Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 3, 2024

Fatigue Driving Detection Based on Driver Facial Temporal Sequences

Yonglin Qian, Guoqiang Zheng, Yifan Xie, Xiangshuai Lv and Weizhen Zhang

School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China

Abstract: Addressing the issues of low real-time performance and high false positive rates in driver fatigue detection methods based on deep learning, this paper proposes a temporal sequence Transformer-based fatigue detection method grounded in the localization of facial landmarks in drivers. Initially, the facial positions are obtained using the single-stage face detection algorithm RetinaFace. Subsequently, a lightweight GM module is designed as the principal feature extraction module for constructing a multi-scale fusion facial landmark detection network, and facial fatigue feature parameters based on temporal sequences are calculated according to the facial landmarks. Finally, a fatigue driving classification method based on temporal sequences Transformers is developed for classifying the sequences of fatigue feature parameters. Experimental results demonstrate that the inference time required for facial detection and landmark detection is merely 16.8 milliseconds, with the per-frame inference time for facial landmark detection being only 2.5 milliseconds, thus fulfilling the real-time requirements of fatigue driving detection during the feature extraction phase. A temporal sequence of fatigue feature parameters built on the NTHU-DDD dataset and the trained temporal sequence Transformer model resulted in an accuracy rate of 91.4% for the proposed method.

Keywords: Fatigue driving detection; facial landmark detection; Transformer.

1. Introduction

According to estimates by the National Highway Traffic Safety Administration, fatigue driving contributes to 100,000 traffic accidents annually, resulting in over 1,550 deaths, 71,000 injuries, and a loss of $12.5 billion[1]. Therefore, real-time detection of a driver's state during driving and timely alerts when the driver is fatigued or drowsy are crucial for reducing traffic accidents caused by fatigue driving. Currently, fatigue driving detection methods can be mainly classified into three categories based on the type of detection parameters: detection based on vehicle behavior characteristics[2], detection based on physiological characteristics of the driver[3], and detection based on facial features of the driver. Detection of driver fatigue using facial features is more practical and promising compared to methods based on vehicle behavior and physiological characteristics, as it is non-contact, real-time, accurate, and low-cost.

Yi et al.[4]used the Dlib library to evaluate driver fatigue by measuring facial landmarks, blink frequency, eyelid closure, yawning, and nodding. Aicha et al.[5]analyzed fatigue using the eye and mouth aspect ratios and head movement, applying machine learning techniques like MLP and KNNs. However, determining driver fatigue through blink rate and frequency involves predefined, driver-specific thresholds that limit the method's general applicability and resistance to interference. Moreover, machine learning methods often fail to automatically learn drowsiness features from videos, decreasing their effectiveness in complex scenarios. Dua et al.[6]proposed an ensemble learning model to assess the driver's state, which, despite showing excellent performance in improving accuracy, possesses over 300 million trainable parameters, requiring substantial computational power support, thereby limiting its practical applicability. Liu et al.[7]proposed a fatigue detection method using CNN-LSTM, which utilizes CNN to extract features related to the eyes, mouth, and facial orientation, and then combines these with steering wheel feature parameters SA as inputs to the LSTM, with the level of fatigue as the output. Although the LSTM network is capable of extracting temporal features from the input data, enhancing the accuracy of driver drowsiness detection to some extent, it is better suited for processing short-term memory and short-term temporal dependencies. The main contributions of this paper are as follows:

1) A lightweight GM feature extraction module is designed, and based on this module, a multi-scale fusion facial landmark detection algorithm is constructed. This algorithm extracts fatigue feature parameters from the driver's facial temporal sequence through facial landmarks;

2) The Encoder part of the Transformer model is applied to the classification of fatigue driving temporal sequences, thereby effectively capturing the temporal characteristics of fatigue;

3) Experiments on facial landmarks and fatigue driving detection datasets validate the effectiveness of the proposed method.

2. Methodology

The temporal sequence Transformer-based fatigue driving detection method primarily consists of three steps: face detection, facial landmark localization, and fatigue state assessment. Initially, the driver's face is acquired using the RetinaFace[8] face detection algorithm. Then, based on the detected facial region, the driver's facial landmarks are localized using the landmark detection network MSGM-Net, and a temporal sequence of fatigue feature parameters is calculated based on these facial landmarks. Finally, the temporal sequence of fatigue feature parameters is fed into the Transformer classification method.

2.1. MSGM-Net Facial landmark Detection Algorithm

After acquiring the facial region of the driver, a further step involves extracting 98 landmarks from that area. The efficiency of facial landmark detection hinges on the design of the backbone network, which can both increase processing speed and alleviate the model's computational load, thus facilitating efficient and rapid inference. MobileOne[9], developed by Apple Inc., is an optimized backbone network for mobile devices that uses structural re-parameterization and a branching topology to enhance inference speed. This paper develops a lightweight feature extraction GM module based on MobileOne, incorporating phantom channels from the Ghost[10] module to increase inference speed without reducing feature map output. The structure of the GM module is depicted in Figure 2, divided into the training phase and the inference phase. Figure 2(a) illustrates the training structure based on the GM module. Figure 2(b) shows the inference structure after structural re-parameterization.

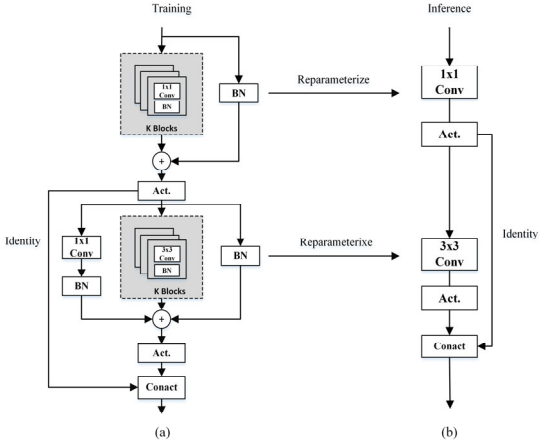

Figure 2. GM module structure diagram

The GM bottleneck layer structure is shown in Figure 4, composed by stacking GM modules. Figure 3(a) depicts the structure of the GM-bottleneck with a stride of 1. Figure 3(b) shows the structure of the GM-bottleneck with a stride of 2. The detailed configuration of the multi-scale fusion facial landmark detection network structure is shown in Table I, where n indicates the number of repetitions of that layer structure, s represents the stride of the first layer's convolution, c indicates the final output channel number for that row, and the expansion factor t is used to determine the output channel number applied to the intermediate layers.

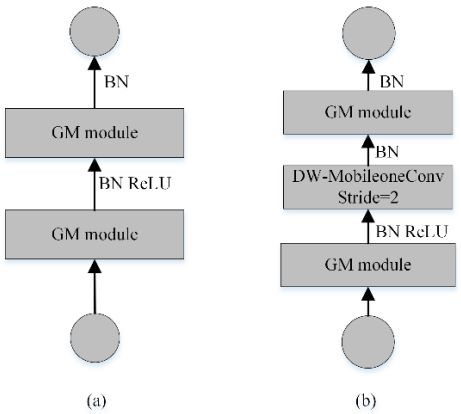

Figure 3. GM-bottleneck module structure

Table 1. Structure of landmark detection network

input	Operator	*t*	*c*	*n*	*s*
$112^2\times3$	Conv3×3	-	64	1	2
$56^2\times64$	Conv3×3	-	64	1	1
$56^2\times64$	GM-Bottleneck	2	80	3	2
$28^2\times80$	GM-Bottleneck	2	96	3	2
$14^2\times96$	GM-Bottleneck	4	144	4	2
$7^2\times144$	GM-Bottleneck	2	16	1	1
$7^2\times16$	Conv3×3	-	32	1	1
$7^2\times32$	Conv7×7	-	128	1	1
(*S1*)$56^2\times64$	AvgPool	-	64	1	-
(*S2*)$28^2\times80$	AvgPool	-	80	11	-
(*S3*)$14^2\times96$	AvgPool	-	96	1	-
(*S4*)$7^2\times144$	AvgPool	-	144	1	-
(*S5*)1×1×128	--	-	128	-	-
S1,S2,S3,S4,S5	Full Connection	-	196	1	-

2.2. Fatigue Feature Parameter Extraction

Based on the detection of 98 landmark, 12 points are selected to extract eye fatigue feature parameters. The formulas for calculating fatigue feature parameters of the left and right eyes are shown in Equations (2) and (3). This paper employs the Mouth Aspect Ratio (MAR) as a fatigue feature parameter, with its calculation formula presented as shown in Equation (4).

$$EAR_{left} = \frac{|y_{69}-y_{75}|+|y_{71}-y_{73}|}{2|x_{68}-y_{72}|} \tag{2}$$

$$EAR_{right} = \frac{|y_{61}-y_{67}|+|y_{63}-y_{65}|}{2|x_{60}-y_{64}|} \tag{3}$$

$$MAR = \frac{H}{W} = \frac{|y_{78}+y_{80}-y_{86}-y_{84}|}{2|x_{76}-x_{82}|} \tag{4}$$

This paper introduces head Euler angles, including pitch, yaw, and roll, as key feature parameters for fatigue detection. The calculation method for head Euler angles is as follows: (1) Utilize a 2D facial landmark detection network for facial landmarks; (2) Match a 3D face model; (3) Solve for the transformation relationship between 3D points and corresponding 2D points; (4) Calculate Euler angles based on the rotation matrix. The transformation formula for Euler angles is as equation(5). Where θ,ϕ,φ represent pitch, roll and yaw angles respectively.

$$\begin{cases} \theta = a\tan 2(r_{21}, r_{21}) \\ \phi = a\tan 2(-r_{20}, \sqrt{r_{21}^2 + r_{22}^2}) \\ \varphi = a\tan 2(r_{10}, r_{00}) \end{cases} \tag{5}$$

In fatigue driving detection, we extract a series of fatigue-related feature parameters by analyzing facial landmark information in each frame of the video. We use Equation (7) to represent the matrix of fatigue feature parameters for this multivariate time series.

$$X=\begin{bmatrix} t_1 & ear_{left} & ear_{right} & mar & \phi & \theta & \varphi \\ t_2 & ear_{left} & ear_{right} & mar & \phi & \theta & \varphi \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ t_n & ear_{left} & ear_{right} & mar & \phi & \theta & \varphi \end{bmatrix} \quad (7)$$

2.3. Fatigue Driving Discrimination Model

This paper selects the encoder part of the Transformer[11] to cater to the needs of fatigue driving discrimination. Initially, the fatigue feature parameter matrix undergoes a linear transformation to convert it into an embedding sequence, which is then fed into multiple layers of the Transformer encoder as input. For each fatigue feature parameter sample $X \in \mathbb{R}^{n\times m}$, which is a multivariate temporal sequence of length n and number of variables m ($m=7$ in this paper), constitutes a multivariate temporal sequence feature vector $x_t \in \mathbb{R}^m : X \in \mathbb{R}^{n\times m} = [x_1, x_2, ..., x_n]$. The original feature vector x_t is first normalized, and then linearly projected into the d dimensional vector space, with the linear transformation formula shown in Equation (8). Where $\mathrm{W_p} \in \mathbb{R}^{d\times m}, b_{\mathrm{p}} \in \mathbb{R}^d$ are parameters that can be learnt and $\mathbf{u_t} \in \mathbb{R}^d, t=0,...,n$ is the input part of the Transformer model Enconder. Since the

$$\mathrm{u_t} = \mathrm{W_p x_t} + \mathrm{b_p} \quad (8)$$

Transformer is an architecture that is not sensitive to the order of inputs, in order to incorporate the sequential nature of temporal sequences, in this paper, the position encoding $W_{pos} \in \mathbb{R}^{n\times d}$ is added to the input vector. $U \in \mathbb{R}^{n\times d} = [u_1, ..., u_n]$ represents the linearly transformed feature vector, and Z^0 represents the resulting time-aware feature sequence.

$$Z^0 = U + W_{pos} \quad (9)$$

To capture intricate interactions among fatigue feature parameters in time series, the input Z^0 undergoes encoding through a multilayer Transformer.

$$\begin{aligned} head_j &= Attention(Q_j, K_j, V_j) \\ &= soft\max(\frac{Q_j K_j^T}{\sqrt{d}})V_j \\ &= soft\max(\frac{Z^0 W_j^Q (Z^0 W_j^K)^T}{\sqrt{d}})Z^0 W_j^V \end{aligned} \quad (10)$$

where $Q_j = Z^0 W_j^Q, K_j = Z^0 W_j^K, V_j = Z^0 W_j^V$ and W represent learnable linear projection parameters. Specifically it is the projection of Z^0 into a different subspace of N_h via Q, K, V. The multi-head self-attention can be expressed as Equation(11). N_h denotes the number of distinct heads, $concat$ represents the splicing operation, and W^o represents the linear projection parameters.

$$MSHA(Z^0) = concat(head_1, ..., head_{N_h})W^o \quad (11)$$

Each Transformer's Enconder contains N multi-head self-attention blocks. The standard Transformer encoder forward computation is as follows:

$$\widehat{Z^i} = MHSA(LN(Z^{i-1})) + Z^{i-1} \quad (12)$$

$$Z^i = MLP(LN(\widehat{Z^i})) + \widehat{Z^i} \quad (13)$$

$\widehat{Z^i}$ and Z^i denote the output and final output of the intermediate layer of layer i, respectively. MLP composed of two linear feedforward layers and a GELU non-linear activation function. To adapt it to the classification task of fatigue driving detection, the final output of the Transformer Encoder model Z_0^N is normalized to obtain the output. Further, the model passes through fully connected layers $Softmax$ and to obtain the final probability of the fatigue state.

$$Y = LN(Z_0^N) \quad (14)$$

$$y = soft\max(Y) \quad (15)$$

Where Y is the output of the multilayer encoder. The Transformer model for time series can obtain the correlation between the fatigue feature parameters before and after each frame in the video, which can better capture the dynamic changes of the fatigue state, thus improving the accuracy and reliability of the detection.

3. Experiment

3.1. Dataset

To assess the performance of the facial landmark detection algorithm, experiments were conducted on the WFLW dataset. The NTHU-DDD dataset, created by National Tsing Hua University, aims to simulate typical scenarios under nearly real driving conditions. To construct temporal sequence samples, each video was divided into consecutive 6-second interval samples, with a 3-second interval between each sampling, ultimately creating a new dataset.

3.2. Evaluation Metrics

In order to evaluate the performance of the proposed landmark detection method, Normalised Mean Error (NME) is used. NME is commonly used as a common metric to measure the performance of landmark detection algorithms. The NME for each image is defined as (16). The accuracy of the fatigued driving discrimination model is represented using

Equation (17).

$$NME\left(P,\hat{P}\right)=\frac{1}{M}\sum_{i=1}^{M}\frac{\left\|p_i-\hat{p}_i\right\|_2}{d} \tag{16}$$

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{17}$$

3.3. Results Analysis

Landmark Detection: To evaluate the performance of the proposed multi-scale fusion facial landmark detection algorithm, this paper compares the algorithm with current landmark detection algorithms on the WFLW dataset. The algorithm's normalized mean error (NME) values, model size, and inference speed were compared with other algorithms, with results presented in Table 2.

Table 2. Landmark Detection Result Diagram

Methods	Model Size(MB)	Inference Time(ms)	NME(%)
LAB[12]	50.7	2600	5.27
PFLD[13]	5.0	5.5	5.52
Ours	2.71	2.5	5.21

Research [12] achieved a 5.27% normalized mean squared error on the dataset by adopting boundary-aware feature extraction, structured output networks, multi-task learning, and boundary refinement techniques. This method demonstrated good robustness in handling situations involving complex expressions and pose changes. Despite its robustness to complex expressions and pose changes, the model size is as large as 50.7MB, and the inference time reaches 2600ms. Research[13] utilized MobileNet as the backbone network for landmark detection, with a model size of 5.0MB and an NME of 5.52%. Although there was improvement in inference speed and model size, there is still room for further optimization. The method presented in this paper surpasses the above two methods in terms of model size and inference time, requiring only 2.71MB for the model and achieving an inference speed of just 2.5ms, while maintaining an NME performance of 5.21%. The performance improvement of this paper's algorithm is attributed to the replacement of traditional convolutional networks with pointwise and depthwise convolutions and the introduction of ghost channel structures in the GM module, effectively reducing computational load. Furthermore, by employing a re-parameterization structure during the training phase to enhance the model's expressive capability and re-parameterizing it into a linear structure for the inference phase, this method is able to accelerate the model's inference speed while maintaining high accuracy.

Fatigue Discrimination Model: The performance of our fatigue driving detection algorithm compared to existing fatigue detection methods is shown in Table 3.

Table 3. Landmark Detection Result Diagram

Methods	Feature Extraction	Temporal Feature	Accuracy (%)
Literature[14]	HOG	MLP	74.9
Literature[15]	CNN		87.5
Literature[16]	3D cGAN	BILSTM	91.2
Ours	RetinaFace+MSGM-Net	Transformer	91.4

The comparison and analysis focus on both facial spatial features and fatigue temporal characteristics. Research [14] utilized Histogram of Oriented Gradients (HOG) technology to locate faces and landmarks, extracting spatial features such as the Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and head posture angles. It employed Multi-Layer Perceptron (MLP) and K-Nearest Neighbors (KNN) algorithms for fatigue state classification, achieving an accuracy of 74.9%. Research[15] used a 2D Convolutional Neural Network to detect faces, employed Hough Transform to locate the driver's eyes, and determined the eye's open or closed state. Fatigue temporality was analyzed using the PERCLOS threshold method, ultimately achieving an accuracy of 87.5%. Compared to these methods, our approach combines RetinaFace and the multi-scale fusion facial landmark detection network (MSGM-Net) for spatial feature extraction, and processes temporal features through the Transformer model, achieving a classification accuracy of 91.4%. Although the detection accuracy in Research[16] is close to our algorithm, differing by only 0.02%, their method utilized a 3D convolutional generative network, with inference times reaching up to 25 seconds, which does not meet the real-time requirement for fatigue driving detection.

4. Conclusion

This paper proposes a temporal sequence Transformer model for fatigue driving detection based on the localization of driver facial landmarks. By integrating the MobileOne module with ghost channel technology, we designed a lightweight GM feature extraction module for a multi-scale fusion facial landmark detection algorithm. Fatigue feature parameter sequences for the face, mouth, and head posture angles are extracted based on landmark information, and a temporal sequence Transformer is utilized to classify these fatigue feature parameter sequences. Experiments demonstrate that the proposed algorithm has good generalization performance and high real-time detection capability. However, this study only utilized a subset of facial landmarks. In future work, considering more landmarks is planned to further improve accuracy.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62072158, U2004163), the Key Research and Development Special Projects of Henan Province(231111221500), and Science and Technology Project of Henan Province(232102210158, 242102210197).

References

[1] Facts and stats, May 2018. [Online]. Available: http://drowsydriving.org/about/facts-and-stats/.

[2] Su-xian C A I, Chao-kan D U, Si-yi Z, et al. Fatigue driving state detection based on vehicle running data[J]. Journal of Transportation Systems Engineering and Information Technology, 2020, 20(4): 77.

[3] Shujuan Gong, Yongxiang Zhao, Deming Huang. Research on fatigue driving detection based on multi-physiological signal fusion. Journal of Transportation Systems Engineering and Information Technology. 2023, Vol.143, p. 4002.

[4] Yi Y, Zhang H, Zhang W, et al. Fatigue working detection based on facial multifeature fusion[J]. IEEE Sensors Journal, 2023, 23(6): 5956-5961.

[5] Ghourabi A, Ghazouani H, Barhoumi W. Driver drowsiness detection based on joint monitoring of yawning, blinking and nodding[C]. 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing. IEEE, 2020: 407-414.

[6] Dua M, Shakshi, Singla R, et al. Deep CNN models-based ensemble approach to driver drowsiness detection[J]. Neural Computing and Applications, 2021, 33: 3155-3168.

[7] Liu M Z, Xu X, Hu J, et al. Real time detection of driver fatigue based on CNN-LSTM[J]. IET Image Processing, 2022, 16(2): 576-595.

[8] Deng J, Guo J, Zhou Y, et al. Retinaface: Single-stage dense face localisation in the wild[J]. arXiv preprint arXiv:1905.00641, 2019.

[9] Vasu P K A, Gabriel J, Zhu J, et al. Mobileone: An improved one millisecond mobile backbone[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7907-7917.

[10] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[12] WU W, QIAN C, YANG S, et al. Look at boundary: A boundary-aware face alignment algorithm[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2129-2138.

[13] GUO X, LI S, YU J, et al. PFLD: A practical facial landmark detector[J]. arXiv preprint arXiv: 1902.10859, 2019.

[14] A. Ghourabi, H. Ghazouani, and W. Barhoumi, Driver drowsiness detection based on joint monitoring of yawning, blinking and nodding. 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing. IEEE, 2020, pp. 407–414.

[15] J. Xing, G. Fang, J. Zhong, and J. Li, Application of face recognition based on cnn in fatigue driving detection. Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing, 2019, pp. 1–5.

[16] Y. Hu, M. Lu, C. Xie, and X. Lu, Driver drowsiness recognition via 3d conditional gan and two-level attention bi-lstm. IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 12, pp.4755–4768, 2019.