Gaze Behavior Patterns for Early Drowsiness Detection

Hongfei Gao, Ruimin Hu(✉), and Zijun Huang

National Engineering Research Center for Multimedia Software (NERCMS), School of Computer Science, Wuhan University, Wuhan, China
{gaohongfei,hrm,huangzijun}@whu.edu.cn

Abstract. Early drowsiness detection could be crucial for some occupations such as drivers and monitors, as it can greatly improve safety and efficiency. However, most existing drowsiness detection methods do not consider the early stages of drowsiness or the practical feasibility of detection. To address this issue, we propose a gaze behavior pattern-based drowsiness detection model that effectively distinguishes early drowsiness. First, we extract the gaze behavior features of subject from the video, which is composed of eye aspect ratio, head pose and gaze direction. Then we perform a preliminary analysis of the correlation between the gaze behavior features and different stages of drowsiness and propose a multi-stream Transformer model to obtain the classification result from the feature sequences. Our proposed model uses two encoders to encode the temporal and channel information respectively from the gaze behavior features. We conducted experiments on the largest publicly available multi-stage drowsiness video dataset RLDD. Preliminary analysis of the dataset showed the distribution of the features of our selected gaze behavior patterns over different drowsiness stages had relatively significant differences. For early drowsiness detection problem, experiments on real dataset demonstrate the effectiveness of our approach compared to state-of-the-art methods.

Keywords: Early drowsiness detection · Gaze behavior patterns · Transformer · Deep learning

1 Introduction

Drowsiness detection is an important and difficult problem, successful solutions could be used in occupations such as drivers and monitors. Based on careful analysis, experts believe the real number of annual fatalities due to drowsy driving in the U.S. may be closer to 6,000. This would mean drowsiness is involved in approximately 21% of fatal crashes every year. Between hospital admissions, property damage, and other costs, the estimated societal cost of drowsy driving in the U.S. may be anywhere between $12.5 billion and $109 billion per year. In addition, studies show that, when driving for a long period of time, drivers lose their self-judgment on how drowsy they are [17], and this can be

L. Iliadis et al. (Eds.): ICANN 2023, LNCS 14254, pp. 220–231, 2023.
https://doi.org/10.1007/978-3-031-44207-0_19

one of the reasons that many accidents occur close to the destination. Research has also shown that sleepiness can affect workers' ability to perform their work safely and efficiently [15]. These troubling facts above prompted us to look for a method to detect and alert people before they fall into drowsiness completely. It is commonly recognized [11,13] that there are three main types of sources of information in drowsiness detection: Performance measurements, physiological measurements, and the behavioral measurements.

Performance measurements focus on subjects' performance of work. For instance, in the driving domain, it is reflected as steering wheel movements, driving speed, brake patterns, and lane deviations, etc. An example is the Autopilot system of Tesla, by measuring the grip on the steering wheel or directly using the lane departure warning system (LDWS) driving data it can obtain the time and degree of vehicle deviation from the lane, and then analyze and project the driver's drowsiness level or whether the driver is distracted. In addition to being expensive, the similar kind of solutions are difficult to redeploy to different workplaces. Some other performance measurements at workplace can be obtained by testing workers' reaction time and short-term memory [3]. Many of these methods can also not be used in other workplaces and the measurement itself can have an impact on the results.

Physiological measurements can use heart rate, electrocardiogram (ECG), electromyogram (EMG), electroencephalogram (EEG) [6,8,14] and electrooculogram (EOG) [8] to monitor drowsiness. However, these methods are intrusive and not practical to deploy in the car or workspace even though they have higher accuracy. Some wearable and convenient devices like hats and watches have been proposed as an alternative for such measurements, but they are still not practical to be used for long time.

Behavioral measurements are mostly obtained from subject's facial movements and expressions which could be captured non-intrusively by a single camera. This data acquisition method not only require lower cost but also is highly versatile, and can be used in almost any workplace, including the field of driver drowsiness detection. And with the rapid development of deep learning and computer vision techniques, behavioral measurements method will play a more important role in the field of drowsiness detection.

Comparing the above three types of methods, the most potential and practicable is the drowsiness detection based on behavioral measurements. However, it is rather difficult to detect early drowsiness in generic workplace using only the behavioral data recorded by the video. The challenges for the recognition are mainly in the early stage. Early drowsiness in real workplace is not evident externally and is highly susceptible to be misclassified into normal or drowsy stage. Most of the existing methods for drowsiness detection are based on videos of pretend drowsiness in laboratory scenarios and most of them are aimed for drowsiness driving only. And in some researches, the early drowsiness stage is often ignored directly.

Thus, in this paper, we propose a vision-based early drowsiness recognition method that aims to extract the behavior patterns of the different drowsiness

stages of the subjects in the video to detect early drowsiness in real workplace scenarios. In summary, this paper makes the following contributions:

- We extracted sequences data of eye aspect ratio (EAR), gaze direction, and head posture data for the RLDD dataset. By analyzing the distribution and correlation between these feature sequences, we verified that the sequences of these features in different drowsiness states express different behavioral patterns of the subjects, which can be used for early drowsiness detection better.
- We design a multi-stream transformer model for early drowsiness detection, which learns the gaze behavior patterns of the subjects in the videos by the channel and temporal information from the feature sequences to classify different drowsiness states.
- The experimental results show that our method has a strong advantage in early drowsiness detection in real workplace scenarios based on video. And our model significantly outperforms the baseline method of the RLDD dataset and other multivariate time series classification methods.

2 Preliminary

In this section, we first present the RLDD dataset and our preprocessing operation of it. Then we perform an exploratory analysis to disclose the gaze behavior patterns and further demonstrate the motivation for our proposed model.

2.1 Dataset

The RLDD dataset proposed by Ghoddoosian et al. [5] is the largest to date realistic drowsiness dataset. It was created for the task of multistage drowsiness

Table 1. KSS drowsiness scale and drowsiness state categories

Drowsiness State	Description	Score
Normal	Extremely alert	1
	Very alert	2
	Alert	3
Early Drowsy	Rather alert	4
	Neither alert nor sleepy	5
	Some signs of sleepiness	6
	Sleepy, no difficulty remaining awake	7
Drowsy	Sleepy, some effort to keep alert	8
	Extremely sleepy, fighting sleep	9

detection, targeting not only extreme and easily visible cases, but also subtle cases of drowsiness. The RLDD dataset consists of around 30 h of RGB videos of

60 healthy participants. For each participant they obtained one video for each of three different classes: alertness, low vigilance, and drowsiness, for a total of 180 videos. Subjects were undergraduate or graduate students and staff members who were from different ethnicities and ages. Videos were taken from roughly different angles in different real-life environments and backgrounds. Each video was self-recorded by the participant, using their cell phone or web camera. We reclassified the videos in the dataset into the three categories in Table 1. based on the KSS table [1] and the original labels. Our exploratory analysis and the experimental part are all performed on this RLDD dataset.

2.2 Preprocessing and Feature Extraction

The motivation behind using gaze behavior features: eye aspect ratio, gaze direction, and head pose, was to capture temporal patterns that appear naturally in human gaze behavior and could easily be overlooked by spatial feature detectors like CNNs. We used dlib's pre-trained face detector based on a modification to the standard Histogram of Oriented Gradients + Linear SVM method for object detection [4]. Then we calculate eye aspect ratio with the six facial landmarks per eye (Fig. 1), and use the average value of two eyes as the EAR. For each eye, we denote:

$$EAR = \frac{\overline{AE} + \overline{BD}}{2\overline{FC}} \tag{1}$$

where $\overline{AE}$, $\overline{BD}$ and $\overline{FC}$ is the length of the line segment connecting the corresponding points in the Fig. 1.

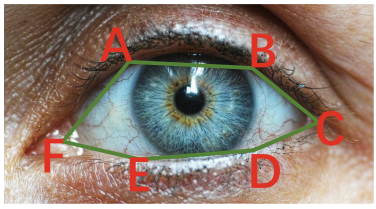


Fig. 1. The eye landmarks to define EAR for each eye.

In contrast to the blink features of Ghoddoosian et al. [5], we used original eye aspect ratio of each frame of videos to preserve the continuity of the time series so that it could be able to form multidimensional time series with other features of gaze behavior.

For the head pose and gaze direction, we use the preprocessing pipeline from [12] to obtain 3D head pose since the dataset does not provide camera parameters and we plug ResNet50 [7,19] to the PnP-GA framework [10] to obtain 3D gaze direction. Both of the extracted head pose and gaze direction are presented by

pitch and yaw angle. All angles are converted to the camera-based coordinate system so that they can be used together for drowsiness detection. A visual representation of these two features is shown in Fig. 2 and Fig. 3.

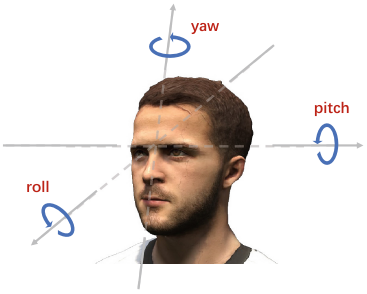


Fig. 2. The pitch, yaw and roll angle of head pose.

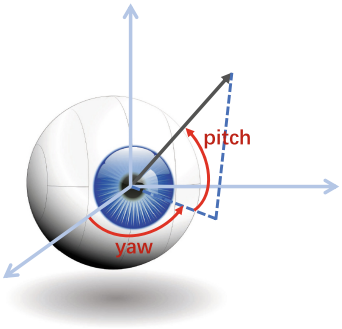


Fig. 3. The pitch and yaw angle of gaze direction

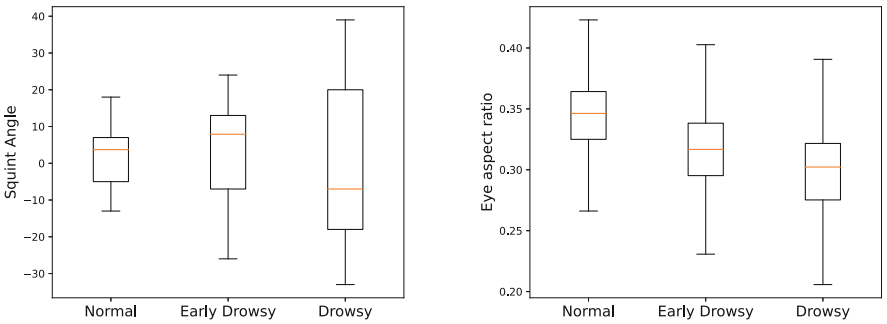
It is worth noting that for head pose and gaze direction, their pitch angle and yaw angle could be corresponded one by one in that they are both used to express the direction to which the head or the eye is directing. As for the roll angle of head pose, because the head can be tilted, for example, with the palm of the hand propped up diagonally, we need a roll angle to measuring its degree of inclination. And obviously, this angle could be significantly different in different drowsiness phrase. For the direction of gaze, we do not consider the complex eye structure here, we simply treat the eye as a sphere or a point, so the direction of gaze does not need a roll angle.

For each frame of the videos of RLDD dataset, we have a 6-dimention feature: $\{EAR,\ pitch_h,\ yaw_h,\ roll_h,\ pitch_g,\ yaw_g\}$, in which EAR means the average eye aspect ratio, $pitch_h$, yaw_h and $roll_h$ means the pitch angle, yaw angle and

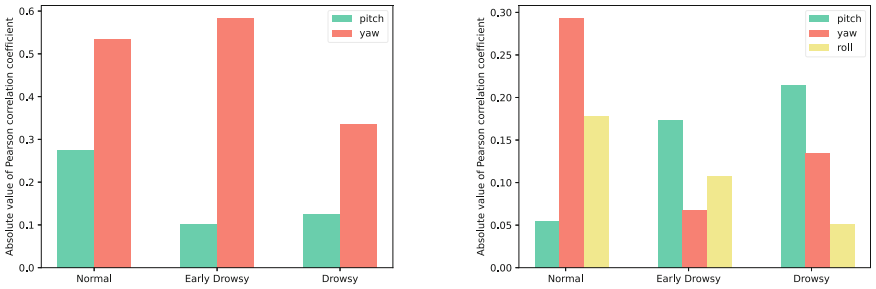
roll angle of head pose, $pitch_g$ and yaw_g means the pitch angle and yaw angle of gaze direction. These are the gaze behavior features that we use as the input of our early drowsiness detection model.

2.3 Exploratory Analysis

Given the real-world early drowsiness dataset, we next convey several exploratory analyses on all subjects from different perspectives to distinguish different drowsiness state.



(a) Squint Angle (b) Eye Aspect Ratio



(c) EAR and gaze direction (d) EAR and head pose

Fig. 4. Feature comparison between different drowsiness states

From $\{pitch_g,\ yaw_g\}$ and $\{pitch_h,\ yaw_h\}$, the squint angle θ_s of gaze direction and head pose can be calculated. In general, when a person looks at someplace, the head will also turns to the direction of looking accordingly, so the direction of a person's gaze direction and the direction of the head pose should be basically the same, that is, the θ_s is not large. But if a person is in a drowsy state, then he is likely to lean on the chair or tilt the body to look at the screen, when the squint angle of gaze direction and head pose could be larger. We made

a statistic of the squint angle under different drowsiness states for all subjects in the dataset, as shown in Fig. 4(a). The distribution of the data fits well with our assumption: Gradual increase in squint angle with increasing drowsiness.

As for EAR, it is a widely used feature in drowsiness detection. Here we show its distribution over 3.24 M frames of the dataset in the Fig. 4(b). Even though there is some individual variability in the value of EAR, it is still possible to see a certain degree of differentiation in its distribution across the different drowsiness phases. As drowsiness increases, the eye aspect ratio tends to decrease. Then we show visualization of EAR in relation to gaze direction and head posture for all subjects in the dataset for the three drowsiness states in Fig. 4(c) and Fig. 4(d). Here we calculate the absolute value of the Pearson correlation coefficient of the EAR with each angular component of gaze direction and head pose. It can be seen that EAR has a higher linear correlation with the yaw angle component of the gaze direction, while the linear correlation with all other angular components is low (<0.3).

Through the above analysis of the components of gaze behavior, we can find that the distribution of EAR and Squint Angle at different drowsiness states has been significantly distinguishable. Although the linear correlation between EAR and head pose and gaze direction is not very high, the distribution of their values in different drowsiness states can still reflect slight difference of different drowsiness states. Therefore, we have reason to believe that the gaze behavioral features composed of the above-mentioned components can effectively extract the behavioral pattern of different drowsiness states and help us effectively classify the early drowsiness state, and our subsequent experimental results also corroborate our assumption.

3 Proposed Model

Our drowsiness detection model is based on the Transformer Network [16]. For natural language processing problems, traditional Transformer has encoder and decoder stacking on the word and positional embedding for sequence generation and forecasting task. As for multivariate time series classification, we have several modifications to adapt the Transformer for our need. The overall architecture of our early drowsiness detection transformer model is shown in Fig. 5.

3.1 Embedding

We use the gaze behavior feature extracted from the video in the preprocessing section as the input of our early drowsiness detection model, it's a 6-dimensional time series. We divide the input into temporal stream feature and channel stream feature by time step and channel.

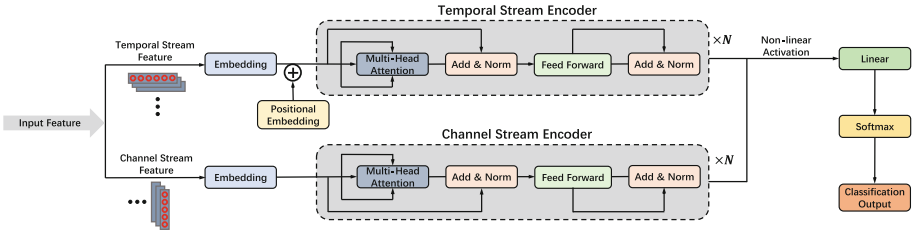


Fig. 5. Overview of our multi-stream transformer model

In the original Transformers [16], all the tokens are projected to an embedding layer. For the time series data is continuous, we replace the embedding layer with fully connected layer. Instead of the linear projection, we use a non-linear activation *tanh*. The positional encoding is added with the temporal stream feature to encode the temporal information to utilize the sequential correlation of time step better.

3.2 Multi-stream Encoder

Gaze behavior features we extract has multiple channels where each channel is a multi-variate time series. The common assumption is that there exists some hidden correlation between different channels. Capturing both the temporal and channel information is the key for our early drowsiness detection.

One of the usual approaches is to apply convolutions, that is, the reception field integrates both channel and temporal feature by the 2D kernels or the 1D kernels with fixed parameter sharing. We design a multi-stream extension where the encoders in each stream explicitly capture the channel and temporal correlation by attention and masking, as shown in Fig. 5.

Different from the natural language processing task, our task in this step is actually a multi-variate time series classification task, so we do not need the decoder [16] part of traditional Transformer.

Then, to merge the information of the two streams which encodes temporal and channel correlations respectively, we use a fully connect layer after out put of both encoders with the non-linear activation as T and C. Then we use a linear projection layer to get h:

$$h = W \cdot Concat(T, C) + b \tag{2}$$

Through the softmax function, the streaming weight of each stream are computed as s_T and s_C:

$$s_T,\ s_C = Softmax(h) \tag{3}$$

Each streaming weight is attending on the corresponding stream's output and we get the final feature vector f:

$$f = Concat(T \cdot s_T,\ C \cdot s_C) \tag{4}$$

3.3 Loss Function

To verify the effectiveness of our model on early drowsiness detection task, we design a supervised leaning framework on labeled video dataset. the loss function of this framework is as follows:

$$y_{pre} = \sigma(W^{'} f + b^{'}) \tag{5}$$

$$Loss = -\frac{1}{N} \sum_{N} \sum_{i=0}^{2} y_i ln(p_i) \tag{6}$$

It's the loss function for our triple classification task. y_{pre} is classification probability of model output: $y_{pre} = [p_0,\ p_1,\ p_2]$, p_i is predictive probability of category i. And y_i is the onehot representation of the sample y: $y = [y_0,\ y_1,\ y_2]$, when the sample y belongs to category i, $y_i = 1$, otherwise $y_i = 0$.

4 Experiment

4.1 Implementation Details

We used one fold of the RLDD dataset as our test set, and the remaining four folds for training. After repeating this process for each fold, the results were averaged across the five folds. All experiments are carried out with 6-dimensional time series with a window length of 256 frames on the server, the step size of the window movement is set to 128 frames and the batch sizes are 128. The optimizer is uniformly used Adam, the learning rate is $4e-5$, the experiments were conducted on a server with two NVIDIA A40 GPUs.

4.2 Evaluation Metrics

Accuracy, *Precision*, *Recall*, and $F1$-*score* are used to evaluate models. Four evaluation indexes are computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

For our target of early drowsiness detection problem, we choose *Accuracy* and *Precision*, *Recall*, and *F*1-*score* of early drowsy category to evaluate our model.

4.3 Results

We used SVM [2], HM-LSTM [5], and two state-of-the-art multivariate time series classification methods: LSTM-FCNs [9] and TapNet [18] as our comparison methods. The overall metrics of the RLDD dataset are shown in Table 2. From this table, we observe that our method achieves significant results in the overall *accuracy* and *Precision*, *Recall*, and *F*1-*score* of early drowsiness state respectively, proving that our multi-stream Transformer model has good performance on early drowsiness drowsiness detection problem. Compared with the baseline HM-LSTM, all other methods using our proposed gaze behavior features performer much better on early drowsiness classification, and this indicates that our proposed gaze behavior features have a large contribution in early drowsiness detection. Among all methods using gaze behavior features for classification, our proposed model have the best performance in overall accuracy and three metrics of early drowsiness classification.

Table 2. Performance of early drowsiness detection using different classification methods on the RLDD dataset.

Method	*Accuracy*	*Precision**	*Recall**	*F*1-*score**
HM-LSTM	0.6522	0.5105	0.3233	0.3959
SVM	0.6333	0.5112	0.5333	0.5220
LSTM-FCNs	0.7311	0.7213	0.6900	0.7053
TapNet	0.7411	0.6698	0.7033	0.6861
Ours	**0.7833**	**0.7309**	**0.7333**	**0.7321**

* The matrics of the early drowsy category

4.4 Ablation Study

We study the effectiveness of the channel stream encoder and each part of our proposed gaze behavior features:1)**Channel Stream Encoder:**We remove the channel stream encoder from the model and use the same gaze behavior features for training and classification. The results in Table 3 show that there is large reduction in the overall metrics of the model after the removal of channel stream encoder. This result indicates that the Channel stream Encoder part of our model plays a significant role in our early drowsiness detection model. 2)**Gaze Behavior Features:** Our proposed gaze behavior features is composed of EAR, gaze direction and head pose. To discover the extent to which each of these features contributes to the final classification result, we remove 1 or 2 of the three

input features and get the corresponding classification results. The results show a drop in classification results when either 1 or 2 features are removed which means each part of the gaze behavior features contributes to the final classification result, of these, EAR is the most important, followed by gaze direction and head pose.

Table 3. Performance after removing channel stream encoder or parts of gaze behavior features on the RLDD dataset

Method	*Accuracy*	*Precision**	*Recall**	*F*1-*score**
Ours	0.7833	0.7309	0.7333	0.7321
Ours(remove channel stream encoder)	0.6188	0.5570	0.5700	0.5634
Ours(EAR)	0.6511	0.5980	0.6000	0.5990
Ours(head pose)	0.5889	0.4685	0.5200	0.4929
Ours(gaze direction)	0.6289	0.5710	0.5767	0.5738
Ours(EAR + head pose)	0.6933	0.6063	0.6367	0.6211
Ours(EAR + gaze direction)	0.6922	0.5868	0.6533	0.6183
Ours(head pose + gaze direction)	0.7244	0.6495	0.6733	0.6612

5 Conclusion

This paper addressed the problem of early drowsiness detection, and found an approach to early drowsiness detection on gaze behavior feature sequence learning by learning the variability of subjects of different drowsiness state in gaze behavior through data analysis. Based on the found variability in features of gaze behaviors between subjects of different drowsiness state, we propose a new variant model of Transformer by changing the internal structure of it to consider gaze behaviors of subjects in combination with drowsiness state and using an channel stream encoder a spatial stream encoder to better recognize gaze behavior patterns which can improve the accuracy of early drowsiness detection effectively. Experiments on real dataset demonstrate the effectiveness of the gaze behavior features and model we proposed compared to state-of-the-art methods, especially for the early drowsiness state. Our research may provide a reference for the drivers and monitors' drowsiness monitoring and warning system.

References

1. Åkerstedt, T., Gillberg, M.: Subjective and objective sleepiness in the active individual. Int. J. Neurosci. **52**(1–2), 29–37 (1990)
2. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**(2), 121–167 (1998)

3. Caldwell, J.A., Caldwell, J.L., Thompson, L.A., Lieberman, H.R.: Fatigue and its management in the workplace. Neurosci. Biobehav. Rev. **96**, 272–289 (2019)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference On Computer Vision And Pattern Recognition (CVPR'05), vol. 1, pp. 886–893. IEEE (2005)
5. Ghoddoosian, R., Galib, M., Athitsos, V.: A realistic dataset and baseline temporal model for early drowsiness detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 178–187 (2019)
6. Guarda, L., Tapia, J., Droguett, E.L., Ramos, M.: A novel capsule neural network based model for drowsiness detection using electroencephalography signals. Expert Syst. Appl., 116977 (2022)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Jiao, Y., Deng, Y., Luo, Y., Lu, B.L.: Driver sleepiness detection from EEG and EOG signals using GAN and LSTM networks. Neurocomputing **408**, 100–111 (2020)
9. Karim, F., Majumdar, S., Darabi, H., Harford, S.: Multivariate LSTM-FCNS for time series classification. Neural Netw. **116**, 237–245 (2019)
10. Liu, Y., Liu, R., Wang, H., Lu, F.: Generalizing gaze estimation with outlier-guided collaborative adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3835–3844 (2021)
11. Ngxande, M., Tapamo, J.R., Burke, M.: Driver drowsiness detection using behavioral measures and machine learning techniques: a review of state-of-art techniques. In: 2017 pattern recognition Association of South Africa and Robotics and mechatronics (PRASA-RobMech), pp. 156–161 (2017)
12. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: Proceedings of the IEEE/CVF International Conference On Computer Vision, pp. 9368–9377 (2019)
13. Reddy, B., Kim, Y.H., Yun, S., Seo, C., Jang, J.: Real-time driver drowsiness detection for embedded system using model compression of deep neural networks. In: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition Workshops, pp. 121–128 (2017)
14. Reddy, T.K., Behera, L.: Driver drowsiness detection: an approach based on intelligent brain-computer interfaces. IEEE Syst. Man Cybern. Mag. **8**(1), 16–28 (2022)
15. Sadeghniiat-Haghighi, K., Yazdi, Z.: Fatigue management in the workplace. Ind. Psychiatry J. **24**(1), 12 (2015)
16. Vaswani, A., et al.: Attention is all you need. In: Advances Neural Information Processing Systems, vol. 30 (2017)
17. Wheaton, A.G., Shults, R.A., Chapman, D.P., Ford, E.S., Croft, J.B.: Drowsy driving and risk behaviors–10 states and Puerto Rico, 2011–2012. Morb. Mortal. Wkly Rep. **63**(26), 557 (2014)
18. Zhang, X., Gao, Y., Lin, J., Lu, C.T.: TapNet: multivariate time series classification with attentional prototypical network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6845–6852 (2020)
19. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: ETH-XGaze: a Large scale dataset for gaze estimation under extreme head pose and gaze variation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 365–381. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_22