params
Base	12	768	3092	12	86M
Large	24	1024	4096	16	307M
Huge	32	1280	5120	16	632M

4.3 Comparison Results

Results on State Farm dataset We chose four posture-based(Pos) models and three appearance-based(App) models, and all seven models have timing information. The experimental results are shown in Table 2. The baseline in the experiment is the traditional skeleton recognition ST-GCN model [14]. Posture-Based model data preprocessing uses OpenPose to extract skeleton data. It can be seen that the accuracy of our model has been greatly improved compared to the baseline. The results of ST-GCN are not ideal because it only relies on posture and ignores the importance of appearance features. These results validate the validity and rationality of our consideration of appearance features. Through the experimental results, we can see that the Ours-Huge model is better than the Ours-Large and Ours-Base model.

Results on Train Drivers dataset Table 3 shows the results of different methods on the Train Drivers dataset, selecting ST-GCN as the baseline. Our proposed method significantly outperforms baseline because ST-GCN which only relies on posture has certain limitations, and our method uses appearance features. We also have some improvements to the C3D method which only relies on appearance, so it is clear that our method is generally effective. The results

Table 2. Comparisons with state-of-the-art driver distraction recognition methods on the State Farm dataset. "Pos" and "App" denote posture and appearance, respectively.

Model	Modality	Accuracy[%]
BaseLine	Pos	77.4
2S-AGCN [12]	Pos	88.5
NAS-GCN [10]	Pos	89.5
ST-TR [11]	Pos	92.5
C3D, Sports 1M pre-training [7]	App	73.3
I3D-RGB, ImageNet + Kinetics pre-training [8]	App	91.8
I3D-two stream, ImageNet + Kinetics pre-training [8]	App	94.4
Ours-Base	App + Pos	92.5
Ours-Large	App + Pos	94.3
Ours-Huge	App + Pos	**95.1**

of the three models proposed in this paper on the Train Drivers dataset are still the Ours-Huge method with the highest accuracy.

Table 3. Comparisons with state-of-the-art driver distraction recognition methods on the Train Drivers dataset. "Pos" and "App" denote posture and appearance, respectively.

Model	Modality	Accuracy[%]
Baseline	Pos	75.21
NAS-GCN [10]	Pos	88.73
C3D, Sports 1M pre-training [7]	App	85.51
Ours-Base	App + Pos	88.52
Ours-Large	App + Pos	89.42
Ours-Huge	App + Pos	**90.24**

5 Conclusion

In this paper, we propose an MFD-former model to complete the classification of drivers' driving behavior to solve the problem of driver distraction detection. We explore the application of Transformer to driving behavior recognition research. We have verified the feasibility of our model on public datasets and our proposed Train Drivers dataset through multiple sets of experiments, and the accuracy is not lower than the current advanced time-series methods. In future work, we will continue to study the problem of driving behavior recognition, and try to improve the driving recognition speed under the high recognition accuracy.

Acknowledgements This work was supported by Joint Fund of Natural Science Foundation of Anhui Province in 2020 (2008085UD08), Anhui Provincial

Key R&D Program (202004a05020004), Open fund of Intelligent Interconnected Systems Laboratory of Anhui Province (PA2021AKSK0107), Intelligent Networking and New Energy Vehicle Special Project of Intelligent Manufacturing Institute of HFUT (IMIWL2019003, IMIDC2019002).

References

1. Abadal, S., Jain, A., Guirado, R., López-Alonso, J., Alarcón, E.: Computing graph neural networks: A survey from algorithms to accelerators. ACM Computing Surveys (CSUR) **54**(9), 1–38 (2021)
2. Ahuja, K., Shen, V., Fang, C.M., Riopelle, N., Kong, A., Harrison, C.: Controllerpose: Inside-out body capture with vr controller cameras. In: CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2022)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017)
4. Farm, S.: State farm distracted driver detection. Tech. rep., Technical Report. 2016. Available online: https://www. kaggle. com/c/state ... (2016)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. Koesdwiady, A., Bedawi, S.M., Ou, C., Karray, F.: End-to-end deep learning for driver distraction recognition. In: International Conference Image Analysis and Recognition. pp. 11–18. Springer (2017)
7. Lemley, J., Bazrafkan, S., Corcoran, P.: Transfer learning of temporal information for driver action classification. In: MAICS. pp. 123–128 (2017)
8. Moslemi, N., Azmi, R., Soryani, M.: Driver distraction recognition using 3d convolutional neural networks. In: 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA). pp. 145–151. IEEE (2019)
9. Moslemi, N., Soryani, M., Azmi, R.: Computer vision-based recognition of driver distraction: A review. Concurrency and Computation: Practice and Experience **33**(24), e6475 (2021)
10. Peng, W., Hong, X., Chen, H., Zhao, G.: Learning graph convolutional network for skeleton-based human action recognition by neural searching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 2669–2676 (2020)
11. Plizzari, C., Cannici, M., Matteucci, M.: Skeleton-based action recognition via spatial and temporal transformer networks. Computer Vision and Image Understanding **208**, 103219 (2021)
12. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
14. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
15. Zhang, C., Song, D., Huang, C., Swami, A., Chawla, N.V.: Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 793–803 (2019)